ΑΝΝΑ ΜΑΡΙΑ ΚΟΚΛΑ

Μαθηματικός Πανεπιστήμιο Αθηνών

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

και

ΠΡΟΒΛΕΨΗ ΣΤΕΦΑΝΙΑΙΑΣ ΝΟΣΟΥ

Διδακτορική Διατριβή

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΝΟΕΜΒΡΙΟΣ 2000

« Η έγκριση διδακτορικής διατριβής υπό της Σχολής Στατιστικής και Ασφαλιστικής Επιστήμης. του Πανεπιστημίου Πειραιώς δεν υποδηλοί αποδοχή των γνωμών του συγγραφέως».

(Ν. 5343/1932, άρθρ. 202)

ΑΝΝΑ ΜΑΡΙΑ ΚΟΚΛΑ

Μαθηματικός Πανεπιστήμιο Αθηνών

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ *και* ΠΡΟΒΛΕΨΗ ΣΤΕΦΑΝΙΑΙΑΣ ΝΟΣΟΥ

Διδακτορική Διατριβή

Υποβληθείσα στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς

Η εξεταστική επιτροπή που ορίστηκε στη Συνεδρίαση της Γενικής Συνέλευσης του Τμήματος την 28. 6. 2000, εγκρίνει την παρούσα διατριβή ως πληρούσα τις προϋποθέσεις για την απονομή του τίτλου

Διδάκτορα Στατιστικής

Η ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Φώτης Γεωργιακώδης
Αν. Καθηγητής, Καθηγητής,
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης
Πανεπιστήμιο Πειραιώς
(Επιβλέπων Καθηγητής)

Μιχάλης Παπαδάκης
Καθηγητής,
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης
Πανεπιστήμιο Πειραιώς
(Μέλος Συμβουλευτικής Επιτροπής)

Κλέων Τσίμπος
Αν. Καθηγητής,
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης
Πανεπιστήμιο Πειραιώς
(Μέλος Συμβουλευτικής Επιτροπής)

Βασίλης Μπένος
Καθηγητής,
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης
Πανεπιστήμιο Πειραιώς

Τάκης Παπαϊωάννου
Καθηγητής,
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης
Πανεπιστήμιο Πειραιώς

Θεόδωρος Αρτίκης
Καθηγητής,
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης
Πανεπιστήμιο Πειραιώς

Κωνσταντίνος Συριόπουλος
Επίκουρος Καθηγητής,
Τμήμα Οικονομικών
Πανεπιστήμιο Μακεδονίας

ΠΕΙΡΑΙΑΣ

ΝΟΕΜΒΡΙΟΣ 2000

Η εργασία αυτή αφιερώνεται σε όλους όσους βοήθησαν στην πραγματοποίησή της.

Στη μνήμη του πατέρα μου

στην Πηνιώ

στο Σπύρο

και στο Δημήτρη

ΕΥΧΑΡΙΣΤΙΕΣ

Στη διάρκεια της εκπόνησης της εργασίας αυτής, αρκετοί ήταν αυτοί που αφιέρωσαν πολύτιμο χρόνο για να μου προσφέρουν τη βοήθειά τους. Θα ήθελα λοιπόν να ευχαριστήσω:

τον Αναπληρωτή Καθηγητή του Πανεπιστημίου Πειραιώς και υπεύθυνο για την παρακολούθηση της εργασίας, *Φ. Γεωργιακώδη* για το θέμα της διατριβής που πρότεινε, τις γνώσεις που μου προσέφερε, και τις παρατηρήσεις του

τον Καθηγητή *Μ.Παπαδάκη* και τον Αναπληρωτή Καθηγητή *Κ.Τσίμπο* για τις χρήσιμες παρατηρήσεις που ως μέλη της τριμελούς, έκαναν,

τον Καθηγητή *Τ. Παπαϊωάννου* για τις υποδείξεις και οδηγίες του,

τον Καθηγητή *Θ. Αρτίκη* για τις χρήσιμες παρεμβάσεις και συμβουλές του,

τα μέλη Δ.Ε.Π και το λοιπό προσωπικό του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιά για το φιλικό κλίμα συνεργασίας και εμπιστοσύνης που υπήρχε σε όλη τη διάρκεια της προσπάθειας αυτής,

τον Επίκουρο Καθηγητή του Πανεπιστημίου Μακεδονίας *Κ. Συριόπουλο* για τη σημαντική προσφορά του στον τομέα των νευρωνικών δικτύων και την εταιρεία PROFILE A.E. για την παροχή του αντίστοιχου λογισμικού συστήματος,

τον καρδιολόγο Επίκουρο Καθηγητή Πανεπιστημίου Αθηνών *Γ. Βυσσούλη* στον οποίο οφείλεται η συλλογή των δεδομένων,

τον Δρα *Γ. Τζαβέλα* για τη χρήση του λογισμικού συστήματος MATHEMATICA και τις χρήσιμες παρατηρήσεις και συμπληρώσεις που έκανε,

τον αδελφό μου *Σ. Κόκλα*, για την προσφορά του στον τομέα της χρήσης των υπολογιστών, και

το σύζυγό μου *Δ. Γεωργιακώδη*, Δρα Φυσικής Στερεάς Κατάστασης, για την πολύτιμη συμπαράστασή του σε όλη τη διάρκεια της εργασίας, τη συμβολή του στην κατανόηση των νευρωνικών δικτύων καθώς και για τις εύστοχες παρατηρήσεις του.

Νοέμβριος 2000

Α.Μ. Κόκλα

ΠΡΟΛΟΓΟΣ

Σκοποί της παρούσας εργασίας είναι: α) η αξιολόγηση των παραγόντων κινδύνου, που ενοχοποιούνται την εκδήλωση της στεφανιαίας νόσου β) ο καθορισμός κατάλληλων μοντέλων πρόβλεψης ικανών να διαχωρίζουν ένα σύνολο ατόμων με συμπτώματα στεφανιαίας νόσου σε υγιείς και ασθενείς, χρησιμοποιώντας τεχνικές πολυμεταβλητής ανάλυσης ή αρχές των νευρωνικών δικτύων γ) η αξιολόγηση των μετρικών αποτιμήσεων των μοντέλων της πολυμεταβλητής ανάλυσης και των τεχνητών νευρωνικών δικτύων σε πιλοτικό δείγμα και η σύγκριση και απόδοσή της στο σύνολο των δεδομένων και η σύγκριση της απόδοσής τους και δ) η ανάπτυξη μαθηματικού μοντέλου για την περιγραφή του δείκτη Gensini (βαθμός βαρύτητας νόσου) .

Στο πρώτο κεφάλαιο περιγράφεται το δείγμα και οι κύριες μεταβλητές που έχουν καταγραφεί. Εξετάζονται οι παράγοντες κινδύνου που συσχετίζονται με την εκδήλωση της νόσου, και μελετώνται οι παράγοντες αυτοί στο σύνολο ανδρών και γυναικών.

Στο δεύτερο κεφάλαιο επιχειρείται η διάγνωση της στεφανιαίας νόσου και η εξέλιξή της σε σχέση με τους παράγοντες κινδύνου. Για το σκοπό αυτό χρησιμοποιήθηκε πιλοτικό δείγμα 160 ατόμων (80 άνδρες - 80 γυναίκες) όπου με τη βοήθεια παραγοντικής ανάλυσης επισημάνθηκαν οι παράγοντες εκείνοι που ενοχοποιούνται περισσότερο για την εμφάνιση της νόσου. Διαχωριστικές τεχνικές, τα αποτελέσματα των οποίων χρησιμοποιήθηκαν για το διαχωρισμό του συνολικού δείγματος σε ασθενείς και υγιείς, εφαρμόστηκαν επίσης στο πιλοτικό αυτό δείγμα.

Η έννοια των νευρωνικών δικτύων και η χρήση τους στη διάγνωση και εξέλιξη της στεφανιαίας νόσου εισάγεται στο 3° κεφάλαιο. Με τη βοήθεια του προγράμματος Profile Neural Applications-P.N.A επιτυγχάνεται η εκπαίδευση δέκα δικτύων με χρησιμοποίηση των δεδομένων του πιλοτικού δείγματος. Τα αποτελέσματα «εκπαίδευσης» των νευρωνικών δικτύων στο πιλοτικό δείγμα χρησιμοποιήθηκαν για την ταξινόμηση του συνόλου των δεδομένων (660 άτομα) και συγκρίθηκαν με σκοπό την αξιολόγηση της αποτελεσματικότητας των μοντέλων και τη σύγκριση της απόδοσης.

Ο υπολογισμός του βαθμού βαρύτητας της νόσου με τη χρήση του δείκτη Gensini μελετάται στο 4° κεφάλαιο. Στα πλαίσια της μελέτης αυτής επιχειρείται η βελτίωση του δείκτη Gensini μέσα από διαδικασίες εξομάλυνσης, η οποία επιτυγχάνεται με τη χρήση ενός νεοεισαγόμενου δείκτη dsn.

Ο δείκτης αυτός βασίζεται σε κατασκευασθείσα συνάρτηση ds(i, y, j), όπου i η θέση έμφραξης της αρτηρίας j, με κλάσμα έμφραξης y. Ο υπολογισμός των τιμών της συνάρτησης αυτής στο αρχείο ασθενών ανδρών, επιτρέπει τον αντικειμενικό υπολογισμό του βαθμού βαρύτητας της νόσου, στηριζόμενη στις κρίσιμες τιμές στένωσης των αρτηριών όπως προσδιορίστηκαν από τον G. Gensini. Περιγράφονται επίσης τα στατιστικά χαρακτηριστικά της συνάρτησης, χρησιμοποιείται δε ως μέσο πρόβλεψης καρδιακών παθήσεων *όπως η παράπλευρη κυκλοφορία και το έμφραγμα*

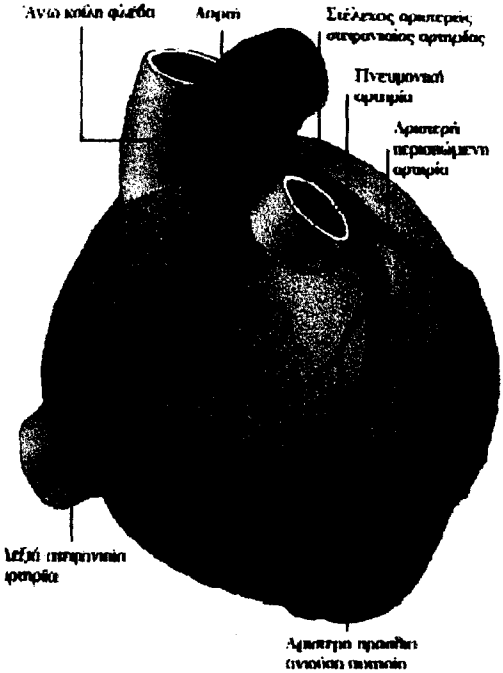
Στο ίδιο κεφάλαιο μελετάται και η συσχέτιση των παραγόντων κινδύνου *στο δείγμα των ασθενών. Στο δείγμα των ασθενών ανδρών συσχετίστηκαν* οι παράγοντες με το ανεύρυσμα, την παράπλευρη κυκλοφορία, το έμφραγμα και την ακινησία τμήματος καρδιακού μυός. Ενώ παράλληλα η χρήση παραγοντικής ανάλυσης επιχειρείται για την ανάδειξη των σημαντικότερων παραγόντων που ευνοούν για την εκδήλωση της νόσου. Η αντίστοιχη ανάλυση στο δείγμα των ασθενών γυναικών δεν έγινε γιατί δεν υπήρχαν επαρκείς πληροφορίες.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΕΙΣΑΓΩΓΗ

Η καρδιά, για να μπορέσει να διεκπεραιώσει τη λειτουργία της έχει ανάγκη από παροχή οξυγονομένου αίματος. Αυτή η λειτουργία γίνεται με τη βοήθεια των δύο στεφανιαίων αρτηριών.

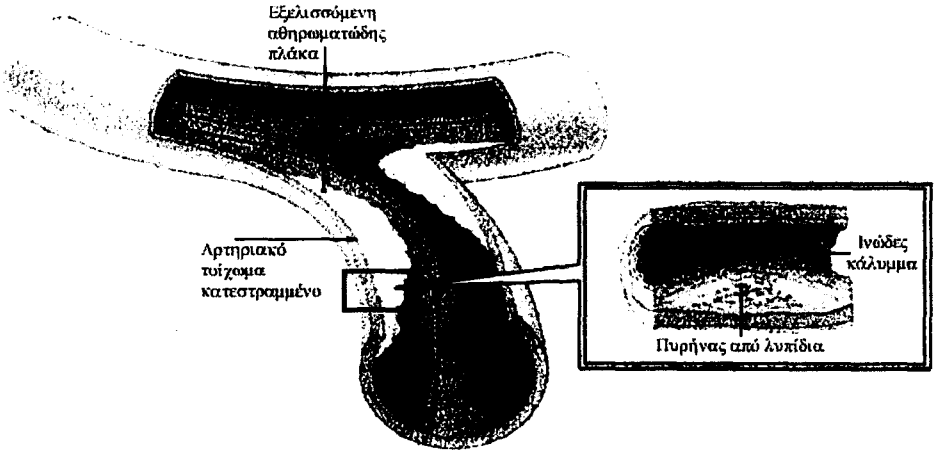
Οι στεφανιαίες αρτηρίες εκφύονται από την αορτή λίγο πιο πάνω από την αορτική βαλβίδα (Σχ. 1.1). Η αριστερή στεφανιαία αρτηρία αιματώνει την αριστερή κοιλία και τον αντίστοιχο κόλπο. Λίγο μετά την έκφυσή της, η αρτηρία αυτή χωρίζεται σε δύο άλλες την πρόσθια ανιούσα και την περισπωμένη αρτηρία, που τροφοδοτούν με αίμα το πρόσθιο και το αριστερό τμήμα της καρδιάς. Η δεξιά στεφανιαία αρτηρία πορεύεται γύρω από το δεξιό τμήμα της καρδιάς



Σχ.1.1

και φθάνει μέχρι την οπίσθια επιφάνεια. Η αρτηρία αυτή αιματώνει τη δεξιά κοιλία και τον αντίστοιχο κόλπο. Λόγω του χωρισμού της αριστερής στεφανιαίας σε δύο κλάδους, οι στεφανιαίες αρτηρίες μπορούν να θεωρηθούν, από λειτουργικής πλευράς, ως τρεις. Ένα δίκτυο βοηθητικών διακλαδώσεων που εκπορεύονται από τις αρτηρίες διεισδύει βαθιά μέσα στη μυϊκή μάζα. Πλήρης έμφραξη, σε οποιοδήποτε τμήμα του συστήματος αυτού των αγγείων, προκαλεί το «θάνατο» στην αντίστοιχη περιοχή του καρδιακού μυός - κατάσταση γνωστή ως έμφραγμα του μυοκαρδίου ή καρδιακό επεισόδιο. Η βαρύτητα του επεισοδίου εξαρτάται από την έκταση της περιοχής του μυοκαρδίου που στερείται αιμάτωσης.

Η πρόοδος της χειρουργικής, τα μέσα συνεχούς παρακολούθησης στις μονάδες εντατικής θεραπείας και η φαρμακευτική αγωγή είχαν ως αποτέλεσμα τη μείωση των θανάτων των εμφραγματιών που εισάγονται στο νοσοκομείο. Όμως μεγάλο ποσοστό θανάτων από στεφανιαία νόσο συμβαίνουν αιφνίδια, σε ανθρώπους ελεύθερους συμπτωμάτων. Αυτά είναι τα θύματα της σιωπηλής επιδημίας του 20ου αιώνα της αθηρωμάτωσης (Σχ. 1.2).



Σχ. 1.2

Θα μπορούσε κανείς να ορίσει την αθηρωμάτωση ως «νόσο πάχυνσης του έσω χιτώνα του αρτηριακού τοιχώματος», με αποτέλεσμα τη στένωση του αυλού και την παρεμπόδιση της αιματικής ροής. Σε όλους τους ανθρώπους, το

αθήρωμα αρχίζει να σχηματίζεται, κάτω από το επένδυμα των αρτηριών, από την παιδική ηλικία.(AMA (1991))

Οι λιποπρωτεΐνες - σωματίδια τα οποία μεταφέρουν τη χοληστερίνη, ουσία υπεύθυνη για το σχηματισμό του αθηρώματος - διεισδύουν και εγκλωβίζονται κάτω από το επένδυμα των αρτηριών με συνέπεια την πρόκληση φλεγμονής και την ανάπτυξη ουλώδους ιστού. Μεγάλοι σχηματισμοί γνωστοί ως αθηρωματώδεις πλάκες σχηματίζονται στις ευαίσθητες περιοχές. Αιματικοί θρόμβοι είναι δυνατό να σχηματιστούν πάνω στις αθηρωματώδεις πλάκες αν η επιφάνειά τους είναι ανώμαλη ή αν η αιματική ροή είναι βραδεία. Επιπλέον, το ασβέστιο προκαλεί σκλήρυνση των πλακών, με αποτέλεσμα τη σκλήρυνση των αρτηριών. Με την πάροδο των ετών, οι πλάκες αυξάνονται σε μέγεθος, ώσπου προοδευτικά η αρτηρία στενώνεται ή φράσσεται τελείως.

Οι αθηρωματικές πλάκες (Σχ. 1.2) έχουν την τάση να σχηματίζονται στους διχασμούς των αρτηριών, όπου το αίμα αποκτά στροβιλώδη ροή. Οι ανώμαλες, κατεστραμμένες επιφάνειες των πλακών επιτείνουν τη στροβιλώδη ροή. Ο στροβιλισμός αυτός διεγείρει τον πηκτικό μηχανισμό του αίματος, με αποτέλεσμα το σχηματισμό ενός θρόμβου. Στην περίπτωση που ο θρόμβος δε διαλυθεί, αυξάνεται σε μέγεθος και τελικά φράσσεται ο αυλός του αγγείου και αποστερεί το όργανο από οξυγονομένο αίμα. Μια εγκατεστημένη πλάκα αναπτύσσεται περαιτέρω με τη συσσώρευση αιματοπεταλίων και λευκών αιμοσφαιρίων. Είναι δυνατόν κάποιο τμήμα της πλάκας να αποσχιστεί και να σχηματίσει θρόμβο στην επιφάνεια του αθηρώματος και τελικά να προκληθεί έμφραξη του αγγείου
(έμφραξη της στεφανιαίας αρτηρίας) και έμφραγμα του μυοκαρδίου. (AMA(1991) και Κυριακίδης(1987)).

ΚΕΦΑΛΑΙΟ 1°

ΠΑΡΟΥΣΙΑΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ

1.1 Συλλογή και κατάταξη στοιχείων

Σε χρονικό διάστημα πέντε ετών συγκεντρώθηκαν από το Ιπποκράτειο Γενικό Περιφερειακό Νοσοκομείο Αθηνών στοιχεία για 660 άτομα από τα οποία 230 ήταν γυναίκες (ποσοστό 34.8%) και 430 άνδρες (ποσοστό 65.2%). Στο σύνολο, τα 80(12,1%) άτομα κρίθηκαν υγιή και τα υπόλοιπα 580(87,9%) ασθενή. Το 66% των ατόμων του δείγματος προέρχονται από το νομό Αττικής και το υπόλοιπο από άλλες περιοχές της χώρας πλην των νήσων Αιγαίου και Μακεδονίας. Η έλλειψη ασθενών από τις δύο τελευταίες περιοχές μπορεί να οφείλεται στην ύπαρξη αντίστοιχης νοσοκομειακής μονάδας στη Θεσσαλονίκη.

Τα άτομα του δείγματος είχαν εμφανίσει συμπτώματα στεφανιαίας νόσου. Καταγράφηκαν τα αρχικά συμπτώματα, η ηλικία και ο τρόπος ζωής η προσωπικότητα και οι συνήθειες κάθε ατόμου το οποίο στη συνέχεια υποβλήθηκε σε εξετάσεις. Οι εξετάσεις περιελάμβαναν μεταξύ άλλων αναλύσεις αίματος, ηλεκτροκαρδιογράφημα (ΗΚΓ), υπερηχογράφημα καρδιάς, μέτρηση της αρτηριακής πίεσης του αίματος καθώς και εξέταση στεφανιογραφίας. Οι περισσότερες μετρήσεις που προκύπτουν από τις ιατρικές εξετάσεις είναι μετρήσεις διαστήματος. Η καταγραφή τους όμως και η καταχώρησής τους σε ιατρικά δελτία (βλ. Παράρτημα Δ) έγινε από τους θεράποντες ιατρούς οι οποίοι και μετέτρεψαν τις μετρήσεις αυτές σε δυαδικές αντιστοιχώντας την τιμή 0 στην απουσία χαρακτηριστικού και την τιμή 1 στην παρουσία χαρακτηριστικού. Κατεγράφησαν επίσης οι τιμές ορισμένων δεικτών και χαρακτηριστικών τα οποία, κατά την άποψη των ιατρών, περιγράφουν την κατάσταση των ατόμων που εξετάστηκαν. Κάθε δείκτης και κάθε χαρακτηριστικό ορίστηκε ως μεταβλητή. Οι μεταβλητές αυτές, που είναι ποσοτικές και ποιοτικές, χωρίστηκαν σε δύο ομάδες Α και Β.

Η ομάδα Α περιλαμβάνει τις μεταβλητές, που σύμφωνα με την ιατρική, ενοχοποιούνται σημαντικά για την εμφάνιση της νόσου και ονομάζονται "*παράγοντες κινδύνου*", καθώς και την μεταβλητή EF η οποία αποτελεί μέτρο της συσταλτικότητας του μυοκαρδίου.

Η ομάδα Β περιέχει τις μεταβλητές που προσδιορίζουν τη θέση και το ποσοστό έμφραξης των αρτηριών. Οι δύο ομάδες περιλαμβάνουν συνολικά 27 μεταβλητές, που περιγράφονται στη συνέχεια. Τα φυσιολογικά όρια των μεταβλητών αυτών, όπου αυτά υπάρχουν, καθορίστηκαν από τους καρδιολόγους ιατρούς του Ιπποκράτειου Νοσοκομείου.

ΟΜΑΔΑ Α

Οι μεταβλητές που ανήκουν στην ομάδα Α είναι δυαδικές μεταβλητές πλην της ηλικίας (AGE) και του κλάσματος εξώθησης (EF) που είναι συνεχείς.

1) Κλάσμα εξώθησης (EF)

Συνεχής μεταβλητή με παρατηρηθείσες τιμές στο διάστημα (0, 0.845). Αποτελεί μέτρο της συσταλτικότητας του μυοκαρδίου με μέγιστη θεωρητική τιμή 1 (100%). Το κλάσμα εξώθησης ορίζεται ως ο λόγος του όγκου του αίματος που εξωθείται από μια κοιλία σε μία συστολή προς τον όγκο του αίματος που εισέρχεται κατά τη διαστολή. Η τιμή του κλάσματος εξώθησης προσδιορίζεται με στεφανιογραφία, αλλά και με μη επεμβατική μέθοδο όπως ο υπέρηχος καρδιάς.

2) Φύλο (SEX)

Δυαδική μεταβλητή (0-1) που ορίζεται ως:

$$(SEX)_i = \begin{cases} 0 \text{ αν το ιοστό άτομο είναι άνδρας} \\ 1 \text{ αν το ιοστό άτομο είναι γυναίκα} \end{cases}$$

3) Ηλικία (AGE)

Συνεχής μεταβλητή, η οποία δηλώνει την ηλικία των ατόμων που εξετάστηκαν και κυμαίνεται από 33 ως και 75 έτη.

4) Λιπίδια (LIPIDS)

Η μεταβλητή λιπίδια έχει καταχωρηθεί ως δυαδική και ορίζεται ως:

$$(LIPIDS)_i = \begin{cases} 0 \text{ αν το ιοστό άτομο είναι μη υπερλιπιδαιμικό} \\ 1 \text{ αν το ιοστό άτομο είναι υπερλιπιδαιμικό} \end{cases}$$

Η υπερλιπιδαιμία προσδιορίζεται από τις εξής παραμέτρους (οι φυσιολογικές τιμές σε mg/dl)

Ολική χοληστερόλη (TC)	<200
Τριγλυκερίδια (TGL)	<160
HDL-Χοληστερόλη (HDL)	> 35 (για άνδρες)
	> 45 (για γυναίκες)
LDL-Χοληστερόλη (LDL)	<160

Όπου LDL = TC – HDL – 0,2TGL.

Με βάση τις τιμές της LDL και της HDL ένας άνδρας χαρακτηρίζεται υπερλιπιδαιμικός αν LDL > 160mg/dl και HDL< 35mg/dl ενώ μια γυναίκα χαρακτηρίζεται ως υπερλιπιδαιμική αν LDL > 160mg/dl και HDL< 45mg/dl. Ο τρόπος προσδιορισμού της υπερλιπιδαιμίας καθιστά τη μεταβλητή λιπίδια, σύνθετη μεταβλητή.

5) Αρτηριακή πίεση (HBP)

Με τον όρο "*αρτηριακή πίεση*" ορίζουμε την πίεση που έχει το αίμα όταν ρέει μέσα από τις αρτηρίες. Η πίεση αυτή υφίσταται κυκλικές μεταβολές που ακολουθούν τις φάσεις της λειτουργίας της καρδιάς. Η *συστολική πίεση* εμφανίζεται στη διάρκεια της φάσης συστολής της καρδιάς. Σχετίζεται με τη συστολική πλήρωση της αριστερής κοιλίας, με την ταχύτητα εξόδου και τη διασταλτικότητα της αορτής. Η *διαστολική πίεση* εμφανίζεται στο τέλος της φάσης ανάπαυσης (διαστολής) της καρδιάς και οφείλεται στο γεγονός ότι τα αγγεία παραμένουν πλήρη από αίμα. Είναι ένας δείκτης της κατάστασης συστολής και διαστολής των περιφερειακών αγγείων.

Σύμφωνα με την Παγκόσμια Οργάνωση Υγείας (ΠΟΥ) οι τιμές κανονικής πίεσης για τους ενήλικες είναι για μεν τη συστολική ≤ 140mmHg για δε τη διαστολική ≤ 90mmHg.

Επομένως ένα άτομο θεωρείται υπερτασικό όταν έχει συστολική πίεση μεγαλύτερη από 140mmHg ή διαστολική μεγαλύτερη από 90mmHg, ή και τα δύο. Σύμφωνα με τα παραπάνω η μεταβλητή HBP έχει καταχωρηθεί ως δυαδική , δηλαδή

$$(HBP)_i = \begin{cases} 0 \text{ αν το ιοστό άτομο δεν είναι υπερτασικό} \\ 1 \text{ αν το ιοστό άτομο είναι υπερτασικό} \end{cases}$$

6) Κάπνισμα (SMOKE)

Ένα άτομο θεωρείται μη καπνιστής αν δεν έχει καπνίσει ποτέ ή αν έχει σταματήσει το κάπνισμα για διάστημα μεγαλύτερο του ενός έτους πριν την εμφάνιση των συμπτωμάτων της στεφανιαίας νόσου. Η μεταβλητή "SMOKE" καταχωρήθηκε ως δυαδική μεταβλητή (0,1), και ορίστηκε ως εξής:

$$(SMOKE)_i = \begin{cases} 0 \text{ αν το ιοστό άτομο δεν είναι καπνιστής} \\ 1 \text{ αν το ιοστό άτομο είναι καπνιστής} \end{cases}$$

7) Διαβήτης (DIABETES)

Ο σακχαρώδης διαβήτης ορίζεται ως μια παθολογική κατάσταση, που χαρακτηρίζεται από πολλαπλές μεταβολικές αλλοιώσεις. Οι αλλοιώσεις αυτές είναι περισσότερο εμφανείς στον κύκλο της γλυκόζης, συμπεριλαμβάνουν όμως και τον μεταβολισμό λευκωμάτων και λιπών, έτσι ώστε να εμφανίζεται μια ανεπαρκής δράση της ινσουλίνης.

Η φυσιολογική τιμή της γλυκόζης του αίματος κυμαίνεται από 60 ως 115mg/dl. Η μεταβλητή "DIABETES" καταχωρήθηκε ως δυαδική και ορίζεται ως

$$(DIABETES) \begin{cases} 0 \text{ αν το ιοστό άτομο έχει τιμή γλυκόζης μέσα στα φυσιολογικά όρια} \\ 1 \text{ αν το ιοστό άτομο έχει τιμή γλυκόζης έξω από τα φυσιολογικά όρια} \end{cases}$$

8) Παχυσαρκία (OBESITY)

Το άτομο χαρακτηρίζεται ως μη παχύσαρκο όταν ο δείκτης μάζας σώματος είναι $\leq 27 kg/m^2$. Ως δείκτης μάζας σώματος ορίζεται το πηλίκο της μάζας του συγκεκριμένου ατόμου προς το τετράγωνο του ύψους του (σε m^2).

Η δυαδική μεταβλητή OBESITY ορίζεται ως

$$(OBESITY)_i = \begin{cases} 0 \text{ αν το ιοστό άτομο δεν είναι παχύσαρκο} \\ 1 \text{ αν το ιοστό άτομο είναι παχύσαρκο} \end{cases}$$

9) Οικογενειακό ιστορικό (FAMILY)

Ως ύπαρξη οικογενειακού ιστορικού ορίζεται η περίπτωση κατά την οποία ένας τουλάχιστον από τους γονείς του ατόμου, ή ένας τουλάχιστον συγγενής ηλικίας κάτω των 60 ετών, πάσχει από στεφανιαία νόσο ή έχει πεθάνει από αυτήν. (Tzung-Dau Wang *et al* (1998)).

Η μεταβλητή "FAMILY" καταχωρήθηκε ως δυαδική που ορίζεται ως

$$(\text{FAMILY})_i = \begin{cases} 0 \text{ αν το ιοστό άτομο δεν έχει οικογενειακό ιστορικό} \\ 1 \text{ αν το ιοστό άτομο έχει οικογενειακό ιστορικό} \end{cases}$$

10) Τύπος Α (TYPE A)

Το 1960 προτάθηκε η άποψη ότι υπάρχει προσωπικότητα επιρρεπής στη στεφανιαία νόσο (AMA (1991)). Αυτή η προσωπικότητα, που χαρακτηρίζεται από επιθετικότητα και ανυπομονησία, ονομάστηκε "τύπος Α".

Καταχωρήθηκε ως δυαδική μεταβλητή που ορίζεται ως εξής:

$$(\text{TYPE A})_i = \begin{cases} 0 \text{ αν το ιοστό άτομο δεν έχει τα χαρακτηριστκά του τύπου Α} \\ 1 \text{ αν το ιοστό άτομο έχει τα χαρακτηριστικά του τύπου ΑΑ} \end{cases}$$

11) Καθιστική ζωή(SEDENTARY)

Έχει πλέον αποδειχθεί ότι ο τρόπος ζωής συμβάλλει στην εμφάνιση ή μη της στεφανιαίας νόσου. Η σωματική άσκηση αναπτύσσει τον καρδιακό μυ και τον κάνει πιο δυνατό, ικανό να αντλεί τον ίδιο όγκο αίματος με λιγότερους παλμούς. Στην ανάπαυση το μυοκάρδιο αντλεί 10-12 λίτρα/min, ενώ κατά τη σωματική άσκηση η αντλητική ικανότητα φτάνει τα 18-70 λίτρα/min. Η μεταβλητή "SEDENTARY" καταχωρήθηκε ως δυαδική, που ορίζεται ως εξής:

$$(\text{SEDENTARY}) = \begin{cases} 0 \text{ αν το ιοστό άτομο ασκείται και δεν κάνει καθιστική ζωή} \\ 1 \text{ αν το ιοστό άτομο δεν ασκείται και κάνει καθιστική ζωή} \end{cases}$$

ΟΜΑΔΑ Β

Στη δεύτερη αυτή ομάδα καταχωρήθηκαν 15 συνεχείς μεταβλητές οι οποίες εντοπίζουν τη φραγμένη αρτηρία και προσδιορίζουν τη θέση και το ποσοστό έμφραξής της. Οι τιμές του ποσοστού έμφραξης κυμαίνονται από 0% έως 100%. Στο σημείο αυτό πρέπει να τονιστεί η διχογνωμία που υπάρχει μεταξύ των ιατρών σχετικά με το ποσοστό έμφραξης, που ορίζει μια αρτηρία ως φραγμένη. Μια σειρά από μελέτες (Kramer et al (1986), Pattillo et al (1996), Chao et al (1996)), αναφέρουν ως καθοριστικό ποσοστό το 50%, ενώ άλλες,

πιο πρόσφατες, (Tzung-Dau Wang et al (1998), Krishnaswami et al (1996), Kasaoka et al (1997)) θεωρούν ως καθοριστικό το 70%. Το ποσοστό αυτό έχει χρησιμοποιηθεί ως κριτήριο φραγμένης αρτηρίας και στην παρούσα εργασία.

Οι μεταβλητές της ομάδας Β είναι οι ακόλουθες:

1) Δεξιά στεφανιαία αρτηρία (RCA)

Τέσσερις μεταβλητές PROX, MID, DIST και PD προσδιορίζουν τη θέση της στένωσης (Σχ. 1.3)

2) Πρόσθιος κατιών (LAD)

Με μεταβλητές PROX, MID, 1-D, APIC και 2-D, που προσδιορίζουν τη θέση της στένωσης (Σχ. 1.3).

3) Περισπώμενη αρτηρία(LCX)

Με μεταβλητές τις PROX, DIST, OM, PL και PD (Σχ.1.3).

4) Στέλεχος (LMCA)

$$(LMCA)_i = \begin{cases} 0 \text{ αν το ιοστό άτομο έχει ποσοστό έμφραξης της LMCA} < 70\% \\ 1 \text{ αν το ιοστό άτομο έχει ποσοστό έμφραξης της LMCA} \geq 70\% \end{cases}$$

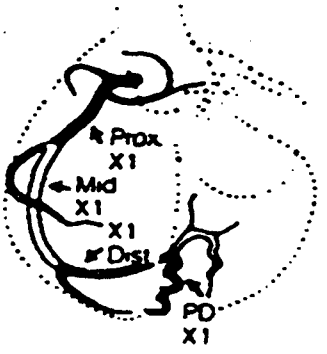
5) Αριθμός φραγμένων αρτηριών (VD)

Κατεγράφησαν ως φραγμένες αρτηρίες αυτές που εμφάνισαν στένωση ≥70% σε ένα τουλάχιστον τμήμα τους.

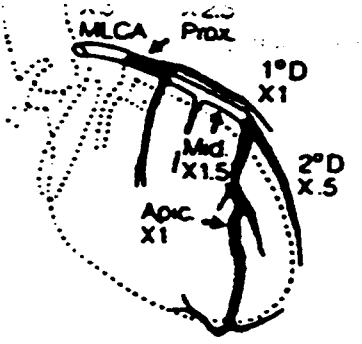
Η μεταβλητή καταχωρήθηκε ως κατηγορική με επίπεδα 0, 1, 2, 3 τα οποία αναφέρονται στον αντίστοιχο αριθμό φραγμένων αρτηριών.

Πιο συγκεκριμένα:

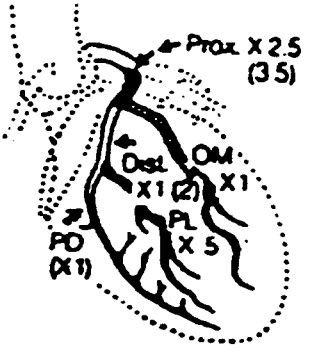
(VD) =κ, όταν το ιοστό άτομο έχει κ φραγμένες αρτηρίες, όπου κ=0,1,2,3.



(α) RCA



(β) LAD



(γ)LCX

Σχ.1.3

Όπως έχει ήδη αναφερθεί, οι μεταβλητές φύλο (SEX), ηλικία (AGE), λιπίδια (LIPIDS), αρτηριακή πίεση (HBP), κάπνισμα (SMOKE), διαβήτης (DIABETES), παχυσαρκία (OBESITY), οικογενειακό ιστορικό (FAMILY), τύπος Α (TYPE A) και καθιστική ζωή (SEDENTARY) αποτελούν παράγοντες κινδύνου (AMA 1991).

Στην ίδια ομάδα μεταβλητών που χρησιμοποιείται για την περαιτέρω ανάλυση περιέχεται και η μεταβλητή κλάσμα εξώθησης (EF) της οποίας οι τιμές συνήθως προσδιορίζονται με στεφανιογραφία. Στην παρούσα εργασία οι τιμές του κλάσματος εξώθησης προκύπτουν από υπερηχογράφημα και θεωρούνται ακριβείς και αποδεκτές από τους ιατρούς.

Από ερευνητικές μελέτες (Kostuk *et al.* (1973)) σε ασθενείς με στεφανιαία νόσο, έχει διαπιστωθεί ότι το κλάσμα εξώθησης (EF) της αριστερής κοιλίας ξεχωρίζει με ακρίβεια τους ασθενείς εκείνους που διατρέχουν αυξημένο κίνδυνο επιπλοκών.

Όσο μεγαλύτερο το κλάσμα εξώθησης τόσο μικρότερος είναι ο κίνδυνος επιπλοκών και θανάτου και αντίστροφα. Η μελέτη του Kostuk έδειξε ότι στους ασθενείς στους οποίους το κλάσμα εξώθησης ήταν μικρότερο του 0.25, η θνησιμότητα τον πρώτο μήνα μετά το έμφραγμα ανερχόταν σε 58% και σε εκείνους στους οποίους το κλάσμα εξώθησης ήταν μεγαλύτερο του 0.25 η θνησιμότητα στο ίδιο διάστημα ήταν μόλις 8%. Αξίζει να σημειωθεί το γεγονός ότι όλοι οι ασθενείς που εμφάνισαν κλάσμα εξώθησης μεγαλύτερο του 0.40 επέζησαν. Θεωρώντας λοιπόν ως κρίσιμη τιμή του κλάσματος εξώθησης την 0.40, ορίστηκε μια δυαδική μεταβλητή (EFGT), με την ακόλουθη κατανομή συχνοτήτων.

ΠΙΝΑΚΑΣ 1. 1: Κατανομή συχνοτήτων της μεταβλητής "EFGT"

EF	EFGT	(f_i)	(f_i) %
[0.00-0.40]	0	168	25.5
(0.40-	1	492	74.5

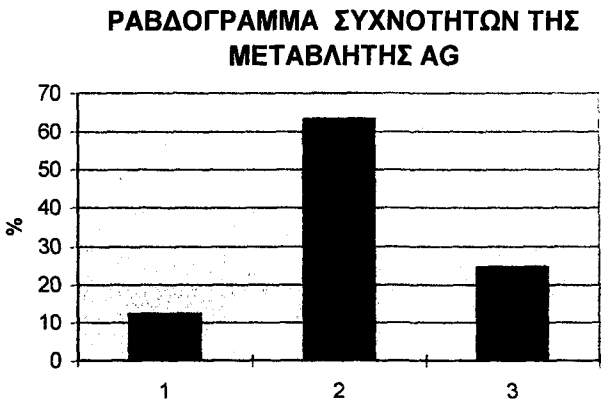
Η μεταβλητή (SEX) αντιπροσωπεύεται στο συνολικό δείγμα από 430 άνδρες (SEX=0) και 230 γυναίκες (SEX=1)

Τα στατιστικά της συνεχούς μεταβλητής (AGE) παρατίθεται στον πίνακα Π.1 του παραρτήματος B_2 .

Για να μελετηθεί η εκδήλωση της νόσου σε σχέση με τους παράγοντες κινδύνου και την ηλικία στο συνολικό αρχείο, ορίστηκε η κατηγορική μεταβλητή ("AG"), που καταχωρεί τα άτομα του δείγματος σε τρεις ομάδες ηλικιών, όπως φαίνεται στον πίνακα 1.2.

ΠΙΝΑΚΑΣ 1.2: Κατανομή συχνοτήτων της μεταβλητής "AG"

ΗΛΙΚΙΑ	AG	(f_i)	(f_i) %
(30-45]	1	81	12.3
(45-60]	2	417	63.2
(60-75]	3	162	24.5



Σχ. 1.4

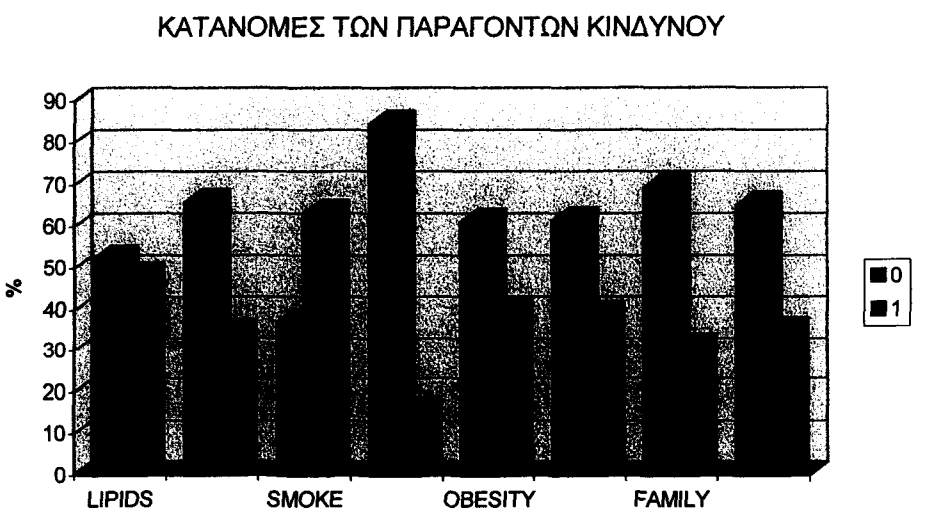
Παρατηρούμε ότι το 63% των ατόμων, που προσήλθαν στο Ιπποκράτειο Νοσοκομείο με συμπτώματα εμφάνισης στεφανιαίας νόσου, έχουν ηλικία από 45 έως 60 έτη.

Οι συχνότητες των παραγόντων κινδύνου που εμφανίζονται στο δείγμα των 660 ατόμων είναι οι ακόλουθες:

ΠΙΝΑΚΑΣ 1.3: Πίνακας κατανομής συχνοτήτων των παραγόντων κινδύνου

Παράγοντες κινδύνου	**0 (απουσία χαρακτηριστικού)**	**1 (παρουσία χαρακτηριστικού)**
LIPIDS	344 (52,1%)	316 (47,9%)
HBR	433 (65,6%)	227 (34,4%)
SMOKE	244 (37,0%)	416 (63,0%)
DIABETES	558 (84,5%)	102 (15,5%)
OBESITY	402 (60,9%)	258 (39,1%)
SEDENTARY	404 (61,2 %)	256 (38,8%)
FAMILY	456 (69,1 %)	204 (30,9%)
TYPE A	429 (65,0%)	231 (35,0%)

Το αντίστοιχο ραβδόγραμμα του πίνακα 1.3 φαίνεται στο σχήμα 1.5



Σχ.1.5

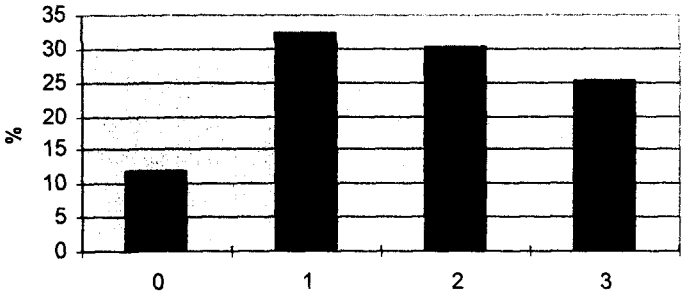
Αξίζει να σημειωθεί ο μεγάλος αριθμός καπνιστών στο δείγμα (63%), καθώς και ο μικρός αριθμός αυτών που πάσχουν από διαβήτη (15,5%).

Οι κατανομές συχνοτήτων των μεταβλητών της ομάδας Β περιέχονται στους πίνακες Π.2-Π.8 του παραρτήματος B_2. Ο πίνακας 1.4 παρουσιάζει την κατανομή συχνοτήτων της μεταβλητής VD (αριθμός φραγμένων αρτηριών), με καθοριστική τιμή έμφραξης το 70%. Το αντίστοιχο ραβδόγραμμα φαίνεται στο Σχ.1.6:

ΠΙΝΑΚΑΣ 1.4: Κατανομή συχνοτήτων της μεταβλητής "VD"

VD	(f_i)	%
0	80	12,1%
1	214	32,4%
2	200	30,3%
3	166	25,2%

ΡΑΒΔΟΓΡΑΜΜΑ ΣΥΧΝΟΤΗΤΩΝ ΤΗΣ ΜΕΤΑΒΛΗΤΗΣ VD



Σχ.1.6

Όπως προκύπτει από τη μελέτη του παραπάνω πίνακα, το μεγαλύτερο ποσοστό αυτών που εξετάστηκαν για πιθανή εμφάνιση στεφανιαίας νόσου, έχουν μια φραγμένη αρτηρία (32,4%) ενώ το 12.1% είναι ελεύθεροι νόσου. Είναι σημαντικό ότι το ποσοστό των ατόμων που εμφανίζουν τουλάχιστον μια φραγμένη αρτηρία και χαρακτηρίζονται σαν πάσχοντα από στεφανιαία νόσο ανέρχεται στο 87,9%.

Η κατάσταση αυτή περιγράφεται από μια δυαδική μεταβλητή, που ονομάστηκε "DCODE", και παίρνει την τιμή 0 για τα ελεύθερα νόσου άτομα και την τιμή 1 για τα άτομα που εμφανίζουν τη νόσο.

ΠΙΝΑΚΑΣ 1.5: Κατανομή συχνοτήτων της μεταβλητής "DCODE"

DCODE	(fi)	%
0	80	12,1%
1	580	87,9%

Παρατηρήθηκε επίσης από τους πίνακες Π2, Π4, Π6 του παραρτήματος Γ, ότι το μεγαλύτερο ποσοστό των ασθενών ανεξαρτήτως φύλλου εμφανίζει έμφραξη στη θέση Proximal των τριών αρτηριών. Ιδιαίτερα αυξημένο εμφανίζεται το ποσοστό των ασθενών στην αρτηρία LAD στη θέση Proximal

το οποίο ανέρχεται σε 39,5%, έναντι του ποσοστού της αρτηρίας RCA (27,9%) και του ποσοστού της αρτηρίας LCX (25,3%).

Η παρατήρηση αυτή μας οδήγησε στον έλεγχο της υπόθεσης ότι: το ποσοστό των ασθενών που εμφανίζουν έμφραξη στη θέση PROXIMAL είναι το ίδιο και για τις τρεις αρτηρίες. Η τιμή του χ^2-test =27,4 σε επίπεδο σημαντικότητας 0,000 οδηγεί στην απόρριψη της υπόθεσης περί ίσων ποσοστών, καθώς και η τιμή του Levene-test =53,8 σε επίπεδο σημαντικότητας 0,000 οδηγεί στην απόρριψη της υπόθεσης περί ίσων διακυμάνσεων μεταξύ των ομάδων.

Άρα θα μπορούσαμε να ισχυριστούμε ότι η αρτηρία που εμφανίζεται περισσότερο επιβαρυμένη στη θέση Proximal (PROX) είναι ο πρόσθιος κατιών (LAD). Η παρατήρηση αυτή επιβεβαιώθηκε με το test Dunnett T3 (έλεγχος πολλαπλών συγκρίσεων σε ομάδες με άνισες διακυμάνσεις), με το οποίο ελέγχθηκε η υπόθεση ότι «οι αρτηρίες LAD, RCA και LCX εμφανίζουν ίσα ποσοστά έμφραξης στη συγκεκριμένη θέση PROX». Τα αποτελέσματα του test είναι τα εξής:

		Διαφορά μέσου (%)	Τυπικό σφάλμα	Επίπεδο σημαντικότητας
LAD	RCA	12,27	0,025	0,000
	LCX	13,64	0,025	0,000
RCA	LAD	-12,27	0,025	0,000
	LCX	1,3636	0,025	0,923
LCX	LAD	-13,64	0,025	0,000
	RCA	-1,3636	0,025	0,923

(*) Το επίπεδο σημαντικότητας στη διαφορά μέσου καθορίζεται στο 0,05.

Από τα επίπεδα σημαντικότητας του test Dunnett T3 που καταγράφονται στον παραπάνω πίνακα επιβεβαιώνεται η μαρτυρία ότι το ποσοστό των ασθενών που εμφανίζουν έμφραξη στη θέση PROXIMAL της αρτηρίας LAD είναι πράγματι αυξημένο έναντι των ποσοστών των ασθενών που εμφανίζουν έμφραξη στην ίδια θέση αλλά στις αρτηρίες RCA και LCX.

1.2 Συσχέτιση παραγόντων κινδύνου με την εκδήλωση νόσου.

Στο δείγμα των 660 ατόμων παρατηρήθηκε ότι ελεύθεροι νόσου είναι 14 άνδρες (2,12%) και 66 γυναίκες (10%) ενώ πάσχουν από στεφανιαία νόσο 416 άνδρες (63,03%) και 164 γυναίκες (24,84%).

Από τον πίνακα 1.6 μπορούμε να υπολογίσουμε το λόγο υπεροχής του αριθμού των ασθενών ανδρών προς τον αντίστοιχο των γυναικών [Παπαϊωάννου Τ. (1999)]

$\text{Odds}_1 = \frac{\alpha}{c} = \frac{416}{164} = 2{,}536$, καθώς και το λόγο του αριθμού των υγιών ανδρών προς

τον αντίστοιχο των υγιών γυναικών , $\text{Odds}_0 = \frac{b}{d} = \frac{14}{66} = 0{,}212$.

ΠΙΝΑΚΑΣ 1.6

DCODE / SEX	1 (ασθενής)	0 (υγιής)	
0 (άνδρας)	416 (α)	14 (b)	430
1 (γυναίκα)	164 (c)	66 (d)	230
Σύνολο	580 (α + c)	80 (b + d)	660

Άρα ο λόγος υπεροχής (OR) για άνδρες στεφανιαίους ασθενείς ορίζεται ως

$$\text{Odds Ratio (OR)} = \frac{\text{Odds}_1}{\text{Odds}_0}$$

και ισούται με $\frac{2,536}{0,212}=11,962$, τιμή που δείχνει ότι ο στεφανιαίος ασθενής είναι 11,96 φορές πιθανότερο να είναι άνδρας.

Ένα 95% διάστημα εμπιστοσύνης για το λογάριθμο του λόγου υπεροχής δίνεται από τη σχέση:

[logOR – 1,96.S.E (logOR), log OR + 1,96.S ·E(logOR)] (Campbell and Machin (1991)):

όπου $$S \cdot E(\log OR)=\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

Αντικαθιστώντας τις τιμές των α,b,c,d του Πίνακα 1.6 στις παραπάνω σχέσεις, προκύπτει ότι: SE(logOR)=0,308. Άρα το διάστημα εμπιστοσύνης για το λογάριθμο είναι:

$$[\log(11,962)-1,96\cdot 0,3082,\quad \log(11,962)+1,96\cdot 0,3082] =$$
$$= (2,48 - 0,604\ ,\ 2,48 + 0,604) = (1.876\ ,\ 3.084)$$

από όπου έπεται ότι ένα 95% διάστημα εμπιστοσύνης για το λόγο υπεροχής OR είναι το (6,527 , 21,846).

Στη συνέχεια ελέγχθηκε η εξάρτηση της εκδήλωσης της νόσου, της τιμής του κλάσματος εξώθησης, του αριθμού των φραγμένων αρτηριών από τους παράγοντες κινδύνου όπως επίσης και η μεταξύ των παραγόντων κινδύνου πιθανή ύπαρξη εξάρτησης. Ο έλεγχος έγινε με το χ^2-τέστ. Οι τιμές του επίπεδου σημαντικότητας(α) καθώς και οι τιμές του επιπέδου σημαντικότητας που προκύπτουν από τη διόρθωση κατά Yates (όταν η μεταβλητή είναι δυαδική) και δείχνουν αλληλεξάρτηση των μεταβλητών παρουσιάζονται στον πίνακα 1.7. και 1.8. Κρίσιμη τιμή του επιπέδου σημαντικότητας θεωρείται η τιμή α=0,05.

ΠΙΝΑΚΑΣ 1.7

ΤΙΜΕΣ ΤΟΥ ΕΠΙΠΕΔΟΥ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ (α) ΤΩΝ ΠΑΡΑΓΟΝΤΩΝ ΚΙΝΔΥΝΟΥ
ΤΙΜΕΣ ΤΟΥ ΕΠΙΠΕΔΟΥ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΜΕ ΔΙΟΡΘΩΣΗ ΚΑΤΑ YATES

	HBP	SMOKE	DIABETES	OBESITY	FAMILY	SEDENTARY	TYPE A	SEX
AG (α)	0,0005	0,0000			0,0074	0,0042	0,0000	0,0000
LIPIDS (α) Yates			0,0022 0,0032				0,0001 0,0001	
HBP (α) Yates		0,0000 0,0000		0,0037 0,0048		0,0181 0.0227	0,0001 0,0001	0,0000 0,0000
SMOKE (α) Yates						0,0000 0,0000	0,0000 0,0000	0,0000 0,0000
DIABETES(α) Yates				0,0254 0,0336				0,0004 0,0007
FAMILY (α) Yates								0,0025 0,0033
SEDENTARY(Yates							0,0000 0,0000	0,0000 0,0000
SEX (α) Yates							0,0000 0,0000	

ΠΙΝΑΚΑΣ 1.8

ΤΙΜΕΣ ΤΟΥ ΕΠΙΠΕΔΟΥ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ (α) ΤΙΜΕΣ ΤΟΥ ΕΠΙΠΕΔΟΥ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΜΕ ΔΙΟΡΘΩΣΗ ΚΑΤΑ YATES							
	LIPIDS	SMOKE	DIABETES	FAMILY	SEDENTARY	TYPE A	SEX
VD (α)		0,0000	0,015				0,0000
EFTG (α)		0,009	0,0000				
Yates		0,0118	0,0000				
DCODE (α)	0,0139	0,0000		0,001	0,0000	0,0000	0,0000
Yates	0,0192	0,0000		0,0016	0,0000	0,0000	0,0000

Τα συμπεράσματα που προκύπτουν από τους πίνακες 1.7 και 1.8 είναι τα ακόλουθα:

- Η εκδήλωση της νόσου στο συνολικό δείγμα ανδρών και γυναικών εμφανίζει συσχέτιση με την υπερλιπιδαιμία (α=0,0139), το κάπνισμα (α=0,000), το οικογενειακό ιστορικό (α=0,001), την καθιστική ζωή (α=0,000), όπως και με τον τύπο προσωπικότητας του ατόμου (α=0,000). Επιπλέον επηρεάζει την τιμή του κλάσματος εξώθησης (α=0.000).
- Υπάρχει εξάρτηση του αριθμού των φραγμένων αρτηριών από το φύλο του ασθενούς (α=0,000), από το κάπνισμα (α=0,000) και από το διαβήτη του ασθενούς (0,015). Ο αριθμός των φραγμένων αρτηριών εμφανίζεται επίσης εξαρτημένος από την τιμή του κλάσματος εξώθησης (α=0,000).
- Το κλάσμα εξώθησης εμφανίζεται εξαρτημένο από το κάπνισμα (α=0,009) και από το διαβήτη (α=0,000).
- Η αλληλεξάρτηση των παραγόντων κινδύνου παρουσιάζεται στον πίνακα 1.7. Παρατηρείται επίσης ότι το φύλο είναι εξαρτημένο από τους περισσότερους παράγοντες κινδύνου.

Τα κενά που υπάρχουν στον πίνακα 1.6 οφείλονται στο ότι οι αντίστοιχες τιμές του επιπέδου σημαντικότητας δείχνουν μη σημαντική στατιστική εξάρτηση των μεταβλητών.

Στις επόμενες παραγράφους επιχειρείται η εύρεση των εξαρτήσεων των παραγόντων κινδύνου κατά φύλο και η εντόπιση εκείνων των παραγόντων κινδύνου που ευνοούν την εμφάνιση της στεφανιαίας νόσου. Για το σκοπό

αυτό το αρχικό δείγμα των 660 ατόμων χωρίστηκε σε αρχείο ανδρών και γυναικών.

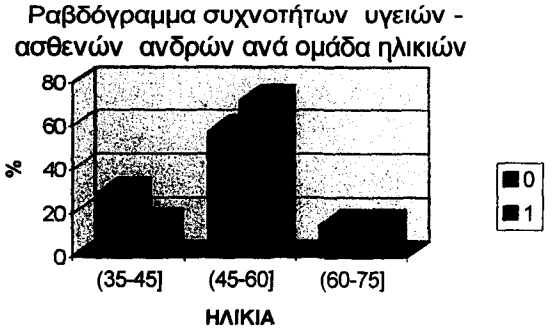
1.3 Ανδρικός πληθυσμός και παράγοντες κινδύνου

Στο αρχείο των ανδρών, το ποσοστό των υγιών ανέρχεται σε (3,2%) με αντίστοιχο ποσοστό ασθενών στο 96,8 %.

Στον πίνακα 1.9 που ακολουθεί, παρουσιάζεται η κατανομή συχνοτήτων της μεταβλητής DCODE (υγιείς=0, ασθενείς=1) ανά ομάδα ηλικιών ενώ το αντίστοιχο ραβδόγραμμα φαίνεται στο Σχ.1.7.

ΠΙΝΑΚΑΣ 1.9

ΗΛΙΚΙΑ	(30-45]	(45-60]	(60-75]	ΣΥΝΟΛΟ
DCODE =0	4 (28,6%)	8 (57.1%)	2 (14,3%)	14 (100.0%)
DCODE =1	61(14,7%)	296(71,2%)	59(14,2%)	416 (100,0%)



Σχ.1.7

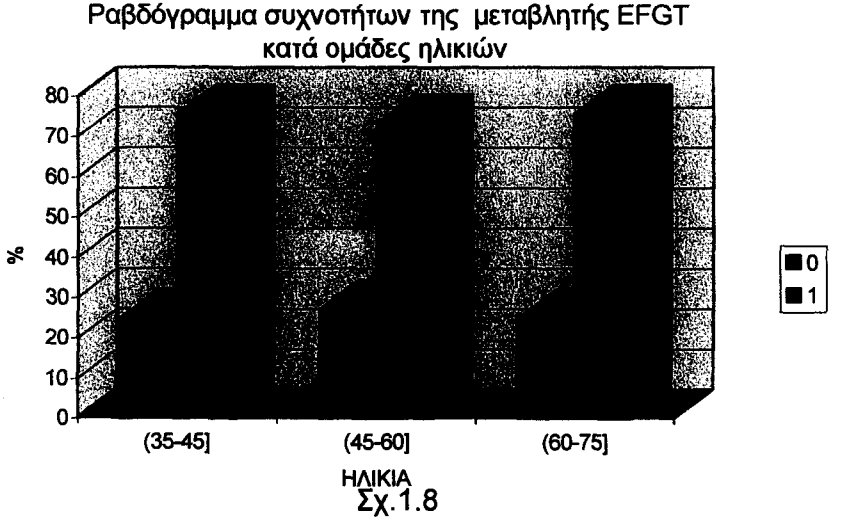
Είναι εμφανές ότι το μεγαλύτερο ποσοστό ανδρών με στεφανιαία νόσο (71,2%) βρίσκεται στο διάστημα ηλικιών (45-60] ετών ενώ στις άλλες δύο κατηγορίες ηλικιών το ποσοστό αυτό μειώνεται σημαντικά.

Θεωρώντας ως τιμή διαχωρισμού της μεταβλητής "EF" την 0.4, κατασκευάστηκε η μεταβλητή EFGT (EFGT =0 αν EF≤ 0,4 και EFGT =1 αν EF> 0,4).

Η κατανομή συχνοτήτων της μεταβλητής αυτής στις τρεις κατηγορίες ηλικιών φαίνεται στον πίνακα 1.10 και παριστάνεται στο ραβδόγραμμα του σχήματος 1.9.

ΠΙΝΑΚΑΣ 1.10

EFGT \ AG	(30-45]	(45-60]	(60-75]	ΣΥΝΟΛΟ
0	16 (24,6%)	82 (26,9%)	15 (24,6)	113
1	49(75,4%)	222(73,1%)	46 (75,6%)	317
ΣΥΝΟΛΟ	65(100%)	304(100%)	61(100%)	430



Σχ.1.8

Στο συγκεκριμένο δείγμα, που αποτελεί και το αντικείμενο της μελέτης αυτής, παρατηρήθηκε ότι όλοι οι άνδρες με τιμή κλάσματος εξώθησης (EF) ≤ 0,4 ήταν ασθενείς.

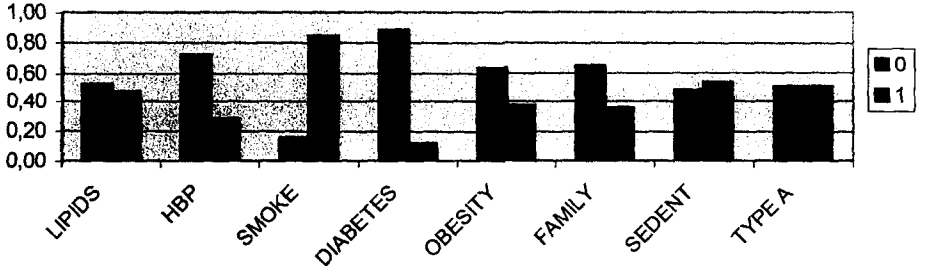
Η κατανομή συχνοτήτων των παραγόντων κινδύνου στους υγιείς – ασθενείς άνδρες φαίνεται στον πίνακα 1.11, που ακολουθεί και αναπαρίσταται στο ραβδόγραμμα το Σχ.1.10, που το συνοδεύει.

ΠΙΝΑΚΑΣ 1.11

Κατανομή των παραγόντων κινδύνου στους άνδρες

ΠΑΡΑΓΟΝΤΕΣ ΚΙΝΔΥΝΟΥ	0	0 (%)	1	1(%)	rank
LIPIDS	227	53	203	47	4
HBP	311	72	119	28	7
SMOKE	67	16	363	84	1
DIABETES	379	88	51	12	8
OBESITY	272	63	158	37	5
FAMILY	280	65	150	35	6
SEDENTARY	203	47	227	53	2
TYPE A	215	50	215	50	3

ΡΑΒΔΟΓΡΑΜΜΑ ΣΥΧΝΟΤΗΤΩΝ ΤΩΝ ΠΑΡΑΓΟΝΤΩΝ ΚΙΝΔΥΝΟΥ ΣΤΟ ΑΡΧΕΙΟ ΑΝΔΡΩΝ



Σχ.1.9

Η παρατήρηση ότι το μεγαλύτερο ποσοστό των ασθενών ανδρών είναι καπνιστές, επιβεβαιώνει προηγούμενες έρευνες (AMA –(1991)) καθώς επίσης και το ρόλο του καπνίσματος ως σημαντικού παράγοντα στην εκδήλωση της στεφανιαίας νόσου. Στον πίνακα 1.12 εμφανίζεται ο αριθμός των καπνιστών σε σχέση με τον αριθμό των υγιών-ασθενών στο συγκεκριμένο δείγμα.

ΠΙΝΑΚΑΣ 1.12

SMOKE / DCODE	0	1	Σύνολο
0	5	9	14
1	62	354	416
Σύνολο	67	363	430

Στον πίνακα 1.13, παρουσιάζονται οι τιμές του επιπέδου σημαντικότητας (α) που προέκυψαν με την εφαρμογή του χ^2 - test κατά τον έλεγχο της εξάρτησης των παραγόντων κινδύνου με την εκδήλωση της νόσου .

ΠΙΝΑΚΑΣ 1.13

DCODE ΠΑΡΑΓΩΝ ΚΙΝΔΥΝΟΥ	χ^2	α
AG	2,109	0,348
LIPIDS	0,110	0,740
HBP	1,666	0,197
SMOKE	**4,489**	***0,035***
DIABETES	1,267	0,260
OBESITY	1,460	0,227
FAMILY	1,153	0,283
SEDENTARY	0,045	0,832
TYPE A	0,295	0,587

Παρατηρούμε ότι το κάπνισμα αναδεικνύεται σε σημαντικό επιβαρυντικό παράγοντα για την εκδήλωση στεφανιαίας νόσου.

Με χ^2 – test έγινε ο έλεγχος για πιθανή ύπαρξη εξάρτησης μεταξύ της μεταβλητής AG (ομάδες ηλικιών) και της μεταβλητής EFGT. Οι τιμές που προέκυψαν είναι: $\chi^2 = 0.258$, βαθμοί ελευθερίας 2, επίπεδο σημαντικότητας $\alpha = 0.873$, γεγονός που δεν παρέχει μαρτυρία υπέρ της άποψης ότι υπάρχει σχέση μεταξύ EF και ηλικίας.

Επειδή το κλάσμα εξώθησης (EFGT) αποτελεί μέτρο της συσταλτικότητας του μυοκαρδίου εμφανίζει εξάρτηση από τον αριθμό των φραγμένων αρτηριών (α= 0.0100). Επίσης η μεταβλητή (EFGT) εμφανίζει εξάρτηση από παράγοντες κινδύνου όπως: DIABETES (α=0.025), FAMILY (α=0.048), TYPE A (α=0.000).

Παρατηρήθηκε ότι για ηλικίες ≤ 45 ετών η εκδήλωση της νόσου παρουσίασε εξάρτηση από το κάπνισμα (χ^2 = 6,826, α=0.009), ενώ για ηλικίες > 45 δεν παρατηρήθηκε εξάρτηση μεταξύ των παραγόντων κινδύνου και της νόσου.

Η παρατήρηση αυτή μπορεί να έχει ως πιθανή εξήγηση το γεγονός ότι ο αριθμός των ανδρών στεφανιαίων ασθενών εμφανίζεται ιδιαίτερα αυξημένος σε ηλικίες μεγαλύτερες των 45 ετών. Είναι πολύ πιθανό τα άτομα αυτά να έχουν διακόψει το κάπνισμα ή να ακολουθούν φαρμακευτική αγωγή οπότε οι τιμές των δεικτών είναι ικανοποιητικές.

Λεπτομερέστερη ανάλυση των επιβαρυντικών παραγόντων στους ασθενείς άνδρες επιχειρείται σε επόμενο κεφάλαιο .

Η συσχέτιση μεταξύ των παραγόντων κινδύνου στο αρχείο ανδρών δίνεται από τον πίνακα 1.14. Ο έλεγχος έγινε με τη χρήση του χ^2 – test. Στον πίνακα παρουσιάζονται οι τιμές του επιπέδου σημαντικότητας (α) και τονίζονται οι τιμές του α που αντιστοιχούν σε μεταβλητές που εμφανίζουν συσχέτιση.

ΠΙΝΑΚΑΣ 1.14: Συσχέτιση των παραγόντων κινδύνου
(τιμές του επιπέδου σημαντικότητας)

ΠΑΡΑΓΩΝ ΚΙΝΔΥΝΟΥ	HBP	SMOKE	DIABETES	OBESITY	FAMILY	SEDENTARY	TYPE A
LIPIDS	0,969	**0,021**	0,0767	0,7499	0,439	0,232	**0,00007**
HBP		0,0549	0,710	**0,0029**	0,4301	0,409	0,105
SMOKE			0,423	**0,0013**	**0,039**	**0,042**	**0,00003**
DIABETES				**0,0015**	0,852	**0,0033**	0,8414
OBESITY					0,8528	**0,0338**	0,8414
FAMILY						0,5184	0,1566
SEDENTAR							0,4989
AG						**0,047**	

1.4. Αρχείο γυναικών

Στο δείγμα των 230 γυναικών, οι 66 (28,7%) είναι υγιείς και οι 164 (71,3%) ασθενείς. Η κατανομή συχνοτήτων της μεταβλητής VD (αριθμός φραγμένων αρτηριών) δίνεται στον πίνακα 1.15:

ΠΙΝΑΚΑΣ 1.15: Αριθμός φραγμένων αρτηριών

VD	Συχνότητα	%
0	66	28,7
1	*75*	*32,6*
2	47	20,4
3	42	18,3
Σύνολο	230	100

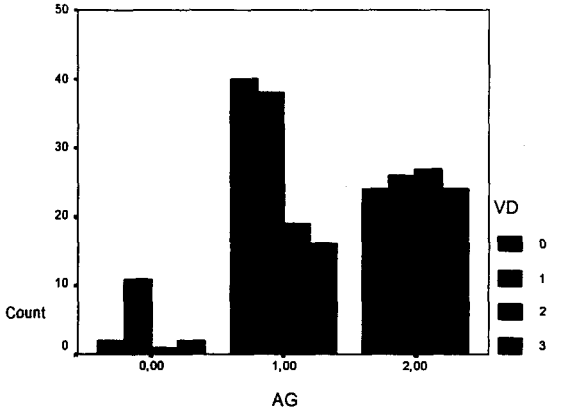
Παρατηρείται αυξημένο το ποσοστό των γυναικών που πάσχουν από νόσο μιας αρτηρίας. Ο αριθμός των φραγμένων αρτηριών (VD) σε σχέση με τις τρεις ομάδες ηλικιών(AG), δίνεται στον πίνακα (1.16) καθώς και στο ραβδόγραμμα του σχήματος 1.10.

ΠΙΝΑΚΑΣ 1.16

AG	VD	0	1	2	3	Σύνολο
0	(30-45]	2	11	1	2	16
1	(45-60]	40	38	19	16	113
2	(60-75]	24	26	27	24	101
	Σύνολο	66	75	47	42	230

Παρατηρείται ότι το μεγαλύτερο ποσοστό των γυναικών που πάσχουν από νόσο μιας αρτηρίας βρίσκεται στην ηλικία 45-60 ετών.

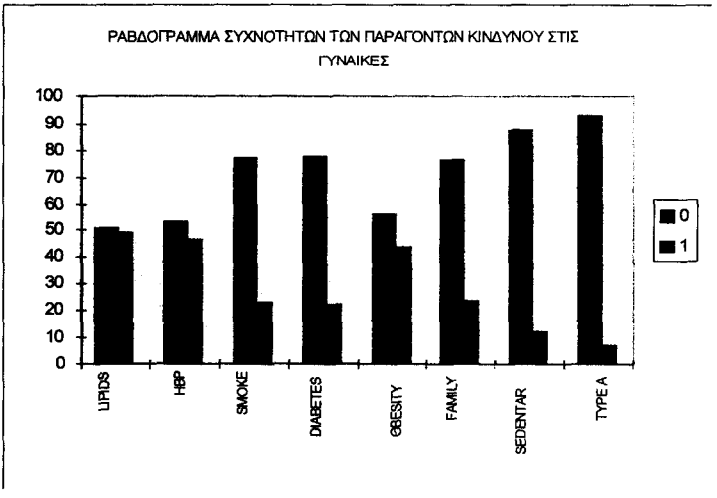
Η συσχέτιση των μεταβλητών (VD) και (AG) ελέγχθηκε με το χ^2 - test του οποίου η τιμή είναι 19,260 με 6 βαθμούς ελευθερίας με αντίστοιχη τιμή του επιπέδου



Σχ 1.10

σημαντικότητας, α=0.004. Συμπεραίνουμε λοιπόν ότι υπάρχει εξάρτηση του αριθμού φραγμένων αρτηριών (VD) με την ηλικία (AG).

Η κατανομή συχνοτήτων των παραγόντων κινδύνου στο αρχείο γυναικών (ασθενών – υγιών) παρουσιάζεται στον πίνακα 1.17 και στο ραβδόγραμμα του σχήματος 1.11, που ακολουθεί:



Σχ.1.11

ΠΙΝΑΚΑΣ 1.17: Κατανομή συχνοτήτων των παραγόντων κινδύνου στις γυναίκες

ΠΑΡΑΓΩΝ ΚΙΝΔΥΝΟΥ	**0**	**0(%)**	**1**	**1(%)**
LIPIDS	117	50.8	113	49,2
HBP	122	53,1	108	46,9
SMOKE	177	76,9	53	23.1
DIABETES	179	77,8	51	22,2
OBESITY	130	56,5	100	43,5
FAMILY	176	76,5	54	23,5
SEDENTARY	201	87,3	29	12,7
TYPE A	214	93	16	7

Στον πίνακα 1.18 επισημαίνονται οι παράγοντες κινδύνου που εμφανίζονται να ευνοούν την εκδήλωση της νόσου και εμφανίζουν εξάρτηση με το κλάσμα εξώθησης και τον αριθμό των φραγμένων αρτηριών. (Η σύγκριση του επιπέδου σημαντικότητας έγινε κι εδώ με α=0.05).

ΠΙΝΑΚΑΣ 1.18: Συσχέτιση των μεταβλητών DCODE, EFTG, VD με τους παράγοντες κινδύνου στις γυναίκες

ΓΥΝΑΙΚΕΣ	**DCODE**		**EFTG**		**VD**	
	x^2	*α*	x^2	*α*	x^2	*α*
LIPIDS	9,24	***0,0023***	0,84	0,35	11,23	***0,01***
HBP	1,36	0,243	0,65	0,79	4,095	0,25
SMOKE	12,48	***0,0004***	5,39	***0,02***	14,1	***0,002***
DIABETES	11,43	***0,0007***	16,16	***0,00***	26,37	***0,00***
OBESITY	0,008	0,9286	1,4	0,22	2,54	0,46
FAMILY	4,99	***0,025***	0,001	0,97	5,59	0,13
SEDENTARY	5,46	***0,019***	0,0009	0,97	5,98	0,11
TYPE A	4,234	***0,039***	3,719	0,053	5,06	0,16
AG	5,733	0,057	7,804	***0,020***	19,260	***0,004***

Παρατηρούμε ότι η εκδήλωση της νόσου στα δύο φύλα, όπως αυτή εμφανίζεται στα συγκεκριμένα αρχεία ανδρών – γυναικών (Πίνακας 1.12 και Πίνακας 1.18), ευνοείται από διαφορετικούς παράγοντες κινδύνου. Έτσι στους άνδρες το κάπνισμα ευνοεί την εμφάνιση της νόσου, ενώ στις γυναίκες εκτός από το κάπνισμα δρουν επιβαρυντικά η υπερλιπιδαιμία, ο διαβήτης, το οικογενειακό ιστορικό, η καθιστική ζωή και η προσωπικότητα του ατόμου. Στις γυναίκες, το κλάσμα εξώθησης εμφανίζεται εξαρτημένο από το κάπνισμα, το διαβήτη και την ηλικία. Στον αριθμό των φραγμένων αρτηριών δρουν επιβαρυντικά η υπερλιπιδαιμία, το κάπνισμα, ο διαβήτης και η ηλικία.

Στον πίνακα 1.19 μελετάται η αλληλεξάρτηση των παραγόντων κινδύνου και επισημαίνονται οι τιμές του επιπέδου σημαντικότητας που δηλώνουν στατιστικώς σημαντική εξάρτηση. Η κρίσιμη τιμή του επιπέδου σημαντικότητας είναι α=0.05.

ΠΙΝΑΚΑΣ 1.19: Πίνακας συσχέτισης μεταξύ των παραγόντων κινδύνου στις γυναίκες (τιμές του επιπέδου σημαντικότητας)

ΓΥΝΑΙΚΕΣ	HBP	SMOKE	DIABETES	OBESITY	FAMILY	SEDENTARY	TYPE A	AG
LIPIDS	0,192	0,763	***0,012***	***0,009***	0,280	0,196	***0,031***	0,259
HBP		0,365	0,331	0,586	0,295	0,582	0,191	0,128
SMOKE			0,109	0,0599	0,869	0,042	***0,000***	***0,022***
DIABETES				0,707	0,992	0,452	***0,000***	0,091
OBESITY					0,159	0,174	0,111	0,105
FAMILY						0,304	0,881	***0,005***
SEDENTARY							***0,000***	0,480
TYPE A								***0,000***

Από τους πίνακες 1.19 και 1.14 παρατηρείται ότι στα δύο φύλα (άνδρες – γυναίκες) οι παράγοντες κινδύνου διαφοροποιούνται στην εξάρτηση που εμφανίζουν .

Μια συγκριτική εικόνα των αλληλεξαρτήσεων των παραγόντων κινδύνου που παρατηρήθηκαν στα δύο φύλα όπως αυτή προέκυψε από την μελέτη των δύο αρχείων παρουσιάζεται στον πίνακα 1.20

ΠΙΝΑΚΑΣ 1.20: ΣΥΓΚΡΙΤΙΚΟΣ ΠΙΝΑΚΑΣ ΑΛΛΗΛΕΞΑΡΤΗΣΗΣ

ΦΥΛΟ (Άνδρες(Α) – Γυναίκες(Γ))																
ΠΑΡΑΓΩΝ ΚΙΝΔΥΝΟΥ	AG		LIPIDS		HBP		SMOKE		DIABETES		OBESITY		FAMILY		TYPE A	
	Α	*Γ*	*Α*	*Γ*	*Α*	*Γ*	*Α*	*Γ*	*Α*	*Γ*	*Α*	*Γ*	*Α*	*Γ*	*Α*	*Γ*
LIPIDS							●			◈		◈			●	◈
HBP											●					
SMOKE		◈									●		●		●	◈
DIABETES											●					◈
OBESITY																
FAMILY		◈														
TYPE A		◈														
SEDENTARY	●						●		●		●					◈

Παρατίθεται ένας συγκριτικός συνοπτικός πίνακας των παραγόντων κινδύνου που ευνοούν την εμφάνιση της νόσου στα δύο φύλα όπως προκύπτει από τη μελέτη που προηγήθηκε.

ΠΙΝΑΚΑΣ 1.21: ΣΥΓΚΡΙΤΙΚΟΣ ΠΙΝΑΚΑΣ ΠΑΡΑΓΟΝΤΩΝ ΚΙΝΔΥΝΟΥ ΜΕ ΤΗΝ ΕΚΔΗΛΩΣΗ ΤΗΣ ΝΟΣΟΥ.

ΦΥΛΟ [Άνδρες (Α) - Γυναίκες (Γ)]		
ΠΑΡΑΓΩΝ ΚΙΝΔΥΝΟΥ	**DCODE**	**DCODE**
	Α	***Γ***
LIPIDS		◈
HBR		
SMOKE	●✳	◈
DIABETES		◈
OBESITY		
FAMILY		◈
SEDENTARY		◈
TYPE A		◈

◈ = ο παράγων κινδύνου που εμφανίζεται στις γυναίκες

● = ο παράγων κινδύνου που εμφανίζεται στους άνδρες

Η λεπτομερής αναφορά των αλληλεξαρτήσεων των μεταβλητών στο συνολικό αρχείο καθώς και στα αρχεία ανδρών –γυναικών σκοπό έχει να δώσει ένα έναυσμα για τη μελέτη της διαφοροποίησης των παραγόντων κινδύνου στα δύο φύλα. Στα επόμενα κεφάλαια μέσω διαχωριστικών τεχνικών και νευρωνικών δικτύων επιχειρείται η εύρεση του καλύτερου δυνατού μοντέλου το οποίο να περιγράφει και να προβλέπει την εκδήλωση στεφανιαίας νόσου συναρτήσει των μεταβλητών που προαναφέρθηκαν.

ΚΕΦΑΛΑΙΟ 2°

ΠΡΟΒΛΕΨΗ ΤΗΣ ΕΜΦΑΝΙΣΗΣ ΤΗΣ ΚΑΡΔΙΑΚΗΣ ΝΟΣΟΥ

2.1 Πιλοτικό δείγμα.

Επειδή στο συνολικό δείγμα οι ασθενείς αντιπροσωπεύονται σε πολύ μεγαλύτερο ποσοστό από τους υγιείς (αριθμός ασθενών=580 άτομα (87,9%), αριθμός υγιών = 80 άτομα(12,1%)), κρίθηκε αναγκαίο να επιλεγεί ένα πιλοτικό δείγμα στο οποίο υπάρχουν 80 υγιή και 80 ασθενή άτομα, ώστε οι δύο ομάδες να αντιπροσωπεύονται με την ίδια αναλογία 50% στο πιλοτικό δείγμα. Από τα 80 υγιή άτομα, 14 είναι άνδρες και 66 γυναίκες. Στα 80 ασθενή άτομα που επελέγησαν, τηρήθηκε η ίδια αναλογία ως προς το φύλο καθώς και ο ίδιος αριθμός ατόμων ανά ομάδα ηλικιών. Ο τρόπος επιλογής του καθώς και η σύνθεση του πιλοτικού δείγματος φαίνεται στον πίνακα που ακολουθεί:

ΠΙΝΑΚΑΣ 2.1: Σύνθεση του πιλοτικού δείγματος

	Α Ν Δ Ρ Ε Σ (Α)		Γ Υ Ν Α Ι Κ Ε Σ (Γ)	
ΗΛΙΚΙΑ	**ΥΓΙΕΙΣ**	**ΑΣΘΕΝΕΙΣ**	**ΥΓΙΕΙΣ**	**ΑΣΘΕΝΕΙΣ**
[40-45)	3	3	2	2
[45-50)	4	4	4	4
[50-55)	3	3	15	15
[55-60)	1	1	12	12
[60-65)	2	2	23	23
[65-70)	1	1	6	6
[70-75)	0	0	4	4
ΣΥΝΟ-	14	14	66	66

Στον πίνακα 2.2 παρατίθενται τα στατιστικά του πιλοτικού δείγματος.

ΠΙΝΑΚΑΣ 2.2: Στατιστικά των δύο ομάδων

SEX	DCODE	N	ΜΕΣΗ ΤΙΜΗ ΗΛΙΚΙΩΝ	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ	ΤΥΠΙΚΟ ΣΦΑΛΜΑ
0(Α)	0	14	50,71	7,56	2,02
	1	66	58,29	6,85	0,84
1(Γ)	0	14	50,71	7,56	2,02
	1	66	58,29	6,87	0,85

Στο προηγούμενο κεφάλαιο παρατηρήθηκε ότι ο αριθμός των παραγόντων κινδύνου που εμφανίζουν εξάρτηση μεταξύ τους είναι αρκετά μεγάλος καθώς και ότι η συμμετοχή τους στην εκδήλωση της νόσου κρίνεται στατιστικά σημαντική. Σε μια προσπάθεια να μειωθεί η διάσταση του διανύσματος των μεταβλητών επιχειρήθηκε ο προσδιορισμός ομάδων "συγγενών" παραγόντων οι οποίοι ενοχοποιούνται για την εκδήλωση της νόσου. Κατάλληλο εργαλείο για το σκοπό αυτό αποτελεί η παραγοντική ανάλυση, αναφορά της οποίας γίνεται στην επόμενη παράγραφο.

2. 2 Παραγοντική ανάλυση

Ο όρος παραγοντική ανάλυση συνήθως περιλαμβάνει τόσο την συνήθη παραγοντική ανάλυση, όσο και την ανάλυση σε κύριες συνιστώσες. Αν και βασίζονται σε διαφορετικά μαθηματικά μοντέλα, μπορούν και οι δύο να χρησιμοποιηθούν για την ανάλυση των ίδιων δεδομένων και να παράγουν παρόμοια αποτελέσματα.

Σε περίπτωση p αρχικών μεταβλητών το ζητούμενο είναι να χρησιμοποιηθεί αριθμός κ νέων μεταβλητών(παράγοντες), με $\kappa < p$, για να μεταφέρει το

μεγαλύτερο μέρος της πληροφορίας που περιέχεται στα αρχικά δεδομένα. Και οι δύο τεχνικές χρησιμοποιούνται για την ελάττωση της διάστασης ενός χώρου R^p (p=αριθμός μεταβλητών), σε $R^κ$ με κ < p υπό την προϋπόθεση ότι οι αρχικές μεταβλητές είναι συσχετισμένες.

Ένας άλλος στόχος και των δύο μεθόδων είναι η μελέτη των συσχετίσεων μεταξύ ενός μεγάλου αριθμού ενδοσυσχετισμένων ποσοτικών μεταβλητών και η ομαδοποίηση των μεταβλητών αυτών σε μικρό αριθμό ασυσχέτιστων παραγόντων. Οι μεταβλητές του ίδιου παράγοντα είναι ισχυρότερα συσχετισμένες μεταξύ τους, παρά με τις μεταβλητές που περιλαμβάνονται σε οποιονδήποτε άλλον παράγοντα. Οι κοινοί παράγοντες που προκύπτουν από την ομαδοποίηση των επιμέρους μεταβλητών περιγράφουν το κοινό χαρακτηριστικό των μεταβλητών που τους απαρτίζουν.

Αν και η παραγοντική ανάλυση καθώς και η ανάλυση σε κύριες συνιστώσες έχουν χαρακτηρισθεί ως τεχνικές μείωσης της διάστασης του χώρου, υπάρχουν σημαντικές διαφορές μεταξύ τους.

Σκοπός της ανάλυσης σε κύριες συνιστώσες είναι να ελαττώσει τον αριθμό των αρχικών μεταβλητών σε ένα μικρότερο αριθμό συνιστωσών έτσι ώστε κάθε συνιστώσα να αποτελεί μια νέα μεταβλητή και ο αριθμός των επιλεγμένων συνιστωσών να εξηγεί το μέγιστο της διακύμανσης των δεδομένων. Αντιθέτως η παραγοντική ανάλυση στόχο έχει, να προσδιορίσει ένα σύνολο μη παρατηρηθέντων ασυσχέτιστων παραγόντων(κρυφοί παράγοντες), που μπορούν να ερμηνεύσουν την ενδοσυσχέτιση των μεταβλητών που περιέχονται σε κάθε παράγοντα.

Θα μπορούσαμε δηλαδή να ισχυρισθούμε ότι η μέθοδος των κυρίων συνιστωσών δίνει έμφαση στην ερμηνεία της διακύμανσης των δεδομένων, ενώ η παραγοντική ανάλυση επιχειρεί να ερμηνεύσει τον συσχετισμό μεταξύ των μεταβλητών.

Στην ανάλυση κυρίων συνιστωσών, κάθε παράγοντας(y_j) είναι γραμμικός συνδυασμός p αρχικών μεταβλητών $x_1, x_2, \ldots x_p$.

$$\text{Δηλαδή} \qquad y_j = \alpha_{j1}x_1 + \alpha_{j2}x_2 + \ldots + \alpha_{jp}x_p$$

Και α_{ji} είναι το φορτίο της i μεταβλητής στον j παράγοντα(κύρια συνιστώσα).

Στην παραγοντική ανάλυση, το παραγοντικό μοντέλο αναδεικνύει την ύπαρξη των κρυφών παραγόντων. Κάθε μεταβλητή είναι γραμμική συνάρτηση των κ κρυφών παραγόντων.

$$x_1 = \lambda_{11}y_1 + \lambda_{12}y_2 + \ldots\lambda_{1k}y_k + \varepsilon_1$$
$$x_2 = \lambda_{21}y_1 + \lambda_{22}y_2 + \ldots\lambda_{2k}y_k + \varepsilon_2$$
$$\vdots$$
$$x_p = \lambda_{p1}y_1 + \lambda_{p2}y_2 + \ldots\lambda_{pk}y_k + \varepsilon_p$$

ε_i είναι ο όρος που εκφράζει το σφάλμα και ονομάζεται μοναδιαίος παράγων (unique factor) και λ_{ij} είναι τα φορτία του j παράγοντα στην i μεταβλητή. (Sharma(1996)).

Το μαθηματικό μοντέλο στην παραγοντική ανάλυση φαίνεται παρόμοιο με την εξίσωση της πολυμεταβλητής παλινδρόμησης. Σε γενικές γραμμές η ανάλυση κατά παράγοντες και ιδιαίτερα η εύρεση κυρίων συνιστωσών έχει ως εξής:

Έστω ότι $x=(x_1,.....,x_p)'$ είναι $p\times1$ τυχαίο διάνυσμα με μέσο μ και πίνακα διακύμανσης $\Sigma(p\times p)$. Ζητείται να βρεθεί ένα καινούργιο σύνολο μεταβλητών $y=(y_1,....,y_p)'$ που να είναι ασυσχέτιστες και για τις οποίες να ισχύει:

$$V(y_{(1)}) > V(y_{(2)}) > \ldots > V(y_{(p)})$$

Κάθε y_j είναι γραμμικός συνδυασμός των x_i, δηλαδή $y_j = a_j' x$ $(j=1,....,p)$ όπου $\alpha_j'=(\alpha_{1j},.....,\alpha_{pj})$ είναι ένα σταθερό διάνυσμα *1xp* , για το οποίο ισχύει $a_j' a_j = 1$.

Η πρώτη κύρια συνιστώσα $y_1 = a_1' x$ προσδιορίζεται έτσι ώστε η διακύμανσή της, $V(y_1)= a_1' \Sigma a_1$, να είναι μέγιστη. Επιλέγουμε το a_1 ώστε η διακύμανση του $y_1 = a_1 x$ να μεγιστοποιείται υπό τον περιορισμό $\alpha_1\alpha_1 = 1$. Παραγώγιση της συνάρτησης: $f(\alpha_1,\lambda) = \alpha_1 \Sigma\alpha_1 - \lambda(\alpha_1\alpha_1 - 1)$, οδηγεί στην $(\Sigma - \lambda I)a_1 = 0$. (2.1)
από την οποία προκύπτει ότι το a_1 είναι ένα ιδιοδιάνυσμα του πίνακα Σ. Επειδή $V(y_1) = \alpha_1' \Sigma\alpha_1 = \alpha_1' (\lambda_1 \cdot I)\alpha_1 = \lambda_1$, η μεγιστοποίηση της διασποράς της y_1, οδηγεί στην επιλογή του a_1 ως το ιδιοδιάνυσμα του πίνακα Σ που αντιστοιχεί στη μέγιστη ιδιοτιμή του.

Για να βρούμε τη δεύτερη κύρια συνιστώσα $y_2 = a_2' x$ μεγιστοποιούμε τη διασπορά της y_2, $V(y_2) = a_2' \Sigma a_2$, κάτω από τους περιορισμούς $a_2' a_2 = 1$ και $Cov(y_2, y_1) = 0$.

Έτσι $Cov(y_2, y_1) = Cov(\alpha_2' x, \alpha_1' x) = E[\alpha_2'(x-\mu)(x-\mu)'\alpha_1] = \alpha_2' \Sigma \alpha_1$

πρέπει να ισούται με μηδέν. Αλλά: $\Sigma a_1 = \lambda_1 a_1$

$$\alpha_2' \Sigma \alpha_1 = \alpha_2' \lambda_1 \alpha_1$$

οπότε $\alpha_2' \Sigma \alpha_1 = \alpha_2' \lambda_1 \alpha_1 = 0$

Η σχέση αυτή δείχνει την ορθογωνιότητα των διανυσμάτων. Η διακύμανση της y_2 μεγιστοποιείται κάτω από τους περιορισμούς: $a_2' a_2 = 1$ και $\alpha_2' \alpha_1 = 0$.

Χρησιμοποιώντας πολλαπλασιαστές Lagrange καταλήγουμε στην εξίσωση

$$(\Sigma - \lambda I) a_2 = 0$$

Έτσι το λ στην περίπτωση αυτή είναι η δεύτερη μεγαλύτερη ιδιοτιμή του πίνακα Σ και a_2 το αντίστοιχο ιδιοδιάνυσμα.

Συνεχίζοντας με την ίδια τεχνική αποδεικνύεται ότι για τον j παράγοντα $y_j = a_j' x$ επιλέγουμε το a_j που είναι το ιδιοδιάνυσμα που αντιστοιχεί στη j μεγαλύτερη ιδιοτιμή του Σ.

Ο αριθμός των παραγόντων που απαιτείται για να περιγράψει καλύτερα τα δεδομένα του προβλήματος, εξαρτάται από το ποσοστό της συνολικής διακύμανσης που ερμηνεύεται από κάθε παράγοντα.

Ως γνωστό, η συνολική διακύμανση είναι το άθροισμα της διακύμανσης κάθε μεταβλητής. Όλες οι μεταβλητές και οι παράγοντες εκφράζονται σε κανονικοποιημένη μορφή, με μέσο μηδέν και τυπική απόκλιση 1. Η συνολική διακύμανση που ερμηνεύεται από κάθε παράγοντα εκφράζεται με την τιμή της ιδιοτιμής του.

Δύο είναι τα επικρατέστερα κριτήρια εξαγωγής κυρίων συνιστωσών:

α) Σύμφωνα με το πρώτο κριτήριο, μόνο κύριες συνιστώσες των οποίων οι ιδιοτιμές είναι μεγαλύτερες της μονάδας (λ >1) συμπεριλαμβάνονται στο παραγοντικό μοντέλο. Παράγοντες με διακύμανση μικρότερη του 1 δε συνεισφέρουν περισσότερο από μια απλή μεταβλητή, δοθέντος ότι κάθε μεταβλητή έχει διακύμανση 1.

β) Η Jolliffee προτείνει ως κριτήριο εξαγωγής κύριων συνιστωσών, η τιμή των ιδιοτιμών να είναι μεγαλύτερη από 0,7.

Λεπτομέρειες για τη χρήση της ανάλυσης κατά κύριες συνιστώσες μπορούν να βρεθούν στις εργασίες των Sharma (1996), Pedhazur (1972) , Morrison (1976).

2.2.1 Χρήση της παραγοντικής ανάλυσης σε ιατρικές έρευνες

Ενδεικτικά παρατίθενται αναφορές από προηγούμενες ιατρικές έρευνες, όπου η παραγοντική ανάλυση χρησιμοποιήθηκε για να προσδιορίσει ομάδες παραγόντων κινδύνου σε άτομα με συμπτώματα στεφανιαίας νόσου μελετώντας τη δομή συσχέτισης μεταξύ των μεταβλητών αυτών.

Η έρευνα της Edwards (1997), σε δείγματα προερχόμενα από τρεις διαφορετικούς πληθυσμούς, αφορά το σύνδρομο IRS (Insulin Resistance Syndrome), το οποίο χαρακτηρίζεται από την ενδοσυσχέτιση του με τη στεφανιαία νόσο και τον μη εξαρτώμενο από ινσουλίνη διαβήτη. Τα αποτελέσματα έδειξαν ότι η παραγοντική ανάλυση κατέληξε σε τέσσερις ασυσχέτιστους παράγοντες και στα τρία δείγματα των διαφορετικών πληθυσμών και είναι οι εξής: α) Βάρος σώματος β) Ινσουλίνη/γλυκόζη γ) Λιπίδια και δ) αρτηριακή πίεση. Σε παρόμοια συμπεράσματα είχε καταλήξει η ίδια έρευνα των Edwards et al (1994) σε 281 γενετικά ασυσχέτιστες μη διαβητικές γυναίκες, όπου οι παράγοντες ερμήνευαν το 66% της συνολικής διακύμανσης των δεδομένων.

Η Rees (1996), χρησιμοποιώντας περιγραφική στατιστική, παραγοντική ανάλυση, βήμα προς βήμα πολυμεταβλητή παλινδρόμηση και κανονική συσχέτιση σε δείγμα 82 στεφανιαίων ασθενών, υποστηρίζουν ότι η ενημέρωση των ασθενών ότι πάσχουν από στεφανιαία νόσο, το άγχος , το κοινωνικό επίπεδο , η αποτελεσματικότητα επηρεάζουν τις διατροφικές συνήθειες, τη σωματική άσκηση και το κάπνισμα.

Η Townsend (1995), χρησιμοποιώντας λογιστική παλινδρόμηση και παραγοντική ανάλυση σε δείγμα 196 ατόμων, υποστηρίζει ότι υπάρχει εξάρτηση μεταξύ της στεφανιαίας νόσου και της οργής ή εχθρικότητας τόσο στους άνδρες όσο και στις γυναίκες. Η παραγοντική ανάλυση καταλήγει σε τρεις ανεξάρτητους παράγοντες συνδεδεμένους με την οργή: α) την εξωτερίκευση της

οργής, παράγων που φαίνεται να είναι άμεσα συσχετισμένος με τη στεφανιαία νόσο β) την αυτοελεγχόμενη οργή και γ) την εχθρικότητα.

Η συσχέτιση της συμπεριφοράς ατόμων τύπου Α, του άγχους και της στεφανιαίας νόσου μελετήθηκε από τις Low και Graff (1991), σε δείγμα 84 εμφραγματιών γυναικών. Η παραγοντική ανάλυση κατέληξε σε πέντε παράγοντες: α) πανικός β) σωματικά συμπτώματα γ) κοινωνικό άγχος δ) ανησυχία και ε) κατάθλιψη. Σε παρόμοια συμπεράσματα καταλήγουν και οι M. De Leon ,C. Francisco(1988), οι οποίοι υποστηρίζουν ότι η συμπεριφορά τύπου Α είναι ένας σημαντικός παράγων κινδύνου για την εκδήλωση εμφράγματος σε πληθυσμούς με ανάμικτο κοινωνικο-οικονομικό και εθνικό υπόβαθρο. Η εξωτερίκευση της οργής φαίνεται να είναι συσχετισμένη με την εκδήλωση εμφράγματος καθώς και με το προκάρδιο άλγος. Η αϋπνία φαίνεται συσχετισμένη με το προκάρδιο άλγος, ενώ το κάπνισμα και η κατανάλωση καφέ εμφανίζει την ισχυρότερη συσχέτιση με το έμφραγμα.

Οι Godsland et al(1998) ερευνούν σε δείγμα 742 ανδρών με μη διαγνωσμένη νόσο, τη σχέση ανάμεσα στους παράγοντες κινδύνου για καρδιαγγειακή νόσο και το κάπνισμα, τη λήψη αλκοόλ, τη σωματική άσκηση. Η παραγοντική ανάλυση διέκρινε ως κύριους παράγοντες: το μεταβολικό σύνδρομο με τη λίγη σωματική άσκηση, τον υψηλό αριθμό των λευκών κυττάρων, την υψηλή συγκέντρωση αιμογλουβίνης, το χαμηλό HDL, το αυξανόμενο κάπνισμα καθώς και τη λήψη αλκοόλ.

2.2.2 Εφαρμογή της παραγοντικής ανάλυσης στο πιλοτικό δείγμα.

Στη συνέχεια, με τη μέθοδο της παραγοντικής ανάλυσης, μελετώνται οι σχέσεις και αλληλοεπιδράσεις μεταξύ των ερμηνευτικών μεταβλητών. Η εφαρμογή της παραγοντικής ανάλυσης στο πιλοτικό δείγμα έχει σκοπό την εύρεση των κυρίων συνιστωσών.

Επειδή ένας από τους σκοπούς της παραγοντικής ανάλυσης είναι να προσδιορίσει παράγοντες που να ερμηνεύουν τις συσχετίσεις των μεταβλητών, οι μεταβλητές πρέπει να είναι συσχετισμένες για να θεωρηθεί κατάλληλο το παραγοντικό μοντέλο. Οι μεταβλητές που εμφανίζουν χαμηλή συσχέτιση

(-0,4, 0,4) δεν είναι πιθανό να ανήκουν σε κοινό παράγοντα.

Για τον έλεγχο της καταλληλότητας ενός μοντέλου εφαρμόζονται έλεγχοι όπως:

α) Bartlett-test (έλεγχος σφαιρικότητας).

Για να εφαρμοσθεί ο έλεγχος απαιτείται τα δεδομένα να είναι δείγμα προερχόμενο από πολυμεταβλητό κανονικό πληθυσμό.

Με τον έλεγχο αυτό, ελέγχεται η υπόθεση, αν ο πίνακας συσχετίσεων του πληθυσμού είναι ο μοναδιαίος. Αν η τιμή του ελέγχου σφαιρικότητας είναι μεγάλη και το αντίστοιχο επίπεδο σημαντικότητας είναι μικρό, τότε η υπόθεση ότι ο πίνακας συσχετίσεων είναι ο μοναδιαίος απορρίπτεται. Αν η υπόθεση αυτή δεν μπορεί να απορριφθεί διότι το παρατηρηθέν επίπεδο σημαντικότητας είναι μεγάλο, η χρήση του παραγοντικού μοντέλου πρέπει να αναθεωρηθεί.

β) Ο μερικός συντελεστής συσχέτισης

Ένας άλλος δείκτης που εκφράζει την δύναμη της συσχέτισης μεταξύ των μεταβλητών είναι ο μερικός συντελεστής συσχέτισης. Οι μερικές συσχετίσεις είναι εκτιμήσεις των συσχετίσεων μεταξύ των μοναδιαίων παραγόντων και πρέπει να παίρνουν τιμές σχεδόν μηδενικές κατά την εφαρμογή της παραγοντικής ανάλυσης. Υπενθυμίζεται ότι απαιτείται οι μοναδιαίοι παράγοντες να είναι ασυσχέτιστοι.

γ) ο αρνητικός μερικός συντελεστής συσχέτισης(anti-image correlation).

Η αρνητική τιμή του μερικού συντελεστή συσχέτισης ονομάζεται "anti-image correlation". Αν το πλήθος των μεγάλων συντελεστών στον πίνακα των αρνητικών μερικών συντελεστών συσχέτισης είναι μεγάλο, πρέπει να αναθεωρηθεί η χρήση του παραγοντικού μοντέλου και

δ) ο έλεγχος των Kaiser-Meyer-Olkin (KMO test).

Ένας δείκτης σύγκρισης των παρατηρηθέντων μέτρων συντελεστών συσχέτισης, με τα μέτρα των συντελεστών μερικής συσχέτισης είναι το μέτρο της δειγματοληπτικής επάρκειας των Kaiser-Meyer-Olkin. Υπολογίζεται από τη σχέση:

$$KMO=\frac{\sum\sum_{i\neq j} r_{ij}^2}{\sum\sum_{i\neq j} r_{ij}^2 + \sum\sum_{i\neq j} \alpha_{ij}^2}$$

Όπου r_{ij} είναι ο συντελεστής συσχέτισης μεταξύ των μεταβλητών i και j και α_{ij} είναι ο συντελεστής μερικής συσχέτισης μεταξύ των μεταβλητών i και j. Αν το άθροισμα των τετραγώνων των συντελεστών μερικής συσχέτισης μεταξύ όλων των ζευγών των μεταβλητών είναι μικρό σε σχέση με το άθροισμα των τετραγώνων των συντελεστών συσχέτισης, τότε ο δείκτης KMO είναι κοντά στη μονάδα. Μικρές τιμές του δείκτη KMO δείχνουν ότι η παραγοντική ανάλυση δεν είναι η βέλτιστη ενδεδειγμένη τεχνική αφού οι συσχετίσεις μεταξύ ζευγών μεταβλητών δεν μπορούν να ερμηνευθούν από άλλες μεταβλητές.

Σύμφωνα με τον Kaiser (1974) οι τιμές του δείκτη κυμαίνονται ως εξής:

KMO	Ποιότητα δείκτη	
≥ 0,90	εξαίρετος	(marvelous)
0,80+	αξιόλογος	(meritotious)
0.70+	καλός	(middling)
0,60+	**μέτριος**	**(mediocre)**
0,50+	ανεπαρκής	(miserable)
< 0,50	μη αποδεκτός	(unacceptable)

Τιμές μικρότερες του 0.5 οδηγούν σε αδυναμία εφαρμογής της παραγοντικής ανάλυσης. Στο συγκεκριμένο πιλοτικό δείγμα, ο δείκτης KMO έχει την τιμή 0.663, γεγονός που μας επιτρέπει να προχωρήσουμε στην παραγοντική ανάλυση. (SPSS 1990).

Για να ελεγχθεί αν το μοντέλο των k παραγόντων περιγράφει ικανοποιητικά τις αρχικές μεταβλητές, υπολογίζουμε το ποσοστό της διακύμανσης της κάθε μεταβλητής που ερμηνεύεται από τους k παράγοντες. Αν οι παράγοντες είναι ασυσχέτιστοι, το ολικό ποσοστό της ερμηνευμένης διακύμανσης είναι απλά το άθροισμα των ποσοστών των διακυμάνσεων που ερμηνεύονται από κάθε παράγοντα. Το ποσοστό της διακύμανσης που ερμηνεύεται από τους κοινούς παράγοντες ονομάζεται "communality" της μεταβλητής.

Κατά την εφαρμογή της παραγοντικής ανάλυσης, μέθοδος κυρίων συνιστωσών, στο πιλοτικό δείγμα εξήχθησαν πέντε ή επτά κύριες συνιστώσες ανάλογα με το κριτήριο που γίνεται αποδεκτό.

Στον Πίνακα 2.4 που ακολουθεί υπάρχουν πέντε ιδιοτιμές (λ) μεγαλύτερες της μονάδας. Σύμφωνα με τα κριτήρια που προαναφέρθηκαν, για $\lambda > 1$ στο παραγοντικό μοντέλο συμπεριλαμβάνονται μόνο οι πρώτες 5 κύριες συνιστώσες ($R^{11} \rightarrow R^5$), ενώ για $\lambda > 0,7$ συμπεριλαμβάνονται 7 κύριες συνιστώσες.

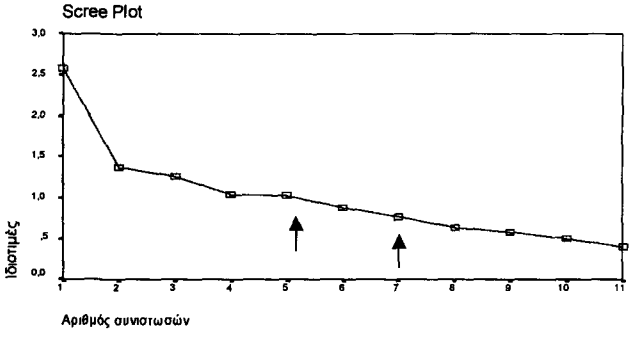
Η ολική διακύμανση που ερμηνεύεται από κάθε συνιστώσα καταγράφεται στη στήλη των ιδιοτιμών (στήλη (II)) και αντίστοιχα το ποσοστό της διακύμανσης που η κάθε συνιστώσα συνεισφέρει στο σύνολο της διακύμανσης που καταγράφεται στη στήλη (III). Στη στήλη (IV) το ποσοστό στην $κ^η$ γραμμή είναι το ποσοστό των κ πρώτων διακυμάνσεων επί του συνόλου των διακυμάνσεων.

ΠΙΝΑΚΑΣ 2.4

	Αρχικές ιδιοτιμές		
(I)	(II)	(III)	(IV)
Παράγο-	Ιδιοτι-	% της διασπο-	Αθροιστική
1	2.569	23.354	23.354
2	1.369	12.449	35.803
3	1.263	11.477	47.281
4	1.040	9.459	56.740
5	1.023	9.302	66.042
6	0.877	7.969	74.011
7	0.763	6.938	80.949
8	0.631	5.738	86.687
9	0.569	5.176	91.863
10	0.495	4.503	96.366
11	0,400	3,634	100

Χρησιμοποιώντας το κριτήριο $\lambda > 1$ οι 5 κύριες συνιστώσες ερμηνεύουν το 66% του συνόλου της διακύμανσης ενώ με το κριτήριο της Jolliffee, οι 7 κύριες συνιστώσες ερμηνεύουν το 80,9% του συνόλου της διακύμανσης. Σύμφωνα με τους Tucker et al (1969) το κριτήριο $\lambda > 1$ δεν παρέχει πάντα τη βέλτιστη λύση.

Οι κύριες συνιστώσες και οι αντίστοιχες ιδιοτιμές απεικονίζονται σε ένα διάγραμμα που ονομάζεται Scree Plot (Catell 1966). Στο διάγραμμα αναδεικνύεται η συνεισφορά κάθε παράγοντα στη συνολική διακύμανση. (Σχήμα 2.1).



Οι συντελεστές που συσχετίζουν τις αρχικές μεταβλητές με τις κύριες συνιστώσες λέγονται φορτία (factor loadings) και έχουν παρόμοιο ρόλο με αυτόν που έχουν οι συντελεστές παλινδρόμησης. Οι παράγοντες που εμφανίζουν μεγάλους συντελεστές (κατ' απόλυτη τιμή) στο παραγοντικό μοντέλο

$$x_i = b_{i1}y_1 + b_{i2}y_2 + ... + b_{ik}y_k + e_i \; i = 1,...,p \qquad (2.2)$$

είναι ισχυρά συσχετισμένοι με την μεταβλητή. Αν οι εκτιμηθέντες παράγοντες είναι ασυσχέτιστοι μεταξύ τους (είναι ορθογώνιοι), τότε τα φορτία είναι οι συσχετίσεις μεταξύ των παραγόντων και των μεταβλητών. Στο παραγοντικό μοντέλο τα φορτία είναι οι τυποποιημένοι συντελεστές παλινδρόμησης σε μία πολυμεταβλητή εξίσωση παλινδρόμησης με εξαρτημένη μεταβλητή την αρχική μεταβλητή και τους παράγοντες ως ανεξάρτητες μεταβλητές.

Συνήθως η αρχική ομαδοποίηση των παραγόντων δεν παρέχει πάντα τη δυνατότητα ουσιαστικής και εύκολης ερμηνείας.

Μια τέτοια δυνατότητα παρέχεται όμως με την περιστροφή των παραγόντων. Το ζητούμενο με την περιστροφή είναι να επιτευχθεί μια απλούστερη δομη (διάρθρωση) των παραγόντων, εφόσον κάθε μια μεταβλητή τελικά εμφανίζει ισχυρή συσχέτιση με ένα μόνο παράγοντα.

Στην ορθογώνια περιστροφή οι παράγοντες που προκύπτουν, είναι ασυσχέτιστοι. Varimax και Quartimax είναι οι συνηθέστερες τεχνικές των ορθογωνίων περιστροφών.

Με την τεχνική Varimax που χρησιμοποιείται στην παρούσα εργασία, επιχειρείται η εύρεση μιας τέτοιας δομής παράγοντα έτσι ώστε κάθε μεταβλητή να συμμετέχει με μεγάλο φορτίο σε ένα και μόνο παράγοντα και με φορτία κοντά στο μηδέν στους άλλους παράγοντες. Μια τέτοια δομή παράγοντα θα έχει σαν αποτέλεσμα κάθε παράγοντας να αντιπροσωπεύει μια διακεκριμένη ομάδα μεταβλητών. Όταν δηλαδή, τα φορτία και οι σημαντικές μεταβλητές έχουν τιμή κοντά στη μονάδα και οι μη σημαντικές μεταβλητές έχουν τιμή κοντά στο μηδέν, τότε η διάρθρωση του κάθε παράγοντα έχει προσδιορισθεί.

Τα φορτία για κάθε ένα από τους 5 παράγοντες πριν την περιστροφή των αξόνων δίδονται από τον Πίνακα 2.5

ΠΙΝΑΚΑΣ 2.5

Πίνακας των φορτίων πριν την περιστροφή a

	Παράγοντες				
	1	2	3	4	5
AGE	-0,665	-4,944-02	-8,442-02	0,214	-0,247
DIABETES	-0,210	0,646	0,269	0,307	7,961-03
EF	0.393	-0,314	0,618	1,695-02	0,431
FAMILY	8,844-02	9,666-02	-0,563	-0,472	0,429
HBP	-0,307	0,450	0,409	0,391	0,280
LIPIDS	-0,102	0,667	0,236	-0,344	0,284
SEDENTARY	0,526	7,446-02	-1,406-02	0,351	-0,177
SEX	0,784	6,726-02	-9,948-02	0,156	0,105
SMOKE	0,665	0,293	0,126	-5,187-02	-0,329
TYPE A	0,695	0,195	1,379-02	8,338-02	-3,259-02
OBESITY	6,644-03	0,244	0,485	-0,489	-0,532

Μέθοδος προσδιορισμού των παραγόντων: Ανάλυση κατά παράγοντες a.
5 παράγοντες εξήχθησαν

Αν και ο πίνακας των φορτίων πριν την περιστροφή δείχνει τη σχέση μεταξύ των μεταβλητών και των κυρίων συνιστωσών, είναι δύσκολο να προσδιοριστούν σημαντικές κύριες συνιστώσες.

Αποδεχόμενοι το κριτήριο ότι : "Μεταβλητές με φορτία μικρότερα του 0,5 κατ' απόλυτη τιμή, δε συμμετέχουν στον παράγοντα", παρατηρούμε ότι ο τέταρτος παράγοντας δε συσχετίζεται ικανοποιητικά με καμία μεταβλητή.

Επίσης η μεταβλητή HBP λόγω της ασθενούς συσχέτισης που εμφανίζει και με τους πέντε παράγοντες μπορεί να εκφρασθεί ως γραμμικός συνδυασμός τουλάχιστον δύο παραγόντων.

Σύμφωνα με τα παραπάνω οδηγούμεθα σε περιστροφή των παραγόντων. Τα φορτία των παραγόντων μετά την περιστροφή που εμφανίζουν απόλυτη τιμή μεγαλύτερη του 0,5 καταγράφονται στον πίνακα 2.6

ΠΙΝΑΚΑΣ 2.6

Πίνακας των φορτίων μετά την περιστροφή [a]

	παράγοντες				
	1	2	3	4	5
AGE	-,544				
DIABETES		,583	,512		
EF		-,878			
FAMILY				,841	
HBP			,766		
LIPIDS			,659		
SEDENTAR	,601				
SEX	,768				
SMOKE	,743				
TYPEA	,717				
OBESITY					,899

Μέθοδος προσδιορισμού των παραγόντων: Ανάλυση κατά παράγοντες.
Μέθοδος περιστροφής: Varimax

a. Η περιστροφή ολοκληρώθηκε στις 10 επαναλήψεις.

Όπως παρατηρεί κανείς στον Πίνακα 2.6 **ο πρώτος παράγοντας** y_1 σχετίζεται με τις μεταβλητές AGE, SEDENTARY, SEX, SMOKE, TYPE A και περιγράφει τον τρόπο ζωής του ατόμου. Με άλλα λόγια ο "κρυφός" παράγοντας που ερμηνεύει τη συσχέτιση των παραπάνω μεταβλητών είναι ο τρόπος ζωής του ατόμου.

Ο δεύτερος παράγοντας y_2, σχετίζεται με τις μεταβλητές DIABETES, EF, και μπορεί να θεωρηθεί ότι περιγράφει την προδιάθεση του ατόμου. **Ο τρίτος παράγοντας** y_3 περιγράφει την κατάσταση υγείας του ατόμου, γιατί σχετίζεται

με τις μεταβλητές DIABETES, HBP, LIPIDS. **Ο τέταρτος παράγοντας** y_4 σχετίζεται με το ιστορικό του ασθενούς και ο **πέμπτος παράγοντας** y_5 είναι η παχυσαρκία (Obesity).

Το άθροισμα των τετραγώνων των συντελεστών ανά γραμμή δίνει το ποσοστό της μεταβλητότητας της αρχικής μεταβλητής που ερμηνεύεται από τους παράγοντες αυτούς.

Για παράδειγμα το $0,583^2+0,512^2 = 0,602$ ή 60,2% είναι το ποσοστό της μεταβλητής DIABETES που ερμηνεύεται από τη κατάσταση της υγείας του ατόμου(παράγων y_3) και την προδιάθεσή του σε καρδιακά νοσήματα (παράγων y_2). Αν χρησιμοποιήσουμε και τα φορτία της μεταβλητής DIABETES ως προς τους άλλους παράγοντες τότε θα προκύψει:

$$0,602+0,0822^2 +0,02117^2+(-0,136)^2=0,6277.$$

Το ποσοστό αυτό λέγεται communality και για κάθε μεταβλητή δίδεται στον πίνακα 2.7, που ακολουθεί

ΠΙΝΑΚΑΣ 2.7

Μεταβλητές	**Communalities**
AGE	0,550
DIABETES	0,628
EF	0,821
FAMILY	0,741
HBP	0,696
OBESITY	0,826
SEDENDARY	0,437
SEX	0,664
SMOKE	0,654
TYPE A	0,529
LIPIDS	0,710

Παρατηρείται ότι και μετά την περιστροφή η μεταβλητή DIABETES εμφανίζεται ικανοποιητικά συσχετισμένη με δύο κύριες συνιστώσες. Το γεγονός αυτό μας οδηγεί στον προσδιορισμό των κύριων συνιστωσών και με το κριτήριο της Jolliffee.

Στη συνέχεια παρατίθεται ο πίνακας των φορτίων των παραγόντων που εμφανίζουν απόλυτη τιμή μεγαλύτερη του 0,5 πριν και μετά την περιστροφή.

ΠΙΝΑΚΑΣ 2.8 Πίνακας συνιστωσών

	Φορτία πριν την περιστροφή						
	1	2	3	4	5	6	7
SEX	0,784						
AGE	-0,665						
EF			0,618				
LIPIDS		0,667					
HBR							
SMOKE	0,665						
DIABETES		O,646					
OBESITY					-0,532		
FAMILY			-0,563				
SEDENTARY						0,620	
TYPE A	0,695						
	Φορτία μετά την περιστροφή						
	1	2	3	4	5	6	7
SEX	0,704						
AGE		-0,647					
EF		0,844					
LIPIDS			0,867				
HBR			0,627				
SMOKE	0,859						
DIABETES					0,935		
OBESITY						0,970	
FAMILY				0,963			
SEDENTARY							0,846
TYPE A	0,661						

Τα φορτία μετά την περιστροφή των επτά παραγόντων μας οδηγούν στο συμπέρασμα ότι εκτός από τις τρεις πρώτες κύριες συνιστώσες, οι οποίες μπορεί να θεωρηθούν ότι περιγράφουν αντίστοιχα τον τρόπο ζωής του ατόμου (SEX, SMOKE, TYPE A) την προδιάθεση (AGE, EF) και την κατάσταση υγείας του (LIPIDS, HBP), οι υπόλοιποι παράγοντες συμπεριφέρονται ουσιαστικά όπως η μεταβλητή με την οποία εμφανίζονται ισχυρά συσχετισμένοι. Αξίζει να παρατηρήσουμε ότι παράγοντες με διακύμανση μικρότερη της μονάδας δεν συνεισφέρουν περισσότερο στο μοντέλο από ότι μια απλή μεταβλητή, γνωστού όντος κάθε μεταβλητή έχει διακύμανση ίση με τη μονάδα.

2.3 Διαχωριστικές τεχνικές

Ο πρωταρχικός σκοπός της ιατρικής είναι η πρόβλεψη και πρόληψη παρά η θεραπεία. Στην περίπτωση μάλιστα των καρδιοπαθών, λόγω της σοβαρότητας της νόσου η σωστή πρόβλεψη είναι ζωτικής σημασίας. Όπως έχει ήδη αναφερθεί κάθε μία από τις 660 περιπτώσεις που μελετώνται επιχειρείται να ταξινομηθεί σε μία από τις δύο κατηγορίες ασθενής-υγιής. Με βάση το δείγμα αυτό, αναπτύσσονται στη συνέχεια και συγκρίνονται διάφορα μοντέλα (Zhuo et al (1991,1994,1995)) τα οποία ομαδοποιούν την κάθε περίπτωση (υγιής, ασθενής), σύμφωνα με κάποιες μεταβλητές οι οποίες είναι διαθέσιμες από το ιστορικό του ασθενούς.

Οι διαχωριστικές τεχνικές χρησιμοποιούνται ευρύτατα στη στατιστική ανάλυση ιατρικών δεδομένων. Η Rosenberg (1995) υποστηρίζει ότι η διαχωριστική ανάλυση μπορεί να ταξινομήσει ισχαιμικά άτομα με αποτελεσματικότητα 69% με βάση την έκφραση οργής του προσώπου του ατόμου ή η μη ύπαρξη έκφρασης χαράς. Αντίστοιχα η Myers (1993) μελέτησε τρία ενδεχόμενα: α) τη σχέση της παχυσαρκίας, της κατανάλωσης οινοπνεύματος και της διατροφής ως προς την εκδήλωση της στεφανιαίας νόσου β) τη σχέση της μεταβολής βάρους σε άτομα άνω των 65 ετών και γ) τη σχέση των διατροφικών συνηθειών ως προς την εκδήλωση της στεφανιαίας νόσου. Η λογιστική παλινδρόμηση ανέδειξε θετική συσχέτιση ανάμεσα στην εκδήλωση της στεφανιαίας νόσου και την ηλικία, το εισόδημα και το δείκτη σώματος.

Αν συμπεριληφθούν οι διατροφικές συνήθειες, τότε η εμφάνιση της νόσου συσχετίζεται θετικά με το είδος της διατροφής και αρνητικά με το αλκοόλ. Ισχυρός παράγων κινδύνου στις γυναίκες αναδείχθηκε η παχυσαρκία και το κάπνισμα. Παρόμοιες έρευνες έγιναν από τους: Robinson (1992), Suhnoon (1991), Arroyo (1993), Derby (1989), Twisk et al (1997), Morise (1996), Roeters et al (1999), Pyorala etal (2000).

Στη συνέχεια, χρησιμοποιείται η διαχωριστική ανάλυση (discriminant analysis), η λογιστική παλινδρόμηση (logistic regression) καθώς και η επεξεργασία με τη βοήθεια των νευρωνικών δικτύων (neural network) που θα αναπτυχθεί στο τρίτο κεφάλαιο.

Η διαχωριστική ανάλυση και η λογιστική παλινδρόμηση βασίζονται σε διαφορετικές υποθέσεις, αλλά έχουν τους εξής κοινούς στόχους:

α) Να αναδείξουν εκείνες τις μεταβλητές οι οποίες διαφοροποιούν τις δύο κατηγορίες.

β) Με τη χρήση των μεταβλητών αυτών, να προσδιοριστεί μία συνάρτηση με την οποία βρίσκεται μία νέα μεταβλητή ή δείκτης η οποία θα διαφοροποιεί τις δύο ομάδες.

γ) Με τη βοήθεια των α) και β) αναπτύσσεται ένας κανόνας με τον οποίο ταξινομείται κάθε νέα περίπτωση σε μία από τις δύο ομάδες. [Everitt (1989), B. Hazard Murno (1993), Cacoulos (1972)]

Οι μεταβλητές τις οποίες θα λάβουμε υπόψη είναι οι παράγοντες κινδύνου: Φύλο (SEX), Ηλικία (AGE), Κλάσμα Εξώθησης (EF), Λιπίδια (LIPIDS), Αρτηριακή Πίεση (HBP), Κάπνισμα (SMOKE), Διαβήτης (DIABETES), Παχυσαρκία (OBESITY), Οικογενειακό ιστορικό (FAMILY), Τύπος Α (TYPE A), Καθιστική ζωή (SEDENTARY), καθώς και το Κλάσμα εξώθησης (EF).

Η κλασική διαχωριστική ανάλυση απαιτεί από κοινού κανονική κατανομή των εξαρτημένων μεταβλητών και τον ίδιο πίνακα συνδιασποράς για τις δύο ομάδες.

Όμως σε πολλές περιπτώσεις οι υποθέσεις αυτές δεν μπορεί να ισχύουν [Krzanowski (1975), (1980), (1982), (1986), (1987)]. Σε πολλά ιατρικά προβλήματα για παράδειγμα κάποιες από τις εξαρτημένες μεταβλητές είναι συνεχούς τύπου και κάποιες διακριτές [Jerwood et al (1989)]. Αντίθετα, η λογιστική

παλινδρόμηση δεν απαιτεί αυτές τις υποθέσεις γι' αυτό και χρησιμοποιείται περισσότερο.

Η περίπτωση της συνύπαρξης συνεχών και διακριτών μεταβλητών και μελετήθηκε και από τον Knoke (1982). Το βασικό συμπέρασμα είναι ότι αν η εξαρτημένη μεταβλητή δεν εξαρτάται από τις αλληλεπιδράσεις των ανεξαρτήτων μεταβλητών, τότε η γραμμική διαχωριστική συνάρτηση έχει σχεδόν την ίδια ασυμπτωτική αποτελεσματικότητα με την λογιστική παλινδρόμηση. Για την περίπτωση που οι εξαρτημένες μεταβλητές έχουν από κοινού κανονική κατανομή αλλά η διασπορά για τις δύο ομάδες δεν είναι η ίδια, προτείνει την τετραγωνική διαχωριστική συνάρτηση.

Το μειονέκτημα είναι ότι η τεχνική αυτή δεν έχει κωδικοποιηθεί στα γνωστά στατιστικά πακέτα.

Οι Press και Wilson (1978) απαριθμούν τους λόγους για τους οποίους όταν η συνθήκη κανονικότητας δεν ισχύει, η λογιστική παλινδρόμηση είναι πιο αξιόπιστη από τη γραμμική διαχωριστική ανάλυση. Συγκεκριμένα:

1) Η διαχωριστική ανάλυση παράγει μη συνεπείς εκτιμητές των συντελεστών της διαχωριστικής συνάρτησης.
2) Η διαχωριστική συνάρτηση δε δίνει αξιόπιστα επίπεδα σημαντικότητας. Έτσι μη σημαντικές μεταβλητές μπορεί να περιλαμβάνονται στο μοντέλο ενώ δεν είναι στατιστικώς σημαντικές.
3) Συγκρίσεις μεταξύ των παρατηρούμενων και των προβλεπόμενων από τα μοντέλα ποσοστών σωστής πρόβλεψης, δείχνουν ότι η λογιστική ανάλυση έχει καλύτερες δυνατότητες πρόβλεψης.
4) Η λογιστική συνάρτηση έχει στατιστική συνάρτηση που σχετίζεται με την εκτίμηση των μεταβλητών, ενώ η διαχωριστική δεν έχει.
5) Υπάρχουν ενδείξεις ότι η διαχωριστική συνάρτηση μεροληπτεί στην ταξινόμηση των περιπτώσεων.

Οι Byth και McLachlan (1980) μελετούν την περίπτωση που αντί της κανονικής κατανομής η μεταβλητή $\underline{x}$ ακολουθεί κατανομή της μορφής:

$$f(\underline{x}) = c_1\, \xi(x) \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu})\Sigma^{-1}(\underline{x}-\underline{\mu})\right\} \quad (2.3)$$

Για $\xi(\underline{x})=1$ έχουμε την κανονική κατανομή.

Σ' ότι ακολουθεί γίνεται χρήση της γραμμικής διαχωριστικής ανάλυσης και τα αποτελέσματα συγκρίνονται με αυτά της λογιστικής παλινδρόμησης.

2.4 Διαχωριστική Ανάλυση

Έστω $\{\Pi_i: 1 \le i \le k\}$, ένα σύνολο k πληθυσμών. Σε κάθε πληθυσμό Π_i αντιστοιχεί μια συνάρτηση πυκνότητας πιθανότητας $f_i(\underline{x})$, με πεδίο ορισμού το χώρο R^p, $\underline{x}$ είναι ένα px1 τυχαίο διάνυσμα.

Η αναγνώριση των διαφορών ανάμεσα σε k πληθυσμούς είναι απαραίτητη για την κατασκευή ενός κατάλληλου κανόνα ταξινόμησης του τυχαίου ατόμου σε έναν από τους k πληθυσμούς. Υποθέτουμε ότι σε κάθε άτομο αντιστοιχεί ένα διάνυσμα $\underline{x}$, px1 συνιστωσών. Ένας κανόνας ταξινόμησης θα κατατάξει το άτομο αυτό σε έναν από τους k πληθυσμούς $\{\Pi_i: i=1,2,...k\}$. Για συντομία θα αναφέρουμε ότι το διάνυσμα $\underline{x}$ κατατάσσεται στον πληθυσμό Π_i, αντί του ότι το άτομο κατατάσσεται στον πληθυσμό Π_i.

Θεωρούμε ότι ο p-διάστατος πραγματικός χώρος R^p, αποτελείται από k αμοιβαία αποκλειόμενες περιοχές $R_1, R_2, ... R_k$ έτσι ώστε:

$$\bigcup_{i=1}^{k} R_i = R^p$$

και

$$R_i \cap R_j = \varnothing, \quad i \neq j$$

Ένας κανόνας ταξινόμησης που αντιστοιχεί στην διαμέριση του R^p, ορίζεται ως εξής: *"ταξινόμησε το $\underline{x}$ στον πληθυσμό Π_i αν το $\underline{x} \in R_i$ για i=1,...,k."*

Κατά συνέπεια ο κανόνας ταξινόμησης μπορεί να ορισθεί από μία διαμέριση του δειγματόχωρου R^p σε k αμοιβαίως αποκλειόμενες περιοχές. Ένας κανόνας ταξινόμησης τοποθετεί το $\underline{x}$ στην περιοχή R_i, αν η πιθανότητα του $\underline{x}$ να προέρχεται από τον πληθυσμού Π_i είναι η μέγιστη, δηλαδή:

Αν, P[το άτομο με διάνυσμα $\underline{x}$ να προέρχεται από τον Π_i]= maximum ,

τότε το $\underline{x} \in R_i$

Ας θεωρήσουμε δυο σημεία $\underline{x}$, $\underline{y}$ στον p-διάστατο χώρο R^p ,όπου $\underline{x}$ ακολουθεί κατανομή με μέσο $\underline{\mu}_1$ και πίνακα διακύμανσης-συνδιακύμανσης Σ_1 και

$\underline{y}$ ακολουθεί κατανομή με μέσο $\underline{\mu}_2$ και πίνακα διακυμάνσεων-συνδιακυμάνσεων Σ_2, έτσι ώστε $\Sigma_1 = \Sigma_2 = \Sigma$.

Μεταξύ των σημείων $\underline{x}$, $\underline{y}$ ορίζεται η απόσταση MAHALANOBIS με μετρική Σ, ως η θετική τετραγωνική ρίζα της:

$$\delta^2(\underline{x}, \underline{y}) = (\underline{x} - \underline{y})' \Sigma^{-1} (\underline{x} - \underline{y})$$

όπου ο πίνακας Σ είναι ο πίνακας διακυμάνσεων-συνδιακυμάνσεων των $\underline{x}$, $\underline{y}$. [Γεωργιακώδης (1986)]

Η απόσταση MAHALANOBIS αποτελεί σημαντικότατο εργαλείο στη διαχωριστική ανάλυση.

2.4.1 Γραμμική διαχωριστική συνάρτηση του Fisher

Η διαχωριστική ανάλυση εφαρμόστηκε κατ' αρχήν από το Sir Ronald Fisher (1936), ο οποίος με τη βοήθεια της γραμμικής διαχωριστικής συνάρτησης που εισήγαγε, διεχώρησε τις τρεις ομάδες αγριόκρινων που μελετούσε με βάση το μήκος και το πλάτος των σέπαλων τους. Ο ορισμός γραμμική ή τετραγωνική διαχωριστική συνάρτηση είναι ανάλογος αυτού με την καμπύλη ελαχίστων τετραγώνων.

Σύμφωνα με το Fisher η γραμμική διαχωριστική συνάρτηση είναι ένας γραμμικός συνδυασμός των μεταβλητών που παρατηρήθηκαν και χρησιμοποιείται για το διαχωρισμό του αρχικού πληθυσμού. Κατά συνέπεια, η συνάρτηση αυτή μπορεί να περιγράψει και να ερμηνεύσει τις διαφορές μεταξύ πληθυσμών ή ομάδων του ιδίου πληθυσμού. Μορφή της γραμμικής διαχωριστικής συνάρτησης είναι:

$$Z(\underline{x}) = \beta_o + \underline{\beta}' \underline{x} \qquad (2.4)$$

όπου $\underline{\mathbf{x}}$ είναι το διάνυσμα του p-διάστατου χώρου, $\underline{\beta}'$ είναι διάνυσμα των μη τυποποιημένων συντελεστών της διαχωριστικής συνάρτησης και η σταθερά της διαχωριστικής συνάρτησης είναι το β_o.

Η προσέγγιση του Fisher μπορεί να εφαρμοσθεί κατ' αρχήν σε δύο πληθυσμούς Π_1 και Π_2. Έστω ο γραμμικός συνδυασμός $z = \underline{\beta}' \underline{x}$ για τον οποίο τα κεντροειδή δίνονται από :

$$E(z/\Pi_1) = \underline{\beta}' \underline{\mu}_1 \quad \text{και} \quad E(z/\Pi_2) = \underline{\beta}' \underline{\mu}_2 \quad \text{για κάθε πληθυσμό,}$$

και $Var(z)=\underline{\beta}'\Sigma\underline{\beta}$, έτσι ώστε $\Sigma_1=\Sigma_2=\Sigma$, όπου Σ_i (i=1, 2) οι πίνακες διακυμάνσεων-συνδιακυμάνσεων των πληθυσμών. Αν μεγιστοποιήσουμε το λόγο $\varphi=\frac{(\underline{\beta}'\underline{\mu}_1-\underline{\beta}'\underline{\mu}_2)^2}{\underline{\beta}'\Sigma\underline{\beta}}$, μπορούμε να υπολογίσουμε την τιμή του $\underline{\beta}$ με αποτέλεσμα να επιτευχθεί ο μέγιστος διαχωρισμός μεταξύ των κεντροειδών $\bar{z}_1$, $\bar{z}_2$ των δύο πληθυσμών.

Με παραγώγιση του λόγου φ ως προς $\underline{\beta}$ προκύπτει ότι :

$$\underline{\mu}_1-\underline{\mu}_2=\Sigma\underline{\beta}\left[\frac{\underline{\beta}'\underline{\mu}_1-\underline{\beta}'\underline{\mu}_2}{\underline{\beta}'\Sigma\underline{\beta}}\right]$$

Οι τιμές των $\underline{\mu}_1$, $\underline{\mu}_2$ και Σ μπορούν να εκτιμηθούν από τους δειγματικούς μέσους $\underline{\bar{x}}_1$, $\underline{\bar{x}}_2$ και S αντίστοιχα. Ο κανόνας ταξινόμησης που προκύπτει είναι ο ακόλουθος:

Αν $|z-\bar{z}_1|<|z-\bar{z}_2|$, τότε ταξινόμησε το διάνυσμα $\underline{x}$ στο Π_1.
Σε κάθε άλλη περίπτωση, ταξινόμησε το διάνυσμα $\underline{x}$ στο Π_2,
όπου

$$z=(\underline{\bar{x}}_1-\underline{\bar{x}}_2)'S^{-1}\underline{x},\quad \bar{z}_1=(\underline{\bar{x}}_1-\underline{\bar{x}}_2)'S^{-1}\underline{\bar{x}}_1 \quad \text{και} \quad \bar{z}_2=(\underline{\bar{x}}_1-\underline{\bar{x}}_2)'S^{-1}\underline{\bar{x}}_2$$

Το κρίτιμο σημείο του κανόνα ταξινόμησης ορίζεται από :

$$c=\frac{\bar{z}_1+\bar{z}_2}{2}$$

οπότε, *αν z > c, τότε ταξινόμησε το* $\underline{x}$ *στο* Π_1
αλλιώς ταξινόμησε το $\underline{x}$ *στο* Π_2

υποθέτοντας ότι $\bar{z}_1>\bar{z}_2$ χωρίς βλάβη της γενικότητας.

Αν θεωρήσουμε τη διαφορά $\bar{z}_1-\bar{z}_2$, τότε αποδεικνύεται ότι $\bar{z}_1-\bar{z}_2=D^2$ όπου D^2 το τετράγωνο της απόστασης Mahalanobis μεταξύ των δειγματικών μέσων. Αποδεικνύεται επίσης ότι $Var(z)=D^2$. Ο Fisher (1936) απέδειξε ότι ο γραμμικός συνδυασμός p-παρατηρήσεων, ο οποίος μεγιστοποιεί τη μεταξύ-δειγμάτων διακύμανση σε σχέση με την εντός των δειγμάτων διακύμανση είναι: $z=(\underline{\bar{x}}_1-\underline{\bar{x}}_2)'S^{-1}\underline{x}$,

Στα προηγούμενα καμιά υπόθεση δεν έγινε όσον αφορά τις κατανομές $f_1(\underline{x})$ και $f_2(\underline{x})$ των πληθυσμών Π_1 και Π_2 αντίστοιχα. Έτσι η προσέγγιση του Fisher μπορεί να θεωρηθεί ως μέθοδος που μπορεί να εφαρμοστεί ανεξάρτητα από το είδος της κατανομής των πληθυσμών, αλλά αριστοποιείται στην περίπτωση κανονικής κατανομής.

Ο κανόνας ταξινόμησης σκοπό έχει την ελαχιστοποίηση του σφάλματος λανθασμένης ταξινόμησης για όλες τις δυνατές κατατάξεις ενώ η διαχωριστική συνάρτηση σκοπό έχει το μέγιστο διαχωρισμό μεταξύ των ομάδων των ατόμων.

Υποθέτουμε ότι σε κάθε άτομο αντιστοιχεί ένα τυχαίο διάνυσμα $\underline{x}$ px1 συνιστωσών και έστω η συνάρτηση πυκνότητας πιθανότητας για τον πληθυσμό Π_1 είναι $f_1(\underline{x})$ και για τον Π_2 είναι $f_2(\underline{x})$.

Ένα μοντέλο ταξινόμησης μπορεί να προσδιορισθεί σαν μια διαμέριση του δειγματόχωρου R σε δύο αμοιβαίως αποκλειόμενες περιοχές R_1 και R_2 με το διαχωριστικό κανόνα ότι τα άτομα του πληθυσμού Π_1 αντιστοιχούν στην περιοχή R_1 και τα αντίστοιχα του Π_2 στην R_2.

Ο απλούστερος κανόνας ταξινόμησης είναι ο εξής:

«Το $\underline{x}$ θα ταξινομείται στο R_1 αν η πιθανότητα να προέρχεται από τον Π_1 είναι μεγαλύτερη από την πιθανότητα να προέρχεται από τον Π_2, και θα ταξινομείται στον R_2 αν η πιθανότητα να προέρχεται από τον Π_2 είναι μεγαλύτερη ή ίση από την πιθανότητα να προέρχεται από τον Π_1".

Επομένως προκύπτει ότι ο υπόχωρος R_1 είναι ένα σύνολο σημείων για τα οποία

$f_1(\underline{x}) > f_2(\underline{x})$ και ο R_2 ένα σύνολο σημείων για τα οποία

$f_1(\underline{x}) \leq f_2(\underline{x})$. Είναι προφανές ότι $R_1 \cap R_2 = \varnothing$ και $R_1 \cup R_2 = R$

Τα παραπάνω μπορούν να διατυπωθούν ως εξής:

Αν $f_1(\underline{x}) / f_2(\underline{x}) > 1$, τότε το $\underline{x} \in$ στον Π_1
αν $f_1(\underline{x}) / f_2(\underline{x}) \leq 1$, τότε το $\underline{x} \in$ στον Π_2

Ο κανόνας ταξινόμησης στηρίζεται στο λόγο πιθανοφάνειας (likelihood ratio) και αναπτύχθηκε από το Fisher κάτω από τις υποθέσεις κανονικότητας των $\underline{x}$ και ισότητας των πινάκων Σ_1 και Σ_2 .

Αποδεικνύεται ότι το παραπάνω κριτήριο ισοδυναμεί με την ελαχιστοποίηση της απόστασης του Mahalanobis

$$D = (\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu}) \qquad (2.5)$$

Με απλές πράξεις αποδεικνύεται ότι το κριτήριο του λόγου πιθανοφάνειας θεωρεί ότι $\underline{x}$ ανήκει στο R_1 όταν

$$(\underline{x} - (\mu_1 + \mu_2))' \Sigma^{-1} (\underline{x} - \frac{\underline{\mu_1} + \underline{\mu_2}}{2}) > 0 \qquad (2.6)$$

Όταν οι πυκνότητες πιθανοτήτων δεν είναι γνωστές τότε οι συντελεστές $\underline{\beta}'$ και β_0 της γραμμικής διαχωριστικής συνάρτησης είναι τέτοιοι ώστε να μεγιστοποιείται ο λόγος

$$\lambda = \frac{\beta' B \beta}{\beta' W \beta} \qquad (2.7)$$

όπου ο πίνακας B αναφέρεται σαν "Άθροισμα τετραγώνων μεταξύ των ομάδων" και ορίζεται

$$B = \begin{bmatrix} ssb_1 & scpb_{1,2} & . & scpb_{1,p} \\ scpb_{1,2} & ssb_2 & . & scpb_{2,p} \\ . & . & . & . \\ scpb_{1,p} & scpb_{2,p} & . & ssb_p \end{bmatrix} \qquad (2.8)$$

όπου ssb_i το "Άθροισμα τετραγώνων μεταξύ των ομάδων " για τη μεταβλητή x_i και το $scpb_{i,j}$ το " Άθροισμα τετραγώνων μεταξύ των ομάδων" για τις μεταβλητές x_i και x_j .

Αν x_i η i-οστη μεταβλητή, και G ο αριθμός των ομάδων, n_g ο αριθμός των παρατηρήσεων επί της iης μεταβλητής στην ομάδα g, $\bar{x}_{ig}$ είναι ο μέσος της iης μεταβλητής για την ομάδα g, και $\bar{x}_{i.}$ ο μέσος της μεταβλητής x_i στο σύνολο του δείγματος, τότε

και

$$ssb_i = \sum_{g=1}^{G} n_g (\bar{x}_{ig} - \bar{x}_{i.})^2$$

$$scpb_{ij} = \sum_{g=1}^{G} n_g (\bar{x}_{ig} - \bar{x}_{i.})(\bar{x}_{jg} - \bar{x}_{j.}) \qquad (2.9)$$

Ο Πίνακας W ορίζεται

$$W=\begin{bmatrix} ssw_1 & scpw_{1,2} & . & scpw_{1,p} \\ scpw_{1,2} & ssw_2 & . & scpw_{2,p} \\ . & . & . & . \\ scpw_{1,p} & scpw_{2,p} & . & ssw_p \end{bmatrix} \qquad (2.10)$$

όπου ssw_i ορίζεται ως το "pooled sum of squares within groups" για τη μεταβλητή x_i και το $scpw_{i,j}$ σαν το "pooled within-groups sum of products" για τις μεταβλητές x_i και x_j.

Πιο συγκεκριμένα:

$$SSW_i = \sum_{g=1}^{G}\sum_{l=1}^{n_g}(x_{il} - \overline{x_{ig}})^2$$

και

$$scpw_{ij} = \sum_{g=1}^{G}\sum_{l=1}^{n_g}(x_{il} - \overline{x_{ig}})(x_{jl} - \overline{x_{jg}}) \qquad (2.11)$$

Από την (2.7) και από τους ορισμούς των μητρών B και W προκύπτει ότι το λ δίνεται από την σχέση $\left|W^{-1}B - \lambda I\right| = 0$, με συνέπεια το λ να είναι η μέγιστη ιδιοτιμή της μήτρας $W^{-1}B$ και το διάνυσμα $\underline{\beta}$ το ιδιο διάνυσμα του παραπάνω πίνακα που αντιστοιχεί στη μέγιστη ιδιοτιμή λ.

Η σταθερά β_0 βρίσκεται από τη σχέση

$$\beta_0 = -(\beta_1\overline{x}_1 + \beta_2\overline{x}_2 + \ldots + \beta_p\overline{x}_p) \qquad (2.12)$$

Η πιο συνηθισμένη μέθοδος ταξινόμησης η οποία θα χρησιμοποιηθεί εδώ είναι η μέθοδος του κριτικού σημείου (cutoff-value point). Σύμφωνα με αυτή βρίσκουμε την τιμή $Z_0 = \dfrac{\overline{Z}_1 + \overline{Z}_2}{2}$ όπου $\overline{Z}_i$ είναι το κεντροειδές της ομάδας i. Έτσι αν για μονάδα του πληθυσμού $\underline{x}$ η διαχωριστική συνάρτηση έχει τιμή $Z(\underline{x})$, τότε η μονάδα ταξινομείται στον Π_1 αν $Z(\underline{x}) \leq Z_0$. Διαφορετικά ταξινομείται στο Π_2 [Mardia (1979), Lachenbruch (1975), Klecka (1984)].

2.4.2 Εφαρμογή διαχωριστικής ανάλυσης στο πιλοτικό δείγμα

Η ταξινόμηση του συνολικού δείγματος των 660 ατόμων σε υγιείς και ασθενείς έχει επιτευχθεί από τους θεράποντες ιατρούς με χρήση επεμβατικής μεθόδου, όπως η στεφανιογραφία.

Σκοπός της παρούσας εργασίας είναι ο προσδιορισμός εκείνης της διαχωριστικής τεχνικής η οποία θα ταξινομεί άτομα με συμπτώματα στεφανιαίας νόσου στις ομάδες υγιών ή ασθενών με όσο το δυνατόν υψηλότερα ποσοστά σωστής ταξινόμησης. Για το σκοπό αυτό επιχειρείται η εφαρμογή της διαχωριστικής ανάλυσης στο πιλοτικό δείγμα.

Οι διαχωριστικές συναρτήσεις που θα προκύψουν θα χρησιμοποιηθούν για την ταξινόμηση των ατόμων του συνολικού δείγματος σε υγιείς και ασθενείς και θα συγκριθεί η ικανότητα γενίκευσης των διαχωριστικών μοντέλων σε άγνωστα δεδομένα.

α) Ως ανεξάρτητες μεταβλητές θα χρησιμοποιηθούν οι 10 παράγοντες που ενοχοποιούνται για την εκδήλωση της νόσου, και η μεταβλητή EF. Ως εξαρτημένη μεταβλητή, η μεταβλητή DCODE (ασθενείς- υγιείς)

x_1=EF
x_2=FAMILY
x_3= HBP
x_4= LIPIDS
x_5= SEDENTARY
x_6= SMOKE
x_7=TYPE A
x_8=SEX
x_9= AGE
x_{10}= DIABETES
x_{11}=OBESITY

Ο πίνακας των συντελεστών της διαχωριστικής συνάρτησης που προκύπτει με τη βοήθεια του λογισμικού SPSS είναι ο ακόλουθος:

ΠΙΝΑΚΑΣ 2.9: Συντελεστές διαχωριστικής συνάρτησης (μη τυποποιημένοι)

Μεταβλητές	Συντελεστές
EF	-9.690
FAMILY	0.316
HBP	0.429
LIPIDS	0.551
SEDENTARY	0.883
SMOKE	0.707
TYPE A	0.382
SEX	-0.189
AGE	-0.030
DIABETES	0.028
OBESITY	0.185
ΣΤΑΘΕΡΟΣ ΟΡΟΣ	6.153

Η διαχωριστική συνάρτηση σύμφωνα με τους συντελεστές του πίνακα 2.9 είναι η ακόλουθη :

$$Z = -9.690x_1 + 0.316x_2 + 0.429x_3 + 0.551x_4 + 0.883x_5 + 0.707x_6 + 0.382x_7 - 0.189x_8 - 0.030x_9 + 0.028x_{10} + 0.185x_{11} + 6.153 \quad (2.13)$$

Τα κεντροειδή των δύο ομάδων έχουν τιμές αντίστοιχα :
Για την ομάδα των υγιών (DCODE =0). τιμή κεντροειδούς $\bar{Z}_0 = -0.956$
Για την ομάδα των ασθενών (DCODE =1) αντίστοιχη τιμή $\bar{Z}_1 = 0.956$.

Η κριτική τιμή είναι ίση με: $Z_c = \frac{\bar{Z}_0 + \bar{Z}_1}{2} = \frac{-0,956 + 0,956}{2} = 0$

Για τιμές της διαχωριστικής συνάρτησης μεγαλύτερες από 0 το άτομο χαρακτηρίζεται ως μη υγιές, ενώ για τιμές μικρότερες από 0, χαρακτηρίζεται ως υγιές.

Από τους τυποποιημένους συντελεστές, οι οποίοι χρησιμοποιούνται για την

εκτίμηση της διαχωριστικής δύναμης κάθε μιας των μεταβλητών που συμμετέχουν στη συνάρτηση, συμπεραίνουμε ότι η μεταβλητή x_1=EF έχει το μεγαλύτερο κατ' απόλυτη τιμή συντελεστή (-1.017), άρα και τη μεγαλύτερη διαχωριστική δύναμη. Αντιθέτως οι μεταβλητές x_{11}=Obesity, x_{10}= DIABETES, x_8=SEX με τυποποιημένους συντελεστές αντίστοιχα 0.092, 0.011, -0.072 έχουν μικρή συνεισφορά στην διαχωριστική ικανότητα του μοντέλου.

ΠΙΝΑΚΑΣ 2.10: ΤΥΠΟΠΟΙΗΜΕΝΟΙ ΣΥΝΤΕΛΕΣΤΕΣ

Μεταβλητές	**Συνάρτηση 1**
SEX	-0.072
EF	-1.017
LIPIDS	0.271
HBR	0.214
SMOKE	0.312
DIABETES	0.011
OBESITY	0.092
FAMILY	0.128
SEDENTARY	0.342
TYPE A	0.118
AGE	-0.225

Από τον πίνακα ιδιοτιμών προκύπτει ότι μια και μόνο διαχωριστική συνάρτηση απαιτείται για την ανάλυση.

ΠΙΝΑΚΑΣ 2.11: Ιδιοτιμές

Συνάρτηση	Ιδιοτιμή	% της διασποράς	Αθροιστική (%)	Κανονική συσχέτιση
(2.13)	0.926[a]	100.0	100.0	0.693

β) Για τους λόγους που αναφέρθηκαν προηγουμένως (πίνακας 2.10), στο δείγμα των 160 ατόμων εφαρμόσθηκε διαχωριστική ανάλυση 8 μεταβλητών. Οι μεταβλητές που χρησιμοποιήθηκαν ήταν οι του μοντέλου 2.5, εκτός των μεταβλητών x_8 ,x_{10},x_{11}.

Ο πίνακας των τυποποιημένων συντελεστών που χρησιμοποιήθηκαν για την εκτίμηση της διαχωριστικής δύναμης των μεταβλητών που συμμετείχαν, ανέδειξαν την μεταβλητή EF ως ισχυρό διαχωριστικό παράγοντα (-1.029).

Η μεταβλητή AGE αφαιρέθηκε και η διαχωριστική ανάλυση εφαρμόστηκε εκ νέου, γιατί σκοπός της έρευνας είναι η εύρεση μιας διαχωριστικής συνάρτησης με την μεγαλύτερη απόδοση αλλά τον μικρότερο κατά το δυνατόν αριθμό μεταβλητών

Οι μεταβλητές που χρησιμοποιήθηκαν είναι οι ακόλουθες:

x_1 = EF
x_2 = FAMILY
x_3 = HBP
x_4 = LIPIDS
x_5 = SEDENTARY
x_6 = SMOKE
x_7 = TYPE A

Σαν εξαρτημένη μεταβλητή (διαχωριστικός παράγοντας) θεωρήθηκε η Z=DCODE. Η διαχωριστική συνάρτηση σύμφωνα με τους συντελεστές από τον Πίνακα 2.12 είναι:

$$Z = -9{,}333x_1 + 0{,}343x_2 + 0{,}393x_3 + 0{,}595x_4 + 0{,}88x_5 + 0{,}766x_6 + 0{,}481x_7 + 4{,}269$$

(2.14)

Στον πίνακα 2.12 δίνονται οι τιμές των μη τυποποιημένων συντελεστών της διαχωριστικής συνάρτησης:

ΠΙΝΑΚΑΣ 2.12: ΣΥΝΤΕΛΕΣΤΕΣ ΜΗ ΤΥΠΟΠΟΙΗΜΕΝΟΙ

Μεταβλητές	**Συνάρτηση (2.5)**
EF	-9.333
FAMILY	0.343
HBP	0.393
LIPIDS	0.595
SEDENTARY	0.880
SMOKE	0.766
TYPE A	0.481
ΣΤΑΘΕΡΟΣ ΟΡΟΣ	4.269

Οι τιμές της διαχωριστικής συνάρτησης στα κέντρα των ομάδων είναι:

για την ομάδα των υγιών ατόμων (DCODE=0) , $\bar{Z}_0 = -0,933$ ενώ για την ομάδα των νοσούντων (DCODE=1) είναι $\bar{Z}_1 = 0,933$.

Η κριτική τιμή (critical value) είναι $Z_c = \frac{\bar{Z}_0 + \bar{Z}_1}{2} = 0$.

Για θετικές τιμές της διαχωριστικής συνάρτησης το άτομο χαρακτηρίζεται ως μη υγιές, ενώ για αρνητικές τιμές, χαρακτηρίζεται ως υγιές.

από τον πίνακα 2.13 προκύπτει ότι μια διαχωριστική συνάρτηση απαιτείται για την ανάλυση

ΠΙΝΑΚΑΣ 2.13 Ιδιοτιμές

Συνάρ-τηση	**Ιδιοτιμή**	**%της διασποράς**	**Αθροιστική (%)**	**Κανονική συσχέ-τιση**
(2.14)	0.881[a]	100.0	100.0	0.684

- ΑΠΟΔΟΣΗ ΤΩΝ ΔΙΑΧΩΡΙΣΤΙΚΩΝ ΣΥΝΑΡΤΗΣΕΩΝ.

Η διαχωριστική ικανότητα μιας ανεξάρτητης μεταβλητής, καθώς επίσης και της συνάρτησης Z, ελέγχεται μέσω του δείκτη Wilks' Lambda Λ ο οποίος ορίζεται από το πηλίκο $\Lambda = \frac{|W|}{|T|}$, όπου ο πίνακας T ορίζεται ως T=W+B και οι πίνακες B και W ορίζονται από τις σχέσεις (2.8) και (2.10) αντιστοίχως .

Ο έλεγχος της διαχωριστικής ικανότητας μιας μεταβλητής γίνεται με τη βοήθεια του F-test. Αποδεικνύεται ότι:

$$F = \frac{(1-\Lambda)}{\Lambda}\frac{(n_1+n_2-p-1)}{p} \qquad (2.15)$$

όπου n_1, n_2 είναι ο αριθμός των παρατηρήσεων για την πρώτη και δεύτερη ομάδα αντίστοιχα, και p είναι ο αριθμός των μεταβλητών για τις οποίες υπολογίζεται το Λ. Οι βαθμοί ελευθερίας είναι p και n_1+n_2-p-1. Για τη διαχωριστική ικανότητα της συνάρτησης Z, χρησιμοποιείται χ^2-test. Πιο συγκεκριμένα ελέγχουμε τις υποθέσεις:

$$H_0 \quad \underline{\mu}_1 = \underline{\mu}_2$$

$$H_\alpha \quad \underline{\mu}_1 \neq \underline{\mu}_2$$

Αποδεικνύεται ότι:

$\chi^2 = -[n-1-(p+G)/2]\ln\Lambda$, με $p(G-1)$ βαθμούς ελευθερίας και G αριθμό ομάδων . Στη συγκεκριμένη περίπτωση G=2.

α) Η διαχωριστική συνάρτηση (2.13) ταξινόμησε σωστά το 82,5% των περιπτώσεων και από τον πίνακα 2.11 παρατηρήθηκε ότι η κανονική συσχέτιση είναι *CR=0.693* .Αυτό σημαίνει ότι το $CR^2\%=48\%$ της μεταβλητότητας των δύο ομάδων ερμηνεύεται από την διαχωριστική συνάρτηση. Η κανονική συσχέτιση μετρά τη συσχέτιση μεταξύ των τιμών της διαχωριστικής συνάρτησης και των παρατηρηθεισών τιμών των ατόμων των ομάδων.

Στη συγκεκριμένη περίπτωση όπου το ζητούμενο είναι ο διαχωρισμός σε δύο ομάδες (υγιείς- ασθενείς), η κανονική συσχέτιση είναι η συσχέτιση κατά Pearson μεταξύ των τιμών της διαχωριστικής συνάρτησης και των τιμών των ατόμων των ομάδων που έχουν καταχωρηθεί ως υγιείς (0) ή ασθενείς(1) .

Το επίπεδο σημαντικότητας α=0.000 στον Πίνακα 2.14 δηλώνει ότι η διαχωριστική συνάρτηση μπορεί να κάνει διάκριση μεταξύ των δύο κατηγοριών.

ΠΙΝΑΚΑΣ 2.14

Έλεγχος συναρτήσεων	Wilks' Lambda	χ^2-τετράγωνο	βαθμοί ελευθερίας	σημαντικότητα
(2.13)	0,519	99,917	11	0,000

Αναλυτικά τα αποτελέσματα ταξινόμησης είναι τα ακόλουθα:

ΠΙΝΑΚΑΣ 2.15

Αποτελέσματα ταξινόμησης[a]

			Προβλεπόμενη Ταξινόμηση		
		DCODE	0	1	Σύνολο
Αρχική	Συχνότητα	0	72	8	80
		1	20	60	80
	%	0	90,0	10,0	100,0
		1	25,0	75,0	100,0

a. **82,5% των περιπτώσεων ταξινομήθηκε σωστά.**

Αποτελεσματικότητα = (132/160) x100% = 82.5%
Ειδικότητα = (72/80)x100% = 90%
Ευαισθησία = (60/80)x100% = 75%
Δείκτης YOUDEN = Ειδικότητα + Ευαισθησία – 100% = 65%

Η ειδικότητα και η ευαισθησία είναι εκφράσεις της διαχωριστικής ικανότητας και η καθεμιά τους συνδέεται με έναν από τους δύο πληθυσμούς. Η αποτελεσματικότητα όμως καθώς και ο δείκτης Youden είναι εκφράσεις της συνολικής διαχωριστικής ικανότητας του μοντέλου διότι σχετίζονται και με τους δύο πληθυσμούς.

Είναι κοινά αποδεκτό ότι ο δείκτης Youden είναι πιο αξιόπιστο μέτρο από την αποτελεσματικότητα, γιατί λαμβάνει υπόψη την διαφορά του μεγέθους των δειγμάτων που προέρχονται από τους πληθυσμούς Π_1 και Π_2.
[Γεωργιακώδης (1986)].

β) Από τον Πίνακα 2.16 βλέπουμε ότι η διαχωριστική συνάρτηση (2.14) ταξινόμησε σωστά το 82,5% των περιπτώσεων. Από τον Πίνακα 2.13 έχουμε ότι η κανονική συσχέτιση είναι $CR = 0{,}684$ και συνεπώς το $CR^2\% = 46{,}7\%$ της μεταβλητότητας των δύο ομάδων ερμηνεύεται από τη διαχωριστική συνάρτηση.

ΠΙΝΑΚΑΣ 2.16

Αποτελέσματα ταξινόμησης[a]

			Προβλεπόμενη Ταξινόμηση		
		DCODE	0	1	Σύνολο
Αρχική	Συχνότητα	0	72	8	80
		1	20	60	80
	%	0	90,0	10,0	100,0
		1	25,0	75,0	100,0

a. **82,5% των περιπτώσεων ταξινομήθηκε σωστά.**

Αποτελεσματικότητα = (132/160) x100% = 82.5%
Ειδικότητα = (72/80)x100% = 90%
Ευαισθησία = (60/80)x100% = 75%
Δείκτης YOUDEN = Ειδικότητα + Ευαισθησία – 100% = 65%

Το επίπεδο σημαντικότητας α=0.000 στον Πίνακα 2.17 δηλώνει ότι η διαχωριστική συνάρτηση μπορεί να κάνει διάκριση μεταξύ των δύο κατηγοριών.

ΠΙΝΑΚΑΣ 2.17

Έλεγχος συναρτήσεων	Wilks' Lambda	χ^2-τετράγωνο	βαθμοί ελευθερίας	σημαντικότητα
(2.14)	0,531	97,653	7	0,000

Παρατηρούμε ότι οι διαχωριστικές συναρτήσεις (2.13) και (2.14) έχουν την ίδια διαχωριστική ικανότητα .Στη διαχωριστική συνάρτηση (2.14) το πλήθος των μεταβλητών που χρησιμοποιούνται είναι μικρότερο, με αποτέλεσμα το μοντέλο να είναι οικονομικότερο, άρα καταλληλότερο έναντι του (2.13)

2.4.3 Βήμα προς βήμα διαχωριστική ανάλυση.

Στη συνέχεια θα επιχειρηθεί η απλοποίηση της συνάρτησης (2.6) με τη χρήση της βήμα-βήμα διαχωριστικής ανάλυσης. Σύμφωνα με αυτήν την ανάλυση, μία μεταβλητή εισέρχεται στο μοντέλο αν αυξάνει τη διαχωριστική ικανότητα της συνάρτησης σε επίπεδο σημαντικότητας α=0,05, και αφαιρείται από το μοντέλο αν μειώνει τη διαχωριστική ικανότητα της συνάρτησης σε επίπεδο α=0.01.

Στον πίνακα 2.18 παρατηρούμε ότι η διαδικασία σταματά στο τέταρτο βήμα και οι τέσσερις πιο σημαντικές μεταβλητές που εισέρχονται στο μοντέλο είναι:

EF,
SEDENTARY,
SMOKE,
LIPIDS.

ΠΙΝΑΚΑΣ 2.18

				Βαθμοί ελευθερίας		α
ΒΗΜΑ	ΕΙΣΕΡ-ΧΟΝΤΑΙ	Λ	F	1	2	0,000
1	EF	0,663	80,376	1	158	0,000
2	SMOKE	0,604	51,46	2	157	0,000
3	SEDENTARY	0,578	38,034	3	156	0,000
4	LIPIDS	0,548	31,996	4	155	0,000

Χρησιμοποιώντας το στατιστικό πρόγραμμα SPSS κατασκευάζουμε τη διαχωριστική συνάρτηση:

Z=-9,267(EF)+0,976(SEDENTARY) +0,809(SMOKE) +0,701(LIPIDS)+ 4,457 (2.15)

Οι τιμές της διαχωριστικής συνάρτησης στα κέντρα των ομάδων δίνονται στον πίνακα 2.19 Η κριτική τιμή σύμφωνα με την οποία θα ταξινομούμε κάθε περίπτωση είναι πάλι 0.

ΠΙΝΑΚΑΣ 2.19: Τιμές κεντροειδών στις δύο ομάδες

DCODE	Συνάρτηση (2.15)
0	-0,903
1	0,903

Τα αποτελέσματα της βήμα προς βήμα διαχωριστικής ανάλυσης που εφαρμόστηκε στο συγκεκριμένο δείγμα καθώς και ο πλήρης πίνακας ταξινόμησης παρατίθενται στους παρακάτω πίνακες 2.20, 2.21 και 2.22.

ΠΙΝΑΚΑΣ 2.20

Έλεγχος της διαχωριστικής συνάρτησης με την Wilks' Lambda στατιστική

Συνάρτηση	Λ	X^2	βαθμοί ελευ-	α
(2.15)	0,548	93,907	4	0,00

ΠΙΝΑΚΑΣ 2.21

Συνάρτηση	Ιδιοτιμή[α]	% της διασποράς	Αθροιστική %	Κανονική Συσχέτιση
(2.15)	0,826	100	100	0,673

[α] Μία διαχωριστική συνάρτηση χρησιμοποιήθηκε στην ανάλυση.

ΠΙΝΑΚΑΣ 2.22: Πίνακας ταξινόμησης της DCODE

DCODE			Προβλεπόμενη ταξινόμηση		
			0	1	Σύνολο
Αρχική	Συχνότητα	0	70	10	80
		1	17	63	80
	%	0	87,5	12,5	100
		1	21,3	78,8	100

Αποτελεσματικότητα = (133/160) x100% = 83,1%
Ειδικότητα = (70/80)x100% = 87,5%
Ευαισθησία = (63/80)x100% = 78,8%
Δείκτης YOUDEN = Ειδικότητα + Ευαισθησία – 100% = 66,3%

Συγκρίνοντας τις συναρτήσεις (2.14) και (2.15) παρατηρούμε ότι η δεύτερη ταξινομεί σωστά μεγαλύτερο ποσοστό του δείγματος με την χρήση τεσσάρων μόνο μεταβλητών. Η συνάρτηση (2.15) έχει μικρότερο ποσοστό ειδικότητας (87,5% έναντι 90% της συνάρτησης 2.14), αλλά μεγαλύτερο ποσοστό ευαισθησίας. Επίσης η τιμή του δείκτη YOUDEN είναι μεγαλύτερη. Συμπερασματικά , θα μπορούσαμε να ισχυρισθούμε ότι η διαχωριστική συνάρτηση (2.15) έχει καλύτερη απόδοση.

Η τελική αξιολόγηση των μοντέλων (2.14) και (2.15) θα γίνει με την εφαρμογή τους πάνω στο συνολικό δείγμα των 660 ατόμων.

2.5 Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση, όπως και η διαχωριστική ανάλυση, είναι διαχωριστικές τεχνικές, οι οποίες χρησιμοποιούνται για την πρόβλεψη της παρουσίας ή απουσίας κάποιου χαρακτηριστικού στα άτομα του αρχικού πληθυσμού και την ταξινόμηση των ατόμων αυτών σε διαφορετικές ομάδες, με βάση

ένα σύνολο ερμηνευτικών μεταβλητών. Η λογιστική παλινδρόμηση είναι τεχνική παρόμοια με την γραμμική παλινδρόμηση, αποδεικνύεται όμως αποτελεσματικότερη σε μοντέλα όπου η εξαρτημένη μεταβλητή είναι διχοτόμος.

Στην περίπτωση μιας ανεξάρτητης μεταβλητής (x), το λογιστικό μοντέλο περιγράφεται από τη συνάρτηση:

$$p = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}} = \frac{1}{1+e^{-(b_0+b_1x)}} \qquad (2.16)$$

όπου b_0, b_1 είναι συντελεστές που εκτιμώνται από τα δεδομένα. Για n ανεξάρτητες μεταβλητές το μοντέλο μπορεί να γραφεί ως

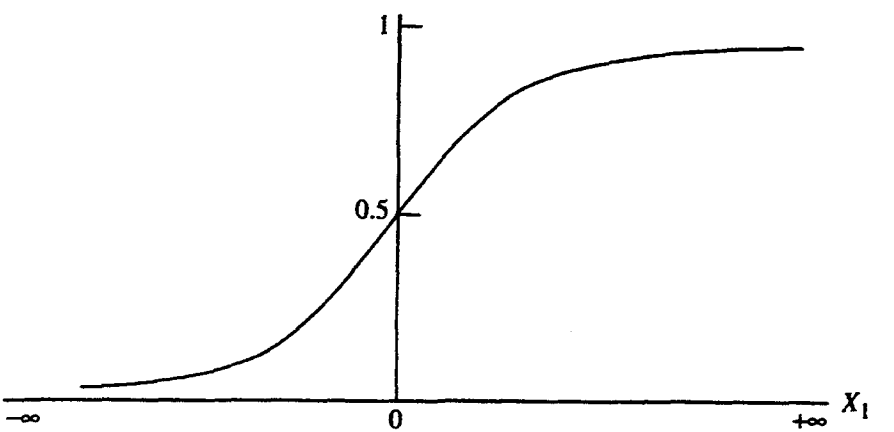
$$p = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} \qquad (2.17)$$

όπου z είναι γραμμικός μετασχηματισμός των x_i ,i=1,2,...,n , άρα

$$z = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n = b_0 + \underline{b}'\underline{x}$$

με $\underline{b}' = (b_1, b_2, \ldots, b_n)$ το διάνυσμα των συντελεστών.

Το λογιστικό μοντέλο έχει το ακόλουθο γράφημα:



Από το παραπάνω γράφημα παρατηρείται η μη- ύπαρξη γραμμικότητας μεταξύ της ανεξάρτητης και της εξαρτημένης μεταβλητής p, καθώς και ότι οι τιμές της συνάρτησης p κυμαίνονται από 0 μέχρι 1 για $z\in(-\infty,+\infty)$. Η τελευταία παρατήρηση μας δίνει τη δυνατότητα να θεωρήσουμε την p ως μια συνάρτηση πυκνότητας πιθανότητας.

Όταν πληρούνται οι συνθήκες της γραμμικής παλινδρόμησης, οι συντελεστές του γραμμικού μοντέλου εκτιμώνται συνήθως με την μέθοδο των ελαχίστων τετραγώνων. Στη λογιστική παλινδρόμηση όπου η εξαρτημένη μεταβλητή είναι διχοτόμος, οι συντελεστές εκτιμώνται με την μέθοδο της μέγιστης πιθανοφάνειας, και επιλέγονται έτσι ώστε να μεγιστοποιείται η πιθανότητα πραγματοποίησης των παρατηρηθέντων αποτελεσμάτων.

Η γραμμική παλινδρόμηση απαιτεί την κανονικότητα της εξαρτημένης μεταβλητής. Ο περιορισμός αυτός αίρεται με τη χρήση των γενικευμένων γραμμικών μοντέλων στα οποία η εξαρτημένη μεταβλητή μπορεί να ακολουθεί οποιαδήποτε κατανομή εκθετικής οικογένειας δηλαδή η συνάρτηση πυκνότητας πιθανότητας μπορεί να γραφεί στη μορφή

$$f(y,p) = h(y)e^{C(p)\cdot y + k(p)} \qquad (2.18)$$

[Lehmann (1983)]

Η εκθετική οικογένεια κατανομών αποδεικνύεται ότι είναι πλήρης οικογένεια και περιλαμβάνει πολλά είδη κατανομών όπως: δυωνυμική, κανονική, γεωμετρική, εκθετική, Poisson [Silvey(1970)].

Η λογιστική παλινδρόμηση αποτελεί περίπτωση του γενικευμένου γραμμικού μοντέλου στο οποίο η μάζα πιθανότητας δίνεται από την συνάρτηση μάζας πιθανότητας της διωνυμικής. Αν λοιπόν υποθέσουμε ότι

$$p = \sum b_i x_i \qquad (2.19)$$

τότε το δεύτερο μέλος πρέπει να βρίσκεται μεταξύ 0 και1. Ο περιορισμός αυτός ικανοποιείται αν εκφράσουμε την 2.19 ως

$$p = \frac{e^{b0+b1x1+.....bnxn}}{1+e^{b0+b1x1+.....bnxn}} \qquad (2.20)$$

από την οποία προκύπτει η:

$$\ln\frac{p}{1-p} = b_0 + \sum_{i=1}^{n} b_i x_i = b_0 + \underline{b}'\underline{x} \qquad (2.20^{\alpha})$$

Η τελευταία έκφραση δίνει τον λογάριθμο του λόγου υπεροχής ως γραμμικό συνδυασμό των x_i και είναι γνωστή ως λογιστικό μοντέλο. Η εκτίμηση των παραμέτρων b_i, i=1,2,...,n του λογιστικού μοντέλου μπορεί να επιτευχθεί και με την γνωστή μέθοδο των ελαχίστων τετραγώνων, σύμφωνα με την οποία επιχειρούμε να ελαχιστοποιήσουμε το άθροισμα

$$\sum_i (y_i - \frac{e^{b0+b1x1+.....bnxn}}{1+e^{b0+b1x1+.....bnxn}})^2 \qquad (2.21)$$

Στη παρούσα εργασία, συμβολίζουμε με p=Pr(DCODE=0) την πιθανότητα ενός ατόμου με διάνυσμα $\underline{x}$ να ταξινομηθεί στην ομάδα των υγιών, και με $\underline{b}'$ το διάνυσμα των συντελεστών. Αν θέσουμε ως Y την εξαρτημένη μεταβλητή DCODE τότε οι τιμές της μεταβλητής Y είναι:

$$Y = \begin{cases} 0, \alpha\nu\ DCODE = 0 \\ 1, \alpha\nu\ DCODE = 1 \end{cases}$$

και η έκφραση $p = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}} = \frac{1}{1+e^{-(b_0+b_1x)}}$,

είναι η δεσμευμένη πιθανότητα p=P(Y=0/ x), κατά συνέπεια 1-p είναι η δεσμευμένη πιθανότητα του Y=1 δοθέντος του x. Έτσι ένας κατάλληλος τρόπος για να εκφράσουμε την συνεισφορά των ζευγών (x_i,y_i) στη συνάρτηση πιθανοφάνειας (likelihood) είναι μέσω του όρου:

$$\zeta(x_i) = p(x_i)^{y_i}\,[1-p(x_i)]^{1-y_i}$$

Με την προϋπόθεση ότι οι παρατηρήσεις είναι ανεξάρτητες, η συνάρτηση πιθανοφάνειας $\ell(b)$ είναι το γινόμενο των όρων που δίνονται στην παραπάνω έκφραση:

$$\ell(b) = \prod_{i=1}^{n} \zeta(x_i).$$

Υπενθυμίζεται ότι $\underline{b}' = (b_0, b_1)$ είναι το διάνυσμα των συντελεστών.

Οι συντελεστές υπολογίστηκαν με τη μέθοδο της μέγιστης πιθανοφάνειας (maximum likelihood). Είναι δηλαδή οι τιμές εκείνες που μεγιστοποιούν τη λογαριθμική συνάρτηση πιθανοφάνειας (loglikelihood), ή ισοδύναμα ελαχιστοποιούν την συνάρτηση (-2)[loglikelihood]. Ένας τρόπος θα ήταν να λυθεί το σύστημα που λαμβάνεται όταν θέσουμε την παράγωγο της λογαριθμικής συνάρτησης πιθανοφάνειας ίσο με μηδέν. Ο τρόπος αυτός δε χρησιμοποιείται στην πράξη.

Αντί αυτού χρησιμοποιείται η επαναληπτική μέθοδος του Newton - Raphson, η οποία τερματίζεται όταν η (-2).[loglikelihood] ελαττώνεται κατά ποσοστό μικρότερου του 0,1%. Αυτό συμβαίνει μετά από 5 επαναλήψεις.

Ο Efron (1975) μελέτησε τις ασυμπτωτικές ιδιότητες των εκτιμητών με τη χρήση της λογιστικής παλινδρόμησης. Ο Menard (1980) δίνει μια πλήρη περιγραφή της λογιστικής παλινδρόμησης και των εφαρμογών της. Οι Brenn και Arnesen (1985) μελετώντας τη θνησιμότητα σε δείγμα 6595 στεφανιαίων ασθενών ανδρών σε διάστημα 9 ετών, συνέκριναν τα αποτελέσματα διαχωριστικής ανάλυσης, λογιστικής παλινδρόμησης και παλινδρόμησης κατά Cox. Κατέληξαν στο συμπέρασμα ότι η διαχωριστική ανάλυση είναι καταλληλότερη για προκαταρτική ή βήμα-προς-βήμα ανάλυση, ενώ η μέθοδος του Cox προτείνεται σε κάθε άλλη περίπτωση.

Θα επιχειρήσουμε να περιγράψουμε το μοντέλο της λογιστικής παλινδρόμησης, με τρόπο ανάλογο με αυτό της διαχωριστικής ανάλυσης χρησιμοποιώντας τις ίδιες μεταβλητές.

Στο πρώτο βήμα επιχειρείται η εισαγωγή των 11 μεταβλητών που χρησιμοποιήθηκαν στην προηγούμενη παράγραφο.

Στη λογιστική παλινδρόμηση διευκρινίζεται η φύση των μεταβλητών. από τις 11 μεταβλητές που χρησιμοποιήθηκαν στη διαχωριστική ανάλυση, οι μεταβλητές EF και AGE είναι συνεχείς. Οι υπόλοιπες μεταβλητές είναι δυαδικές με τιμές 0 και 1, ονομάζονται "dummy variables" και οι τιμές τους δεν έχουν σημασία μετρήσιμου μεγέθους. Στη συγκεκριμένη περίπτωση οι τιμές αυτές δηλώνουν απουσία χαρακτηριστικού (τιμή 0) ή παρουσία χαρακτηριστικού (τιμή 1), εκτός από τη μεταβλητή SEX που δηλώνει το φύλο.

Στον πίνακα 2.23 που ακολουθεί, παρουσιάζονται τα αποτελέσματα της λογιστικής παλινδρόμησης.

ΠΙΝΑΚΑΣ 2.23

Μεταβλητή	b	τυπικό σφάλ-μα	Wald τιμή	βαθμοί ελευθερίας	α	R	Exp(b)
AGE	-0,0517	0,0399	1,674	1	0,1957	0,000	0,9496
DIABETES	0,0457	0,6842	0,0045	1	0,9468	0,000	1,0467
EF	-24,1071	4,3081	31,3123	1	0,0000	-0,3635	0,0000
FAMILY	-1,38888	0,6833	4,1308	1	0,0421	-0,098	0,2494
HBP	-1,3683	0,5751	5,6606	1	0,174	-0,1285	0,2545
LIPIDS	-1,0827	0,5207	4,3236	1	0,0376	-0,1024	0,3387
OBECITY	-0,8003	0,5238	2,3345	1	0,1265	-0,388	0,4492
SEX	0,2163	0,8273	0,683	1	0,7938	0,0000	1,2414
SED/TAR	-2,0292	0,7924	6,5582	1	0,0104	-0,1434	0,1314
SMOKE	-1,3424	0,6304	4,5348	1	0,0332	-0,1069	0,2612
TYPEA	-1,9158	1,1611	2,7222	1	0,0990	-0,0571	0,1472
Σταθερός όρος	23,8683	4,6378	26,4860	1	0,0000		

Η πρώτη στήλη περιέχει τις μεταβλητές, η δεύτερη στήλη την εκτίμηση των συντελεστών, η τρίτη το τυπικό σφάλμα, η τέταρτη την χ^2-τιμή του Wald με την οποία ελέγχεται η υπόθεση μηδενισμού του συντελεστή της μεταβλητής της πρώτης στήλης. Παρατηρούμε ότι μόνο οι συντελεστές των μεταβλητών EF, FAMILY, LIPIDS,SEDENTARY και SMOKE φαίνονται να είναι στατιστικώς σημαντικά διάφοροι του μηδενός, σύμφωνα με το επίπεδο σημαντικότητας του

test που περιέχεται στην έκτη στήλη. Η πέμπτη στήλη αναφέρεται στους βαθμούς ελευθερίας του χ^2-test, η έβδομη περιέχει τον συντελεστή συσχέτισης της μεταβλητής με την εξαρτημένη μεταβλητή, και η τελευταία στήλη περιέχει την έκφραση Exp(b) που εκφράζει τον λόγο υπεροχής για τις μεταβλητές της πρώτης στήλης .

Από τον Πίνακα 2.23 βρίσκουμε ότι το λογιστικό μοντέλο είναι :

$$\ln(p/(1-p)) = 23.8683 - 0.0517*X_1 + 0.0457*X_2 - 24.1071*X_3 - 1.3888*X_4 - 1.3683*X_5 - 1.0827*X_6 - 0,8003*X_7 + 0.2163*X_8 - 2.0292*X_9 - 1.3424*X_{10} - 1,9158*X_{11} \quad (2.22)$$

Από τον πίνακα 2.24 παρατηρείται ότι η αποτελεσματικότητα του λογιστικού μοντέλου (2.22), δεν διαφέρει από αυτήν των διαχωριστικών συναρτήσεων (2.13) και (2.14). Αν και η τιμή του δείκτη YOUDEN είναι μικρότερη από τις τιμές των δεικτών των προηγουμένων διαχωριστικών συναρτήσεων, το ποσοστό ευαισθησίας είναι μεγαλύτερο.

ΠΙΝΑΚΑΣ 2.24

		Προβλεπόμενη ταξινόμηση		Ποσοστό σωστής ταξινόμησης
Παρατηρούμενη ταξινόμηση		0	1	
DCODE	0	66	14	82,5%
	1	14	66	82,5%

Αποτελεσματικότητα = (132/160) x100% = 82,5%
Ειδικότητα = (66/80)x100% = 82,5%
Ευαισθησία = (66/80)x100% = 80,0%
Δείκτης YOUDEN = Ειδικότητα + Ευαισθησία – 100% = 62,5%

Η καλή προσαρμογή του λογιστικού μοντέλου στα εμπειρικά δεδομένα είναι απαραίτητη και για το λόγο αυτό υπάρχουν και τα αντίστοιχα κριτήρια.

Η διαδικασία επιλογής της παραμέτρου Θ της συνάρτησης συχνότητας $f(x, \Theta)$ καταλήγει στην τιμή $\hat{\Theta}$, που μεγιστοποιεί την πιθανότητα εμφάνισης

του συγκεκριμένου δείγματος $x_1, x_2, \ldots, x_n$, το οποίο έχει ήδη παρατηρηθεί. Ως κριτήριο καλής προσαρμογής του εκτιμηθέντος μοντέλου στα εμπειρικά δεδομένα χρησιμοποιείται η τιμή της λογαριθμικής συνάρτησης πιθανοφάνειας (loglikelihood), ή ισοδύναμα της συνάρτησης: $-2LL = -2 \cdot \text{loglikelihood}$.

Στο συγκεκριμένο δείγμα, για το λογιστικό μοντέλο που περιέχει μόνο τη σταθερά, η τιμή της συνάρτησης $-2LL$ είναι 221,8071, ενώ για το μοντέλο των 11 μεταβλητών, η τιμή της συνάρτησης $-2LL$ =111,306 είναι σημαντικά μικρότερη από την προηγούμενη.

Τα στατιστικά των Cox & Snell (R^2) καθώς και του Nagelkerke ($\tilde{R}^2$) προσπαθούν να ποσοτικοποιήσουν το ποσοστό της διακύμανσης που ερμηνεύεται από το λογιστικό μοντέλο και ορίζονται ως εξής:

$$\text{Cox \& Snell: } R^2 = 1 - \left[\frac{L(0)}{L(B)}\right]^{2/N}$$

όπου $L(0)$ είναι η τιμή της συνάρτησης πιθανοφάνειας για το μοντέλο που περιέχει μόνο τη σταθερά, ενώ $L(B)$ είναι η τιμή της συνάρτησης πιθανοφάνειας για το μοντέλο που εξετάζουμε, και N είναι το μέγεθος του δείγματος
Αντίστοιχα :

$$\text{Nagelkerke: } \tilde{R}^2 = \frac{R^2}{R^2_{MAX}} \text{ , όπου } R^2_{MAX} = 1 - [L(0)]^{2/N} .$$

Στο συγκεκριμένο δείγμα, από την τιμή του Nagelkerke $\tilde{R}^2 = 0{,}665$, μπορούμε να συμπεράνουμε ότι το 66,5% της διακύμανσης της εξαρτημένης μεταβλητής ερμηνεύεται από το λογιστικό μοντέλο. Αξίζει να σημειωθεί επίσης ότι η τιμή του κριτηρίου καλής προσαρμογής των Hosmer και Lemeshow – $\chi^2 = 5{,}93$ σε επίπεδο σημαντικότητας 0,655 με 8 βαθμούς ελευθερίας- μας οδηγεί στο συμπέρασμα ότι το μοντέλο φαίνεται να προσεγγίζει ικανοποιητικά τα εμπειρικά δεδομένα.

2.5.1 Βήμα προς βήμα λογιστική παλινδρόμηση

Όπως φαίνεται από τον πίνακα 2.23 δεν είναι όλες οι μεταβλητές στατιστικώς σημαντικές. Η επιλογή των πιο σημαντικών μεταβλητών θα γίνει με την βοήθεια της βήμα προς βήμα (stepwise) λογιστικής παλινδρόμησης.

Τα αποτελέσματα της παλινδρόμησης αυτής παρουσιάζονται στον πίνακα (2.25), που ακολουθεί:

ΠΙΝΑΚΑΣ 2.25

μεταβλητή	b	τυπικό σφάλμα	Wald τιμή	βαθμοί ελευθερίας	α	R	Exp(b)
EF	-18,5616	3,1162	35,4786	1	0,0000	-0,3885	0,0000
LIPIDS	-1,2114	0,4651	6,7840	1	0,0092	-0,1469	0,2978
SED/AR	-2,0044	0,6979	8,2480	1	0,0041	-0,1678	0,1347
SMOKE	-1,3091	0,5264	6,1857	1	0,0129	-0,1374	0,2701
Constant	13,8822	2,2729	37,3035	1	0,0000		$1,07x10^6$

Παρατηρούμε ότι όπως και στη βήμα προς βήμα διαχωριστική ανάλυση, τέσσερις είναι οι πιο σημαντικές μεταβλητές για την περιγραφή του λογιστικού μοντέλου (2.23): η EF (Το κλάσμα εξώθησης), LIPIDS (Λιπίδια), SEDENTARY (Τρόπος ζωής), και SMOKE (Κάπνισμα).
Η συνάρτηση λογιστικής παλινδρόμησης είναι :

ln(p/(1-p))=13,8822 -18,5616(EF) - 1,2114(LIPIDS) - 2,0044(SEDENTARY) - 1,3091(SMOKE)

η οποία ισοδυναμεί με την:

$$\frac{p}{1-p}=\text{Exp}[13,8822-18,5616(EF)-1,2114(LIPIDS)-2,0044(SEDENTARY)-1,3091(SMOKE)]$$

ή

$$\frac{p}{1-p}=\text{Exp}(13,8822)\times\text{Exp}[-18,5616(EF)]\times\text{Exp}[-1,2114(LIPIDS)]\times\text{Exp}[-2,0044(SEDENTARY)]\times\text{Exp}[-1,3091(SMOKE)] \quad (2.23)$$

Το μοντέλο (2.23) είναι οικονομικότερο των προηγουμένων λόγω του μικρού αριθμού μεταβλητών που περιέχει. Στο συγκεκριμένο μοντέλο, όλες οι μεταβλητές έχουν αρνητικό πρόσημο, και κατά συνέπεια αύξηση της τιμής των μεταβλητών σημαίνει ελάττωση του λόγου υπεροχής p/(1-p). Για παράδειγμα ας θεωρήσουμε την μεταβλητή EF. Αυτή είναι η σημαντικότερη μεταβλητή γιατί έχει τον μεγαλύτερο (κατ' απόλυτη τιμή) συντελεστή. Ας θεωρήσουμε έναν ασθενή ο οποίος με σταθερές όλες τις υπόλοιπες μεταβλητές μεταπηδά από τη φάση "EF=1" (υγιής κατάσταση του μυοκαρδίου), στη φάση "EF=0" (πρόβλημα στη συσταλτικότητα του μυοκαρδίου).

Τότε ο λόγος υπεροχής για τους υγιείς $\frac{p}{1-p}$ μεταβάλλεται κατά ένα παράγοντα της τάξης $\text{Exp}(18,5616)=8,686\times10^{-9}\approx0,0$. Στη συγκεκριμένη περίπτωση αυτό σημαίνει ότι η πιθανότητα p=Pr(DCODE=0) γίνεται ουσιαστικά μηδέν.

Η αποτελεσματικότητα του μοντέλου (2.23) παρουσιάζεται στον πίνακα 2.26

ΠΙΝΑΚΑΣ 2.26

		Προβλεπόμενη ταξινόμηση		Ποσοστό σωστής ταξινόμησης
Παρατηρούμενη ταξινόμηση		0	1	
DCODE	0	66	14	82,5%
	1	14	66	82,5%

Αποτελεσματικότητα = (132/160) x100% = 82,5%

Ειδικότητα = (66/80)x100% = 82,5%

Ευαισθησία = (66/80)x100% = 82,5%

Δείκτης YOUDEN = Ειδικότητα + Ευαισθησία – 100% = 62,5%

Παρατηρείται ότι το ποσοστό επιτυχούς πρόβλεψης του λογιστικού μοντέλου (2.23) παραμένει το ίδιο (82,5%) με αυτό που περιγράφεται από τη σχέση (2.22). Θα μπορούσαμε όμως να ισχυρισθούμε ότι είναι οικονομικότερο

(αριθμός μεταβλητών=4), χωρίς να υπάρχει απώλεια της διαχωριστικής του ικανότητας, κατά συνέπεια θεωρείται καλύτερο από το (2.22)

2.5.2 Λογιστικό μοντέλο με αλληλεπιδράσεις

Η λογιστική παλινδρόμηση έχει το πλεονέκτημα ότι μπορεί να ενσωματώσει αλληλεπιδράσεις (interactions) μεταξύ των μεταβλητών.[Hosmer and Lemeshow(1989)] Έτσι στο σύνολο των μεταβλητών προστέθηκαν και οι αλληλεπιδράσεις των εννέα κατηγορικών μεταβλητών ανά δύο. Προέκυψαν 36 αλληλεπιδράσεις οι οποίες προστέθηκαν στο σύνολο των 11 αρχικών μεταβλητών. Το μοντέλο των 47 μεταβλητών δεν πλήρη το κριτήριο της οικονομικότητας, *είναι σαφώς δύσχρηστο αλλά επιτυγχάνει υψηλά ποσοστά σωστής* ταξινόμησης.
Η διαχωριστική ικανότητα του μοντέλου φαίνεται στον πίνακα 2.27.

ΠΙΝΑΚΑΣ 2.27: Ταξινόμηση της DCODE στο πιλοτικό αρχείο

		Προβλεπόμενη ταξινόμηση		Ποσοστό σωστής ταξινόμησης
Παρατηρούμενη ταξινόμηση		0	1	
DCODE	0	76	4	95%
	1	4	76	95%

Αποτελεσματικότητα = (152/160) x100% = 95%
Ειδικότητα = (76/80)x100% = 95%
Ευαισθησία = (76/80)x100% = 95%
Δείκτης YOUDEN = Ειδικότητα + Ευαισθησία – 100% = 90%.

Αξίζει να σημειωθεί επίσης ότι η τιμή της συνάρτησης –2LL = 41,076 για το πλήρες μοντέλο, η τιμή $\tilde{R}^2$=0,902 καθώς και η τιμή του Hosmer και Lemeshow test =3,95 (α=0,8612), μαρτυρούν ότι το λογιστικό μοντέλο προσεγγίζει καλά τα εμπειρικά δεδομένα.

Από τον πίνακα όμως των εκτιμηθέντων συντελεστών του μοντέλου της λογιστικής παλινδρόμησης που περιέχονται στον πίνακα Π9 του παραρτήματος Β, παρατηρούμε ότι οι τιμές του επιπέδου σημαντικότητας του Wald-test καθιστούν μόνο τους συντελεστές των μεταβλητών EF, HBP* LIPIDS και HBP*OBESITY στατιστικώς σημαντικά διάφορους του μηδενός.

Το υψηλό ποσοστό πρόβλεψης και η τιμή του δείκτη YOUDEN, κάνει το μοντέλο με τις αλληλεπιδράσεις ελκυστικό. Το πλεονέκτημα αυτό εξισορροπείται από το μειονέκτημα της πολυπλοκότητας. Πράγματι ο μεγάλος αριθμός των μεταβλητών κάνουν την χρήση του μοντέλου ανέφικτη, επίσης οι περισσότεροι εκτιμηθέντες συντελεστές των μεταβλητών της λογιστικής παλινδρόμησης είναι στατιστικά μη σημαντικοί. Γι' αυτόν το λόγο στη συνέχεια θα επιχειρηθεί η απλοποίησή του.

Τα αποτελέσματα της ανάλυσης με τη χρήση της "βήμα προς βήμα λογιστικής παλινδρόμησης παρουσιάζονται στον πίνακα 2.28, που ακολουθεί και αποκαλύπτουν σημαντική βελτίωση στη διαχωριστική ικανότητα του μοντέλου (2.24) σε σχέση με τα μοντέλα της διαχωριστικής ανάλυσης. Στο λογιστικό μοντέλο περιέχονται εννέα μεταβλητές εκ των οποίων η μία είναι η μεταβλητή EF και οκτώ αλληλεπιδράσεις των κατηγορικών μεταβλητών. Αναλυτικότερα, περιέχονται οι εξής μεταβλητές:

X_1 = EF

INT_1= FAMILY*LIPIDS

INT_2=LIPIDS* SEDENTARY

INT_3=DIABETES*HBP

INT_4=HBP*TYPE A

INT_5=SEDENTARY*SMOKE

INT_6=SEDENTARY*DIABETES

INT_7=FAMILY*OBESITY

INT_8=OBESITY *SEDENTARY

Όπως παρατηρούμε στον πίνακα 2.28 οι τιμές του επιπέδου σημαντικότητας του Wald-test μαρτυρούν ότι οι εκτιμηθέντες συντελεστές είναι στατιστικώς σημαντικά διάφοροι του μηδενός.

ΠΙΝΑΚΑΣ 2.28

Μεταβλητή	b	τυπικό σφάλμα	Wald τιμή	βαθμοί ελευθερίας	α	R	Exp(b)
X_1	-32,738	6,115	28,666	1	0,000	-0,3467	0,000
INT_1	3,143	1,404	5,013	1	0,025	0,117	23,162
INT_2	-4,682	1,433	10,671	1	0,001	-0,198	0,009
INT_3	3,522	1,183	8,858	1	0,003	0,176	33,854
INT_4	-4,389	1,256	12,212	1	0,001	-0,215	0,012
INT_5	-2,151	0,765	7,919	1	0,005	-0,163	0,116
INT_6	-1.934	0,958	4,074	1	0,043	-0,097	0,145
INT_7	-5,060	1,384	13,372	1	0,000	-0,226	0,006
INT_8	4,188	1,282	10,677	1	0,001	0,198	65,842
Σταθερός όρος	23,020	4,330	28,262	1	0,000		

Η μεγάλη τιμή του συντελεστή της μεταβλητής EF την αναδεικνύει σε ισχυρό διαχωριστικό παράγοντα.

Σύμφωνα με τα όσα προαναφέρθηκαν η συνάρτηση λογιστικής παλινδρόμησης είναι η εξής:

$$\ln(p/1-p) = 23,020 - 32,738.X_1 + 3,143.INT_1 - 4,682.INT_2 + 3,522.INT_3 - 4,389.INT_4 - 2,151.INT_5 - 1,934.INT_6 - 5,060.INT_7 + 4,188.INT_8 \quad (2.24)$$

Η διαχωριστική ικανότητα του μοντέλου στο πιλοτικό δείγμα των 160 ατόμων παρουσιάζεται στον πίνακα 2.29.

ΠΙΝΑΚΑΣ 2.29: Ταξινόμηση της DCODE στο πιλοτικό αρχείο

		Προβλεπόμενη ταξινόμηση		**Ποσοστό σωστής**
Παρατηρούμενη ταξινόμηση		0	1	ταξινόμησης
DCODE	0	72	8	90%
	1	8	72	90%

Αποτελεσματικότητα = (144/160) x100% = 90%

Ειδικότητα = (72/80)x100% = 90%

Ευαισθησία = (72/80)x100% = 90%

Δείκτης YOUDEN = Ειδικότητα + Ευαισθησία – 100% = 80%

Σημειώνεται ότι η τιμή της συνάρτησης –2LL = 82,923 είναι ικανοποιητική για το μοντέλο των εννέα μεταβλητών καθώς το ποσοστό της ερμηνευμένης διακύμανσης, που είναι 77,4% ($\tilde{R}^2$=0,774) .

Το απλοποιημένο μοντέλο (2.24) παρουσιάζει υψηλό ποσοστό σωστής ταξινόμησης, όπως επίσης ειδικότητας και ευαισθησίας. Ο αριθμός των μεταβλητών έχει μειωθεί σημαντικά και η τιμή του δείκτη YOUDEN είναι πολύ ικανοποιητική. Το μοντέλο αυτό πληρεί τα κριτήρια της οικονομικότητας, της καλής προσαρμογής, του υψηλού ποσοστού σωστής ταξινόμησης, της υψηλής τιμής του δείκτη YOUDEN, κατά συνέπεια προτείνεται ως το καλύτερο έναντι των άλλων μοντέλων της λογιστικής παλινδρόμησης.

Ο τελικός έλεγχος σωστής ταξινόμησης του καλύτερου λογιστικού μοντέλου (2.24), όπως και αυτών της διαχωριστικής συνάρτησης γίνεται επί του συνόλου του δείγματος. Τα αποτελέσματα του ελέγχου παρατίθενται στον πίνακα (2.32) .

2.6 Εφαρμογή διαχωριστικών τεχνικών στο συνολικό δείγμα.

Οι διαχωριστικές συναρτήσεις (2.14) και (2.15) οι οποίες εμφάνισαν την καλύτερη απόδοση στο πιλοτικό δείγμα των 160 ατόμων εφαρμόστηκαν στο συνολικό δείγμα των 660 ατόμων, με σκοπό να ελεγχθεί η αποτελεσματικότητα τους και σε άγνωστα δεδομένα. Υπολογίστηκαν οι τιμές των διαχωριστικών συναρτήσεων (2.14) και (2.15) για κάθε άτομο του συνολικού δείγματος και οι τιμές αυτές συγκρίθηκαν με την κριτική τιμή των διαχωριστικών συναρτήσεων z_c=0. Για την ευκολότερη ταξινόμηση των 660 ατόμων σε υγιείς και ασθενείς κατασκευάστηκε η μεταβλητή Eval η οποία παίρνει την τιμή 0 για τους υγιείς, των οποίων η τιμή της διαχωριστικής συνάρτησης είναι αρνητική ενώ για τα άτομα που αναμένεται να είναι ασθενή παίρνει την τιμή 1 και η τιμή της διαχωριστικής συνάρτησης είναι θετική. Η μεταβλητή Eval είναι η αξιολόγηση των διαχωριστικών τεχνικών.

Παρατίθεται ο πίνακας (2.30) που είναι η αξιολόγηση της διαχωριστικής συνάρτησης (2.14) και ταξινομεί τις απαντήσεις Eval σύμφωνα με τις σωστές απαντήσεις της μεταβλητής DCODE.

ΠΙΝΑΚΑΣ 2.30:Πινακας ταξινόμησης της EVAL ως προς την DCODE

			EVAL(Προβλεπόμενη ταξινόμηση)		
			0	1	Σύνολο
DCODE	0	Συχνότητα	**72**	8	80
		%	90%	10%	100%
	1	Συχνότητα	121	**459**	580
		%	20,9%	79,1%	100%

Το συνολικό ποσοστό σωστών απαντήσεων είναι ίσο με $\frac{72+459}{660}$ =80,4%.

Αποτελεσματικότητα = (534/660) x100% = 80,4%

Ειδικότητα = (72/80)x100% = 90%

Ευαισθησία = (459/580)x100% = 79,1%

Δείκτης YOUDEN = Ειδικότητα + Ευαισθησία – 100% = 69,1%

Ο πίνακας (2.31) είναι η αξιολόγηση της διαχωριστικής συνάρτησης (2.15)

ΠΙΝΑΚΑΣ 2.31:Πίνακας ταξινόμησης της EVAL ως προς την DCODE

Παρατηρούμενη ταξινόμηση		EVAL(Προβλεπόμενη ταξινόμηση)		
			0	1
DCODE	*0*	Συχνότητα	**70**	*10*
		%	87,5	12,5
	1	Συχνότητα	116	**464**
		%	20	80

Αποτελεσματικότητα = (534/660) x100% = 80,9%

Ειδικότητα = (70/80)x100% = 87,5%

Ευαισθησία = (464/580)x100% = 80,0%

Δείκτης YOUDEN = Ειδικότητα + Ευαισθησία – 100% = 67,5%

Το συνολικό ποσοστό σωστών απαντήσεων είναι ίσο με $\frac{70+464}{660}$=80,9%.

Με ανάλογο τρόπο αξιολογήθηκε το λογιστικό μοντέλο (2.24) στο συνολικό δείγμα το οποίο προτείνεται ως το καλύτερο της λογιστικής παλινδρόμησης. Υπολογίστηκε η τιμή του λογιστικού μοντέλου για κάθε άτομο του συνολικού δείγματος και θεωρήθηκε ως κριτική τιμή, η τιμή p=0,5. Η μεταβλητή EVAL πήρε την τιμή 0 και καταχώρησε ως υγιή το ιοστο άτομο του οποίου η τιμή p_i>0,5.

ΠΙΝΑΚΑΣ 2.32: Έλεγχος του λογιστικού μοντέλου (2.24) στο σύνολο του δείγματος.

		EVAL(Προβλεπόμενη ταξινόμηση)		Ποσοστό σωστής
Παρατηρούμενη ταξινόμηση		0	1	ταξινόμησης
DCODE	0	**71**	9	88,8%
	1	91	**489**	84,3%

Αποτελεσματικότητα = (560/660) x100% = 84,9%
Ειδικότητα = (71/80)x100% = 88,8%
Ευαισθησία = (489/580)x100% = 84,3%
Δείκτης YOUDEN = Ειδικότητα + Ευαισθησία – 100% = 73,1%

Αν και παρατηρείται πτώση της αποτελεσματικότητας των διαχωριστικών συναρτήσεων και του μοντέλου της λογιστικής παλινδρόμησης κατά την εφαρμογή τους στο σύνολο του δείγματος (άγνωστα δεδομένα), οι τιμές της αποτελεσματικότητας (>80%) και του δείκτη YOUDEN(≈70%) των συγκεκριμένων μοντέλων βρίσκονται μέσα στα αποδεκτά όρια. Η μείωση των τιμών αυτών, σε σχέση με τις αντίστοιχες του πιλοτικού δείγματος πιθανώς να οφείλεται στη ιδιαιτερότητα του δείγματος των 660 ατόμων, όπου ο αριθμός των ασθενών είναι πολύ μεγαλύτερος ως προς τον αριθμό των υγιών (87,9% *ποσοστό ασθενών*-12,1% *ποσοστό υγιών*).

Υπενθυμίζεται ότι τα μοντέλα των διαχωριστικών τεχνικών υπολογίστηκαν στο πιλοτικό δείγμα όπου το ποσοστό υγιών –ασθενών είναι το ίδιο (50%).

Από τις διαχωριστικές συναρτήσεις αν και η (2.15) παρουσιάζει ελαφρώς μεγαλύτερη αποτελεσματικότητα και ελαφρώς αυξημένη ευαισθησία, η διαφορά αυτή δεν ελέγχεται ως στατιστικώς σημαντική. Η διαχωριστική συνάρτηση (2.15) εμφανίζει χαμηλότερο δείκτη YOUDEN (67,5%), έναντι της διαχωριστικής (2.14) με δείκτη YOUDEN = 69,14%, άρα η διαχωριστική (2.14) θεωρείται καταλληλότερη.

Αντίστοιχα το μοντέλο λογιστικής παλινδρόμησης (2.24) έχει μεγαλύτερη αποτελεσματικότητα, αυξημένη ευαισθησία, και ικανοποιητικό δείκτη

YOUDEN=73,1% στο συνολικό δείγμα. Μπορούμε επίσης να ισχυρισθούμε ότι το μοντέλο λογιστικής παλινδρόμησης έχει τη μεγαλύτερη ικανότητα σωστής ταξινόμησης στο πιλοτικό δείγμα, άρα θα μπορούσε να θεωρηθεί ως το καλύτερο μοντέλο των διαχωριστικών τεχνικών της παρούσας έρευνας.

Το μοντέλο λογιστικής παλινδρόμησης προτιμάται σε σχέση με το γραμμικό μοντέλο επειδή οι αλληλεπιδράσεις είναι στατιστικά σημαντικές. Αντίστοιχη αναφορά έχει γίνει από τους Byth (1980) και Press (1978) οι οποίοι απέδειξαν ότι όταν υπάρχουν στατιστικά σημαντικές αλληλεπιδράσεις το λογιστικό μοντέλο προτιμάται έναντι του αντίστοιχου γραμμικού.

ΚΕΦΑΛΑΙΟ 3
ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Εισαγωγή

- **Ιστορική εξέλιξη των νευρωνικών δικτύων**

Η έρευνα των τεχνητών νευρωνικών δικτύων (ΤΝΔ) υπολογίζεται να έχει ξεκινήσει από το 1800 περίπου, όπως αναφέρεται σε μια εργασία του Freud (1966), κατά την περίοδο της προ-ψυχανάλυσης. Η πρώτη υλοποίηση ΤΝΔ έγινε από μια υδραυλική συσκευή, όπως περιγράφεται από τον Russel (1913). Στη δεκαετία του 1940, επιστήμονες στο πανεπιστήμιο της Πενσυλβανίας, κατασκεύασαν τον πρώτο ψηφιακό υπολογιστή με το όνομα "ENIAC", ο οποίος αποτέλεσε τον πατέρα των σύγχρονων υπολογιστικών συστημάτων. Στόχος των ερευνητών ήταν η μίμηση των λειτουργιών του αριστερού τμήματος του εγκεφάλου. (Aleksander 1987)

Το 1943 οι Warren McCullock και Walter Pitts δημοσίευσαν μια εργασία με τίτλο "A Logical Calculus of Ideas Immanent in Nervous Activity" [McCullock και Pitts (1943)] η οποία αποτέλεσε τη βάση για τη μεταγενέστερη ανάπτυξη των νευρωνικών δικτύων. Το 1951, ένας φοιτητής του M.I.T. ο Marvin Minsky υλοποίησε το πρώτο νευρωνικό δίκτυο με το οποίο προσπάθησε να επιλύσει το πρόβλημα της εκμάθησης ενός λαβύρινθου. Αυτή ήταν η αρχή του επιστημονικού πεδίου της τεχνητής νοημοσύνης και ο Dr. Marvin Minsky, που είναι ακόμα καθηγητής στο M.I.T. θεωρείται ως ο πατέρας των έμπειρων συστημάτων. Τα συστήματα αυτά μπορούν να παίρνουν διαδοχικές αποφάσεις με βάση προγραμματισμένα βήματα και πληροφορία που έχει αποθηκευτεί στη μνήμη τους. Τα έμπειρα συστήματα μπορούν εύκολα να επεκταθούν για να δεχτούν νέες πληροφορίες και κανόνες, σε αντίθεση με τα συμβατικά προγράμματα που πρέπει στην περίπτωση αυτή να ξαναγραφούν. Σήμερα τα έμπειρα συστήματα χρησιμοποιούνται για ιατρική διάγνωση, έλεγχο παραγωγής, υπολογισμό κόστους, οδηγίες επισκευών κλπ. Τα έμπειρα συστήματα χρησιμοποιούν τη γνώση των ειδικών και χρησιμεύουν σαν άριστοι εκπαιδευτές.

Στα τέλη της δεκαετίας του 1970 με την ανάπτυξη ισχυρών υπολογιστικών συστημάτων, έγινε εφικτή η πρακτική έρευνα πάνω στις εφαρμογές των τεχνητών νευρωνικών δικτύων, ενώ η ανάπτυξη του αλγόριθμου εκπαίδευσης "backpropagation" επέτρεψε τη χρήση των δικτύων για τη λύση καθημερινών προβλημάτων στον επιστημονικό, επιχειρηματικό και βιομηχανικό χώρο.

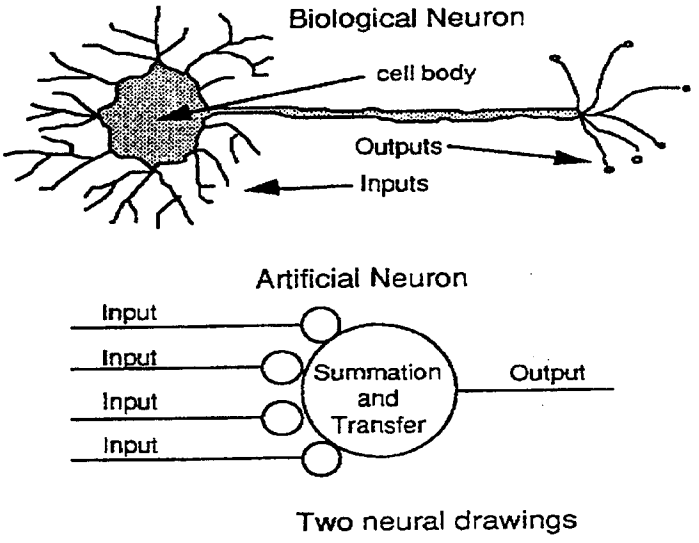
Στο σημείο αυτό αξίζει να αναφερθούν οι έννοιες της «ασαφούς λογικής» (fuzzy logic) και των ασαφών συστημάτων (fuzzy systems). Η ασαφής λογική αποτελεί νέο, σχετικά, τομέα της τεχνητής νοημοσύνης που επιτρέπει την λογική επεξεργασία ασαφών πληροφοριών. Τα ασαφή συστήματα είναι μια μέθοδος ορισμού και χειρισμού δεδομένων. Μοιάζουν με τα νευρωνικά δίκτυα στο ότι και αυτά διαχειρίζονται μη ακριβή πληροφορία με μη ακριβή τρόπο. Τα νευρωνικά δίκτυα χρησιμοποιούν παραδείγματα αντί για κανόνες για να αναγνωρίσουν patterns που μπορεί να είναι αόριστα ή ελαφρά αντιφατικά. [Jain et al (1995)]. Τα ασαφή συστήματα και τα έμπειρα συστήματα μοιάζουν στο ότι και τα δύο χρησιμοποιούν μια λογική βασισμένη σε κανόνες κατά τη διάρκεια της επίλυσης προβλημάτων.

Η καινοτομία των Τ.Ν.Δ. βρίσκεται στην ικανότητά τους να υποδειγματοποιούν μη-γραμμικές διαδικασίες δίχως a priori υποθέσεις γύρω από τη φύση της διαδικασίας που δημιουργεί τις μεταβλητές που θέλουμε να μελετήσουμε [Συριόπουλος (1997)]. Τα Τ.Ν.Δ. είναι στατιστικά εργαλεία, ανάλογα των μη-παραμετρικών, μη - γραμμικών μοντέλων παλινδρόμησης. Η ουσιαστική και σημαντική διαφορά τους είναι ότι αποτελούν μια "data-driven" προσέγγιση του προβλήματος που μελετάμε, αντίθετα με τα άλλα μοντέλα που είναι "model-driven" προσεγγίσεις.

- **Σύγκριση του βιολογικού και του τεχνικού νευρώνα**

Τα νευρωνικά δίκτυα προσπαθούν να μιμηθούν τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος επεξεργάζεται τα δεδομένα. Η βασική δομή του νευρικού συστήματος είναι ο νευρώνας (biological neuron), ένα κέλυφος το οποίο διοχετεύει την πληροφορία από και προς τα διάφορα μέρη του σώματος.

Ο νευρώνας αποτελείται από το σώμα ή κορμό του κέλυφους, μερικές επεκτάσεις μορφής αγκαθιού που ονομάζονται δενδρίτες και ένα απλό νευρικό νήμα το οποίο ονομάζεται άξονας και διακλαδώνεται από το σώμα ενώ συνδέεται με άλλους νευρώνες [Lawrence 1993)] (Σχ.3.1α).



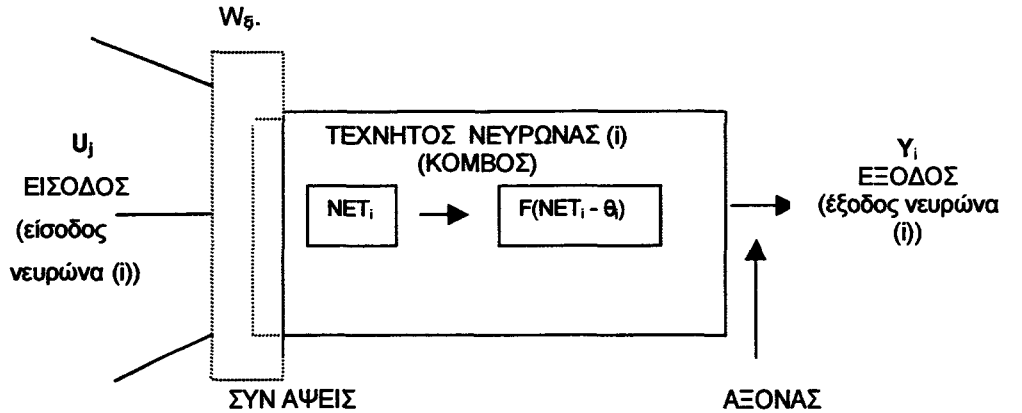
Σχ. 3.1 Βιολογικός και τεχνητός νευρώνας

Οι συνδέσεις που υπάρχουν μεταξύ των νευρώνων, πάνω στο σώμα ή πάνω στους δενδρίτες ονομάζονται συνάψεις. Οι άξονες και οι δενδρίτες μπορούν να παρομοιαστούν με μεμονωμένους αγωγούς διαφορετικού πάχους, οι οποίοι μεταφέρουν ηλεκτρικά σήματα στο νευρώνα. Όλοι οι νευρώνες, οι οποίοι αλληλοσυνδέονται μέσω του άξονα και των δενδριτών, οι οποίοι μεταφέρουν σήματα και ρυθμίζονται από τις συνάψεις, δημιουργούν ένα νευρωνικό δίκτυο.

Σαν μονάδα επεξεργασίας, ο νευρώνας θεωρείται ότι επιτελεί έναν απλό υπολογισμό κατωφλίου ή οριακής τιμής (threshold value). Τα σήματα εξόδου από τους άξονες των γειτονικών νευρώνων συλλέγονται στις συνάψεις και μεταφέρονται μέσω των δενδριτών στο σώμα του κυττάρου, όπου αθροίζονται, διαγείροντας ή όχι το κύτταρο. Όταν η αθροιστική διέγερση στο σώμα του κυττάρου υπερβεί μια οριακή τιμή (κατώφλι), ο νευρώνας ενεργοποιείται, και παράγει το δικό του σήμα το οποίο μεταφέρεται μέσω των αξόνων στο γειτονικό κύτταρο.

Σε αναλογία με το βιολογικό νευρώνα, ο τεχνητός νευρώνας, που ονομάζεται και κόμβος, δέχεται ένα σύνολο δεδομένων εισόδου U_j, j=1,2,3,...,n, τα οποία πολλαπλασιάζει με ένα σύνολο σταθμίσεων (βαρών) W_{ji},

όπου i=1,2,...,q ο αριθμός των κόμβων του κρυφού επιπέδου. Οι σταθμίσεις αυτές λαμβάνονται αρχικά με τυχαίο τρόπο από μια κανονική κατανομή βαρών.



- $NET_i = \sum_{j=1}^{n} u_j w_{ji}$
- $Y_i = F(NET_i - \theta_i)$:συνάρτηση μεταφοράς

θ_i: τιμή κατωφλίου του νευρώνα(i)

Σχ.3.2 Λειτουργία νευρώνα κρυφού επιπέδου.

Οι σταθμισμένες μονάδες εισόδου αθροίζονται για να προσδιορίσουν το επίπεδο ενεργοποίησης του νευρώνα, δημιουργώντας ένα σταθμισμένο άθροισμα: $NET_i = \sum_{j=1}^{n} u_j w_{ji}$ (Σχ.3.2).

Στη συνέχεια στο νευρώνα i, μια συνάρτηση μεταφοράς $F(NET_i - \theta_i)$ παράγει την έξοδο Y_i από το συγκεκριμένο νευρώνα. Η τιμή κατωφλίου θ_i καθορίζεται από έναν κόμβο ο οποίος ονομάζεται «μονάδα πόλωσης» (bias unit). Η τιμή αυτή είναι αντίθετη του βάρους που προσδίδεται στην έξοδο της μονάδας πόλωσης. Το επίπεδο ενεργοποίησης της μονάδας πόλωσης λαμβάνεται ίσο με τη μονάδα. Πρέπει να σημειωθεί ότι αρχικά οι τιμές των βαρών που προσδίδονται τόσο στα δεδομένα εισόδου όσο και στην έξοδο της μονάδας πόλωσης είναι μικροί τυχαίοι αριθμοί.[Sampath. G.(1977)].

Συνήθως ως συνάρτηση μεταφοράς χρησιμοποιείται η σιγμοειδής ή λογιστική συνάρτηση η οποία παράγει την έξοδο του i νευρώνα:

$$Y_i = \frac{1}{1+e^{-(\sum_j u_j w_{ji} - \theta_i)}}$$

Στα επόμενα η τιμή κατωφλίου θ_i τίθεται ίση με το μηδέν, οπότε η συνάρτηση μεταφοράς παίρνει τη μορφή

$$Y_i = \frac{1}{1+e^{-(\sum_j u_j w_{ji})}}$$

Στην περίπτωση που η συνάρτηση μεταφοράς είναι συνάρτηση στάθμισης τότε η έξοδος:

$$Y_i = \begin{cases} 1, \alpha\nu \quad NET_i > \theta_i \\ 0, \alpha\lambda\lambda o\acute{\upsilon} \end{cases}$$

Η παραπάνω διαδικασία γίνεται στο λεγόμενο «κρυφό στρώμα» του δικτύου. Το επίπεδο (στρώμα) αυτό αποτελείται από κόμβους που βρίσκονται μεταξύ επιπέδου εισόδου και επιπέδου εξόδου. Οι κόμβοι στο δίκτυο διατάσσονται σε επίπεδα και καθένας απ' αυτούς συνδέεται με όλους τους κόμβους του επόμενου επιπέδου. Τα σήματα $Y_i = Y(u_j, w_i)$από το κρυφό επίπεδο περνάνε στο επίπεδο εξόδου του δικτύου προσλαμβάνοντας επιπλέον σταθμίσεις β_i. Το επίπεδο εξόδου δίνει την έξοδο του δικτύου, και έχει τη μορφή:

$f(u,\delta) = \alpha + \sum_{i=1}^{q} \beta_i y_i$, όπου $\alpha, \beta_1, \ldots, \beta_q$ είναι οι σταθμίσεις από το κρυφό επίπεδο προς το επίπεδο εξόδου του δικτύου, $w_1, \ldots, w_q$ είναι οι σταθμίσεις από το επίπεδο εισόδου προς το κρυφό επίπεδο και $\delta = (\alpha, \beta_i, \ldots \beta_q, w'_1, \ldots, w'_q)$ είναι η μήτρα των παραπάνω παραμέτρων. [Συριόπουλος (1992), (1997)].

Στο σημείο αυτό πρέπει να διατυπωθεί το Θεμελιώδες θεώρημα των ΤΝΔ [Hornik, Stinchcombe and White (1989), (1990)], σύμφωνα με το οποίο:

«Κάθε συνάρτηση της μορφής $f(u,\delta) = \alpha + \sum_{i=1}^{q} \beta_i y_i$

αποτελεί μια «καθολική προσέγγιση» και υπάρχει πάντα ένας πεπερασμένος αριθμός m κόμβων του (μοναδικού) κρυφού στρώματος και μια μήτρα παραμέτρων δ έτσι ώστε η συνάρτηση της εξόδου να προσεγγίζει οποιαδήποτε μη-γραμμική συνάρτηση ορισμένη στο σύνολο των πραγματικών αριθμών, με την προϋπόθεση ότι η εξεταζόμενη μεταβλητή δεν παρουσιάζει υπερβολικές ασυνέχειες».

Η διαδικασία εκμάθησης λαμβάνει χώρα μέσω μιας συνεχούς αλλαγής των βαρών του δικτύου (weight adaptation) κατά τη φάση της εκπαίδευσής του. Η φάση αυτή αποτελεί μια προσπάθεια σύνδεσης εξωτερικών προτύπων εισόδου με εξωτερικά πρότυπα εξόδου μέσω επαναληπτικής παρουσίασής τους στο δίκτυο. Το σύνολο των προτύπων εισόδου ονομάζεται σύνολο προτύπων εκπαίδευσης (training set) και η πληροφορία που παρέχουν καθορίζει το μέγεθος των αλλαγών στα βάρη του δικτύου. [Kandel (1985)]

Η έξοδος του δικτύου συγκρίνεται με την επιθυμητή έξοδο και στη συνέχεια αρχίζει η διαδικασία της ανάστροφης μετάδοσης σφάλματος (back propagation).

Η κυριότερη διαφορά μεταξύ των βιολογικών και των τεχνητών νευρωνικών δικτύων είναι ο τρόπος εισόδου των δεδομένων.

Στα βιολογικά δίκτυα η είσοδος διαμορφώνεται από τις πέντε αισθήσεις. Στα τεχνητά δίκτυα η είσοδος είναι αποκλειστικά αριθμητική και συνήθως μεταξύ του εύρους [-1,+1]. Οι διαφορές μεταξύ των δυο τύπων δικτύων συνοψίζονται στον παρακάτω πίνακα 3.1.

ΠΙΝΑΚΑΣ 3.1

Παράγοντας	Βιολογικό Δίκτυο	Τεχνητό Δίκτυο
Μέγεθος	Μεγαλύτερο από 100 δισεκατομμύρια νευρώνες	Συνήθως μερικές εκατοντάδες νευρώνες
Είσοδοι	Οι πέντε αισθήσεις	Αριθμοποιημένα σύνολα δεδομένων
Εξειδίκευση	Ικανότητα εκτέλεσης πολλών εργασιών	

- **Εφαρμογές των νευρωνικών δικτύων.**

Τα νευρωνικά δίκτυα χρησιμοποιούνται σε πολλές πρακτικές εφαρμογές που ανάμεσα τους συγκαταλέγονται ιατρικές διαγνώσεις, αναγνώριση προτύπων, έλεγχος προϊόντων και χρηματοοικονομικές προβλέψεις [Zapranis (1998)]. Ακολουθεί ένας σύντομος κατάλογος των ικανοτήτων των νευρωνικών δικτύων και των αντίστοιχων εφαρμογών τους [Profile neural application (1995)].

Ικανότητα	Εφαρμογή
Αναγνώριση προτύπων	Αναγνώριση υποβρυχίων από σόναρ
Γενίκευση	Καθορισμός αξίας ακινήτων
Πρόβλεψη τάσεων	Απόφαση για αγορά ή πώληση μετοχών
Πρόβλεψη συμπεριφοράς	Ιατρικές προβλέψεις
Εκτίμηση	Έγκριση/Απόρριψη αιτήσεων δανειοδότησης
Ανοχή σε κακή ποιότητα στοιχείων.	Αναγνώριση οπτικών χαρακτήρων
Φιλτράρισμα	Καθαρισμός σημάτων video
Ταχύτητα λειτουργίας	Έλεγχος ρομποτικών μελών
Αντίληψη λεπτών σχέσεων	Εξειδικευμένες ιατρικές συμβουλές
Graceful Degradation	Μηχανικός έλεγχος στο διάστημα
Βελτιστοποίηση	Προγραμματισμός πτήσεων
Ανάλυση μεγάλου αριθμού στοιχείων	Συνυπολογισμός ασφαλιστικών διεκδικήσεων
Extrapolation	Διάγνωση μελλοντικών σφαλμάτων

Τα νευρωνικά δίκτυα χρησιμοποιούνται επίσης στην πρόβλεψη χημικών αντιδράσεων, στον έλεγχο βιομηχανικής παραγωγής, στην ταξινόμηση δακτυλικών αποτυπωμάτων, στην αναγνώριση καρκινογόνων κυττάρων, στην πρόβλεψη οικονομικών τάσεων και τιμών. [Συριόπουλος (1997), Siriopoulos, Markellos, Sirlantzis(1996)].

3.1 Στοιχεία θεωρίας νευρωνικών δικτύων

Τα νευρωνικά δίκτυα διακρίνονται σε δίκτυα με ανάδραση (Feedback Network) και σε εμπρόσθια δίκτυα (Feedforward Neural Network). Ο τύπος εμπροσθίων δικτύων, που χρησιμοποιείται ευρύτερα είναι ο «Standard Feedforward Neural Network». Δίκτυα αυτής της αρχιτεκτονικής παρέχει το πακέτο «Profile Neural Applications (P.N.A)» που χρησιμοποιήθηκε για την εκπόνηση της παρούσας εργασίας.

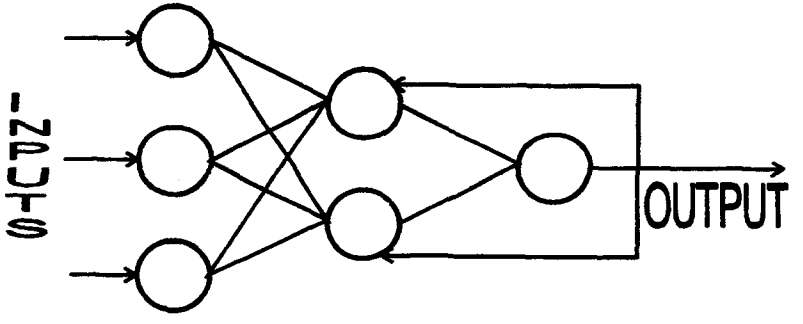
Οι δυο αυτοί τύποι δικτύων πέρα από τις διαφορές στην αρχιτεκτονική τους, διαφέρουν σημαντικά στον τρόπο και τη φιλοσοφία της εκπαίδευσής τους. Στη συνέχεια θα αναφερθούμε στα εμπρόσθια νευρωνικά δίκτυα.

Τα δίκτυα αυτά διακρίνονται σε γραμμικά και μη γραμμικά ανάλογα με τη φύση της συνάρτησης μεταφοράς των νευρώνων τους. Τα μη γραμμικά νευρωνικά δίκτυα μπορεί να συμβάλλουν στο να ξεπεραστούν οι αδυναμίες που

εμφανίζουν τα γραμμικά μοντέλα όταν προσπαθούν να περιγράψουν φαινόμενα με όχι απόλυτα γραμμική συμπεριφορά.

3.1.1 Εμπρόσθια μη γραμμικά νευρωνικά δίκτυα

Στην κατηγορία αυτή ανήκουν τα δίκτυα εκείνα στα οποία τα σήματα κινούνται μόνο προς μια κατεύθυνση, η συνάρτηση μεταφοράς των κόμβων είναι μη γραμμική και δεν περιέχονται κλειστοί βρόγχοι στη δομή τους (Σχ. 3.3).



Ο πιο συνηθισμένος τρόπος εκπαίδευσης είναι ο «επιβλεπόμενος» και υλοποιείται κυρίως με τον αλγόριθμο ανάστροφης μετάδοσης σφάλματος (back propagation).

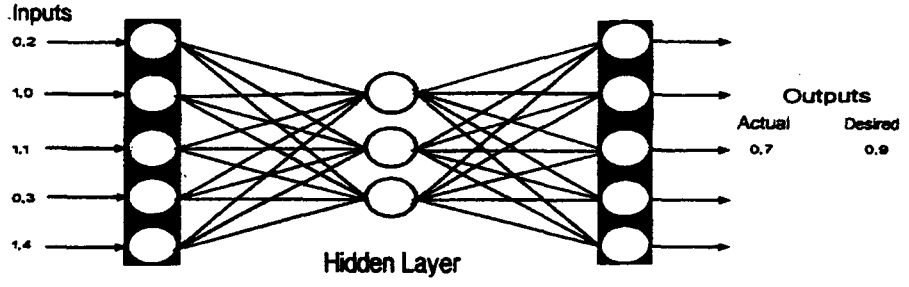
3.1.2 Επιβλεπόμενη μάθηση (Supervised Learning)

Κατά την επιβλεπόμενη μάθηση για την εύρεση των συσχετίσεων μεταξύ των δεδομένων εισόδου-εξόδου, χρησιμοποιούνται εξωτερικά σήματα, τα οποία παίζουν το ρόλο του δασκάλου. Για κάθε ζευγάρι εισόδου-εξόδου η ακριβής διαφορά μεταξύ της επιθυμητής και της πραγματικής τιμής της εξόδου είναι γνωστή. Έτσι ένα σήμα λάθους αναδράται στο δίκτυο μεταβάλλοντας τα βάρη ώστε η διαφορά αυτή να ελαχιστοποιηθεί. Ο στόχος αυτής της διαδικασίας μάθησης είναι να εξαφανίσει τις διαφορές ανάμεσα στην επιθυμητή και στην πραγματική τιμή της εξόδου βρίσκοντας το καθολικό ελάχιστο μιας συνάρτησης επιφάνειας λάθους. Μια συνοπτική αναφορά στον αλγόριθμο "back propagation" [Lawrence (1993)] βοηθά στην καλύτερη κατανόηση των μηχανισμών που καθορίζουν την εκπαίδευση του δικτύου.

Ας θεωρήσουμε το δίκτυο του σχήματος 3.2. Έστω u_j το διάνυσμα εισόδου και w_{ji} τα βάρη μεταξύ των εισόδων και των κόμβων του κρυμμένου επιπέδου. Σε κάθε κόμβο i του κρυφού επιπέδου υπολογίζεται η ποσότητα:

$$X_i = \sum_{j=1}^{n} u_j w_{ji} - \theta_i \qquad (3.1)$$

όπου n ο αριθμός των κόμβων εισόδου [Limin Fu (1994)]. Το σταθμισμένο άθροισμα X_i ονομάζεται συνολική είσοδος του κόμβου i του κρυφού επιπέδου. Η ποσότητα αυτή θα αποτελέσει είσοδο μιας συνάρτησης μεταφοράς, η έξοδος της οποίας θα καθορίσει την κατάσταση του κόμβου.



Σχ. 3.4

Πριν την έναρξη της διαδικασίας της εκπαίδευσης τα βάρη λαμβάνουν τυχαίες τιμές. Η εκπαίδευση του δικτύου συνίσταται στη συστηματική αλλαγή των βαρών μέχρι να βρεθούν αυτά που παράγουν την επιθυμητή έξοδο (μέσα σε ένα δεδομένο διάστημα ανοχής), με βάση ένα συγκεκριμένο πρότυπο εισόδου από όπου πρέπει να παραχθεί ένα επιθυμητό διάνυσμα εισόδου. Η αλλαγή αυτή επαναλαμβάνεται για κάθε διάνυσμα του συνόλου διανυσμάτων εκπαίδευσης. Για κάθε σύνδεσμο στο δίκτυο υπολογίζεται η παράγωγος ως προς το βάρος του συνδέσμου, ενός καθολικού μέτρου του λάθους στην απόδοση του δικτύου. Τα βάρη των συνδέσμων κατόπιν, ρυθμίζονται προς εκείνη την κατεύθυνση που μειώνει το λάθος. Ένα μέτρο της απόδοσης του δικτύου δίνεται από την εξίσωση (3.2), όπου $y_{j,c}$ είναι η πραγματική τιμή του κόμβου εξόδου j, για το πρότυπο εκπαίδευσης c, και $d_{j,c}$ η επιθυμητή τιμή.

$$E = \frac{1}{2}\sum_{j,c}(y_{j,c} - d_{j,c})^2 \qquad (3.2)$$

Ως εκπαίδευση λοιπόν, μπορεί να θεωρηθεί η διαδικασία εύρεσης του καθολικού ελάχιστου της συνάρτησης Ε.

Αυτό επιτυγχάνεται μεταβάλλοντας επαναληπτικά τα βάρη κατά ένα ποσό ανάλογο της μερικής παραγώγου $\partial E/\partial w$ δηλαδή

$$\Delta W_{ji} = \lambda \cdot \delta_{i,c} \cdot y_{j,c} \qquad (3.3)$$

Κατά συνέπεια τα βάρη τη χρονική στιγμή t_{ij} θα δίνονται από τον τύπο:

$$W_{ij} \cdot (t+1) = W_{ij}(t) + \Delta W_{ij}$$

όπου $W_{ij}(t)$ είναι τα βάρη από τον κόμβο i στον κόμβο j κατά τη χρονική στιγμή t και Δw_{ji} η μεταβολή των βαρών.

Η σταθερά λ ονομάζεται ρυθμός μάθησης. Η ποσότητα $\delta_{i,c}$ υπολογίζεται από την παραγώγιση των 3.2 και 3.3 και ισούται με:

$$\delta_{i,c} = (d_{j,c} - y_{j,c}) \cdot f'(y_i) \qquad (3.4)$$

όπου $f(y_i)$ η συνάρτηση μεταφοράς των κόμβων του δικτύου.

Ο παραπάνω αλγόριθμος έχει μια απλή γεωμετρική ερμηνεία [Hinton (1987)] Μπορούμε να παραστήσουμε γραφικά τον πολυδιάστατο χώρο των βαρών χρησιμοποιώντας έναν άξονα για κάθε ανεξάρτητη μεταβλητή βάρους $w_{ji,}$ και έναν άξονα για την εξαρτημένη μεταβλητή του συνολικού λάθους Ε. Σε κάθε συνδυασμό των τιμών των βαρών αντιστοιχεί μια τιμή λάθους. Το σύνολο των δυνατών τιμών που μπορεί να πάρει η συνάρτηση λάθους σχηματίζει την επιφάνεια λάθους.

Αν το δίκτυο δεν περιέχει κρυφούς κόμβους η επιφάνεια αυτή είναι παραβολοειδής με αποτέλεσμα να περιέχει μόνο ένα ελάχιστο το οποίο ο αλγόριθμος είναι βέβαιο ότι θα το προσεγγίσει.

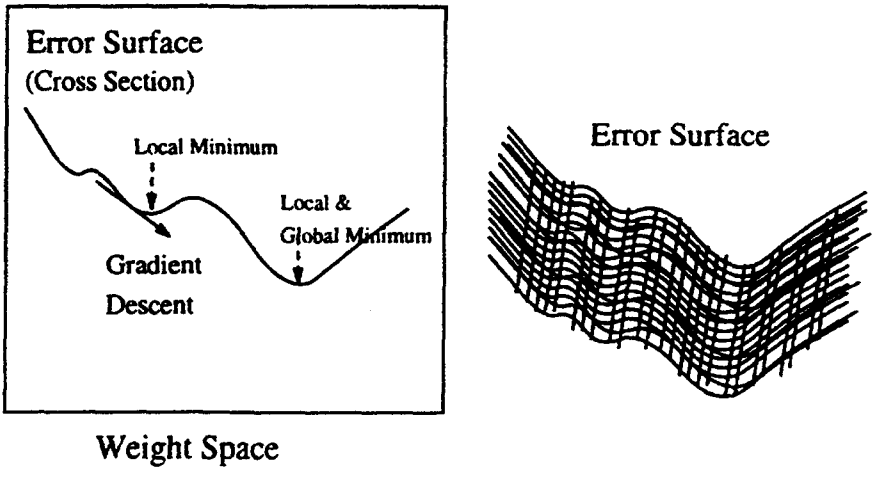
Αν όμως το δίκτυο περιέχει κρυφούς κόμβους, η επιφάνεια λάθους είναι δυνατό να περιέχει πολλά τοπικά ελάχιστα με αποτέλεσμα να υπάρχει ο κίνδυνος παγίδευσης του αλγόριθμου σε ένα από αυτά εμποδίζοντας την προσέγγιση του ολικού ελαχίστου.

3.1.3 Backpropagation (Ανάστροφη μετάδοση σφάλματος)

Το δίκτυο "backpropagation " είναι το πιο ευρέως διαδεδομένο δίκτυο ανάμεσα στους σύγχρονους τύπους των δικτύων των νευρωνικών συστημάτων.

Το δίκτυο αυτό είναι ένα πολυστρωματικό δίκτυο ανάδρασης με μια διαφορετική συνάρτηση μεταφοράς στον τεχνητό νευρώνα και με ισχυρότερο κανόνα εκμάθησης από άλλα δίκτυα (π.χ perceptrons).

Ο αλγόριθμος "backpropagation" παίρνει το όνομά του από τον τρόπο με τον οποίο εκπαιδεύει το νευρωνικό δίκτυο. Είναι ένα είδος τεχνικής κατιούσας κλίσης με ανάστροφη διάδοση σφάλματος (Σχ. 3.5).

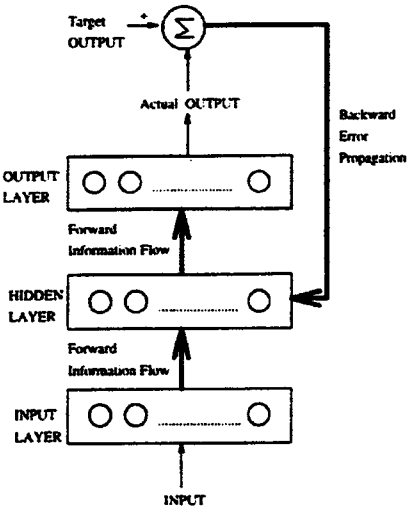


Σχ.3.5: Έρευνα πάνω στην επιφάνεια σφάλματος κατά μήκος της κλίσης

Σε πολλές στιγμές, κατά τη διάρκεια της εκπαίδευσης, το δίκτυο πρέπει να ελέγχεται με σκοπό οι αλληλοσυνδέσεις των βαρών μεταξύ των νευρώνων να είναι σε θέση να ταξινομήσουν σωστά σύμφωνα με τα δεδομένα εκπαίδευσης που δόθηκαν στην είσοδο. Ο έλεγχος αυτός γίνεται σε δύο στάδια:

A) ο αλγόριθμος προσδιορίζει το σφάλμα των αποτελεσμάτων εξόδου του δικτύου, σε σχέση με τα επιθυμητά αποτελέσματα και

B) επιστρέφοντας πίσω μεταδίδει το σφάλμα αυτό μέσω του δικτύου, προσαρμόζοντας κάθε βάρος σε σχέση με το αντίστοιχο σφάλμα του (Σχ.3.6).



Σχ.3.6: Το δίκτυο ανάστροφης διάδοσης

Στη συνέχεια αναφέρονται οι εξισώσεις που χρησιμοποιούνται για την καλύτερη προσαρμογή των βαρών κατά τη διαδικασία της ανάστροφης διάδοσης σφάλματος [Rumelhart et al. (1986), McClelland and Rumelhart, (1988), Carpenter and Hoffman (1995), Limin Fu (1994), Knight (1990)].

Η συνολική είσοδος στον j κόμβο ενός επιπέδου από το προηγούμενο επίπεδο δίνεται από τη σχέση:

$$X_j = \sum_{i=1} \beta_{ji} y_i \qquad (3.5)$$

Η έξοδος του κόμβου αυτού μπορεί να υπολογισθεί με τη χρήση της σιγμοειδούς συνάρτησης μεταφοράς

$$y_j = \frac{1}{1 + e^{-(x_j - \theta_j)}} \qquad (3.6)$$

όπου θ_j η τιμή κατωφλίου του j κόμβου.

Ο στόχος του αλγόριθμου είναι να προσδιορίσει ένα σύνολο βαρών έτσι ώστε οι τιμές του διανύσματος εξόδου, γενικευμένο από κάθε διάνυσμα εισόδου, να είναι ίδιες ή όσο το δυνατό να βρίσκονται πλησιέστερα στις επιθυμητές τιμές του διανύσματος εισόδου [Kingdom (1997)]. Κατ' αρχήν υπολογίζεται το συνολικό σφάλμα γενικευμένο από το δίκτυο:

$$E = \frac{1}{2}\sum_{c}\sum_{j}\left(y_{j,c} - d_{j,c}\right)^2 \qquad (3.7)$$

όπου c είναι ο αριθμός των διανυσμάτων ζευγών εισόδου – εξόδου και j είναι ο αριθμός των μονάδων εξόδου. Το εμπρόσθιο πέρασμα έχει τώρα συμπληρωθεί και το οπίσθιο πέρασμα μπορεί να αρχίσει. Το πρώτο βήμα του οπισθίου περάσματος είναι ο υπολογισμός $\partial E / \partial y$ για κάθε μονάδα εξόδου. Με παραγώγιση της εξίσωσης (3.7) παίρνουμε:

$$\frac{\partial E}{\partial y_j} = y_j - d_j \qquad (3.8)$$

Χρησιμοποιώντας τον κανόνα της αλυσίδας $\partial E / \partial x$

$$\frac{\partial E}{\partial x_j} = \frac{\partial E}{\partial y_j} \cdot \frac{dy_j}{dx_j} \qquad (3.9)$$

Για να υπολογίσουμε το dy_j/dx_j παραγωγίζουμε την εξίσωση (3.6) και αντικαθιστώντας έχουμε:

$$\frac{\partial E}{\partial x_j} = \frac{\partial E}{\partial y_j} \cdot y_j\left(1 - y_j\right) \qquad (3.10)$$

Στο σημείο αυτό γνωρίζουμε πως το σφάλμα μπορεί να επηρεαστεί από μια αλλαγή της εισόδου, κατά συνέπεια υπολογίζουμε κατά πόσο μια αλλαγή των βαρών θα επηρεάσει το σφάλμα. Για κάθε βάρος β_{ji} από το i στο j νευρώνα η παράγωγος είναι:

$$\frac{\partial E}{\partial \beta_{ji}} = \frac{\partial E}{\partial x_j} \cdot \frac{\partial x_j}{\partial \beta_{ji}} = \frac{\partial E}{\partial x_j} \cdot y_j \qquad (3.11)$$

Έτσι η επίδραση στην έξοδο είναι:

$$\frac{\partial E}{\partial x_j} \cdot \frac{\partial x_j}{\partial y_j} = \frac{\partial E}{\partial x_j} \cdot \beta_{ji} \qquad (3.12)$$

Για να υπολογίσουμε την επίδραση όλων των συνδέσμων της μονάδας i έχουμε:

$$\frac{\partial E}{\partial y_j} = \sum_i \frac{\partial E}{\partial x_j} \cdot \beta_{ji} \qquad (3.13)$$

Το τελικό βήμα είναι να γίνει αλλαγή των βαρών κατά μια ποσότητα ανάλογη προς $\partial E / \partial \beta$:

$$\Delta\beta = -\varepsilon \frac{\partial E}{\partial \beta} \qquad (3.14)$$

Ο ρυθμός εκμάθησης (ε) προσδιορίζει πόση από την αλλαγή των βαρών θα εφαρμοστεί σε κάθε διαδρομή. Η προσέγγιση αυτή είναι ικανή μόνο να βρίσκει γεγονός που θα είχε σαν αποτέλεσμα μια πρόσκαιρη αύξηση του σφάλματος.

Με σκοπό να ξεφύγει από τα τοπικά ελάχιστα πρέπει να εισαχθεί ένας παράγοντας ορμής:

$$\Delta\beta(t+1) = -\varepsilon \frac{\partial E}{\partial \beta}(t+1) + \alpha\Delta\beta(t) \qquad (3.15)$$

Ο όρος αυτός (α) προσδιορίζει το βαθμό επίδρασης που είχε η προηγούμενη αλλαγή των βαρών. Οι ένδεκα προηγούμενες εξισώσεις είναι οι απαραίτητες για τον αλγόριθμο "back-propagation".

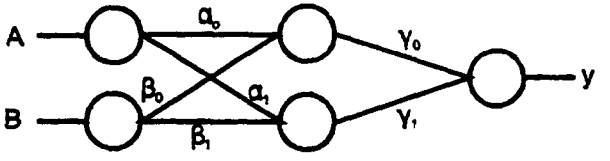
3.1.4 Σύγκριση των νευρωνικών δικτύων με μη παραμετρικά μοντέλα παλινδρόμησης και παρουσίαση της μη γραμμικής συμπεριφοράς τους.

Υποστηρίζεται ότι τα νευρωνικά δίκτυα είναι σε θέση να υπερβούν τα προβλήματα έλλειψης ισχυρού μαθηματικού μοντέλου.

Στην ανάλυση που ακολουθεί μετασχηματίζεται η διαδικασία μάθησης των νευρωνικών δικτύων σε ένα πλαίσιο όμοιο με τη μη γραμμική παλινδρόμηση. Ο μετασχηματισμός αυτός οδηγεί σε μια αναπαράσταση του προσεγγιστικού μαθηματικού μοντέλου με βάση το οποίο το νευρωνικό δίκτυο προσπαθεί να

περιγράψει το πρόβλημα, γεγονός που επιτρέπει τη χρήση της πλούσιας συλλογής αναλυτικών και στατιστικών εργαλείων για τον υπολογισμό της σημαντικότητας των διαφόρων παραμέτρων (βαρών) τον μοντέλου.

Ας θεωρήσουμε το μη γραμμικό εμπρόσθιο δίκτυο του σχήματος 3.7. Για απλούστευση του προβλήματος, θεωρούμε ότι το δίκτυο έχει ένα κρυμμένο επίπεδο δύο κόμβων.



Σχ.3.7

Έστω A και B οι ανεξάρτητες μεταβλητές εισόδου, y η εξαρτημένη μεταβλητή εξόδου, α_o, α_i, β_o, β_i τα βάρη των συνδέσμων μεταξύ των κόμβων εισόδου και του κρυμμένου επιπέδου και γ_o, γ_i τα βάρη των συνδέσμων από το κρυμμένο επίπεδο στον κόμβο εξόδου.

Ο ρόλος της διαδικασίας της εκπαίδευσης του δικτύου είναι ο προσεγγιστικός προσδιορισμός μιας συνάρτησης μεταξύ των διανυσμάτων εισόδου-εξόδου. Η συνάρτηση αυτή εκφράζεται σε παραμετρική μορφή, με παραμέτρους τα βάρη του δικτύου [Kohonen (1984)]

Αν θεωρήσουμε ως συνάρτηση μεταφοράς των κόμβων τη σιγμοειδή, η έξοδος y παίρνει τη μορφή:

$$y = \frac{1}{1+e^{-(\gamma_0 v_0 + \gamma_1 v_1)}} \qquad (3.16)$$

όπου v_o και v_1 οι τιμές των κόμβων του κρυφού επιπέδου. Οι τιμές αυτές μπορούν να γραφούν σε παραμετρική μορφή ως εξής:

$$v_0 = \frac{1}{1+e^{-(\alpha_0 \cdot A + \beta_0 \cdot B)}} \quad \text{και} \quad v_1 = \frac{1}{1+e^{-(\alpha_1 \cdot A + \beta_1 \cdot B)}} \qquad (3.17)$$

Χωρίς να περιορίζεται η γενικότητα τον προβλήματος, αν θεωρήσουμε τον κόμβο εξόδου γραμμικό τότε ισχύει:

$$y=\gamma_0\cdot v_0+\gamma_1\cdot v_1=\gamma_0\cdot\frac{1}{1+e^{-(\alpha_0\cdot A+\beta_0\cdot B)}}+\gamma_1\cdot\frac{1}{1+e^{-(\alpha_1\cdot A+\beta_1\cdot B)}}$$

(3.18)

Σε αρκετές εφαρμογές συνηθίζεται να εφαρμόζεται στις εισόδους και εξόδους ένας μετασχηματισμός εξομάλυνσης πριν από τη φάση της εκπαίδευσης ώστε, για παράδειγμα, να εξαλειφθούν οι επιδράσεις των ακρότατων τιμών (statistical outliers). Ένας κοινός μετασχηματισμός είναι η λογαρίθμηση.

Αντί να υπολογίζεται η συνάρτηση y = f(A,B), χρησιμοποιείται ο μετασχηματισμός ln(y) = f(lnA,lnB). Χρησιμοποιώντας αυτό το μετασχηματισμό οι εκθετικοί όροι στις παραπάνω εξισώσεις μπορούν να γραφούν ως εξής:

$$e^{(\alpha_0\cdot\ln A+\beta_0\cdot\ln B)}=e^{\left(\ln A^{\alpha_0}+\ln B^{\beta_0}\right)}=e^{\ln\left(A^{\alpha_0}\cdot B^{\beta_0}\right)}=A^{\alpha_0}\cdot B^{\beta_0}$$

(3.19)

Λόγω της 3.19 η σχέση 3.18 μπορεί να γραφεί ως

$$\ln y=\gamma_0\cdot\frac{A^{\alpha_0}\cdot B^{\beta_0}}{A^{\alpha_0}\cdot B^{\beta_0}+1}+\gamma_1\cdot\frac{A^{\alpha_1}\cdot B^{\beta_1}}{A^{\alpha_1}\cdot B^{\beta_1}+1}$$

(3.20)

Η συνάρτηση 3.20 περιέχει έξι παραμέτρους. Ο ρόλος της διαδικασίας εκπαίδευσης είναι να υπολογιστούν αυτοί οι παράμετροι έτσι ώστε να ελαχιστοποιείται το ελάχιστο τετραγωνικό σφάλμα. Στη γενική περίπτωση του δικτύου m μεταβλητών εισόδου και n κρυφών κόμβων η (3.20) παίρνει τη μορφή:

(3.21)

$$\ln y=\gamma_0\cdot\frac{A^{\alpha_0}\cdot B^{\beta_0}\cdots M^{\mu_0}}{A^{\alpha_0}\cdot B^{\beta_0}\cdots M^{\mu_0}+1}+\gamma_1\cdot\frac{A^{\alpha_1}\cdot B^{\beta_1}\cdots M^{\mu_1}}{A^{\alpha_1}\cdot B^{\beta_1}\cdots M^{\mu_1}+1}+\ldots+\gamma_n\cdot\frac{A^{\alpha_n}\cdot B^{\beta_n}\cdots M^{\mu_n}}{A^{\alpha_n}\cdot B^{\beta_n}\cdots M^{\mu_n}+1}$$

Η παραπάνω μορφοποίηση της εξόδου τον δικτύου είναι όμοια με τη μορφοποίηση μη γραμμικών πολυπαραμετρικών μοντέλων παλινδρόμησης [Refenes (1993)] και δίνει τη δυνατότητα χρησιμοποίησης των αναλυτικών και στατιστικών εργαλείων που έχουν αναπτυχθεί για αυτά στο πεδίο των νευρωνικών δικτύων.

3.2 Παράμετροι ελέγχου απόδοσης των Τ.Ν.Δ.

Η απόδοση των νευρωνικών δικτύων εκτιμάται ως ο σταθμισμένος μέσος όρος τριών μετρικών εννοιών: της σύγκλισης (convergence), της γενίκευσης (generalization) και της σταθερότητας (stability)

3.2.1 Σύγκλιση - Γενίκευση – Σταθερότητα

Το κριτήριο της σύγκλισης ερευνά αν η διαδικασία μάθησης που εφαρμόζεται είναι ικανή να αποκαλύψει τις συσχετίσεις μεταξύ του συνόλου των δεδομένων εκπαίδευσης (π.χ επιλογή του κατάλληλου αλγορίθμου εκπαίδευσης), κάτω από ποιες καταστάσεις είναι αυτό εφικτό (π.χ έλεγχος των παραμέτρων εκπαίδευσης, αρχιτεκτονική του δικτύου) και ποιες είναι οι υπολογιστικές απαιτήσεις για τη σύγκλιση (π.χ χρόνος εκπαίδευσης).

Ένα νευρωνικό δίκτυο θεωρείται ότι συγκλίνει όταν το συνολικό λάθος εξόδου στα δεδομένα εκπαίδευσης τείνει στο μηδέν, καθώς ο χρόνος εκπαίδευσης τείνει στο άπειρο, ή τουλάχιστον όταν επιτυγχάνει στο σύνολο των δεδομένων εκπαίδευσης το συνολικό λάθος να βρίσκεται εντός ενός ορισμένου εύρους ανοχής.

Η ικανότητα γενίκευσης είναι η πιο σημαντική ιδιότητα των νευρωνικών δικτύων. Μετράει την ικανότητα του νευρωνικού δικτύου να αναγνωρίσει πρότυπα εκτός του συνόλου στο οποίο εκπαιδεύτηκε. Η ιδιότητα αυτή των δικτύων μπορεί να κατανοηθεί καλύτερα παρατηρώντας την αναλογία που υπάρχει μεταξύ της διαδικασίας εκπαίδευσης ενός δικτύου και της διαδικασίας εύρεσης της καμπύλης που περιγράφει καλύτερα τα δεδομένα (curve fitting). Υπάρχουν δυο προβλήματα στην εύρεση αυτής της καμπύλης. Η εύρεση της τάξης του πολυωνύμου που την περιγράφει, και η εύρεση των τιμών των συντελεστών

αυτού του πολυωνύμου. Από τη στιγμή που οι συντελεστές υπολογιστούν μπορεί να γίνει οποιαδήποτε πρόβλεψη, ακόμα και για δεδομένα εκτός του συνόλου εκπαίδευσης. Παρατηρήθηκε ότι ένα νευρωνικό δίκτυο με δομή απλούστερη της απαιτούμενης (π.χ μικρό αριθμό κόμβων στο κρυφό επίπεδο), δεν μπορεί να επιτύχει καλές προσεγγίσεις ακόμα και για τα πρότυπα εντός του συνόλου, ενώ κάποιο με δομή πιο σύνθετη από την απαιτούμενη οδηγεί σε απλή απομνημόνευση των δεδομένων εκπαίδευσης χωρίς να αποκαλύπτει τις πραγματικές συσχετίσεις μεταξύ τους, με αποτέλεσμα την κακή απόδοση σε καινούργια δεδομένα.

Το κριτήριο σταθερότητας ελέγχει την ευαισθησία της εξόδου του δικτύου, όταν μεταβάλλονται οι τιμές των παραμέτρων που επηρεάζουν την απόδοσή του.

Τα νευρωνικά δίκτυα είναι γνωστό ότι παρουσιάζουν μεγάλες διακυμάνσεις στις ιδιότητες πρόβλεψής τους. Για παράδειγμα μικρές αλλαγές, στην αρχιτεκτονική τους, στο χρόνο εκπαίδευσης, στις αρχικές συνθήκες μπορεί να προκαλέσουν μεγάλες αλλαγές στη συμπεριφορά τους.

Στη θεωρία το μόνο κριτήριο σύγκρισης της απόδοσης των νευρωνικών δικτύων και των παραμετρικών μοντέλων παλινδρόμησης είναι το μέσο τετραγωνικό σφάλμα που εμφανίζουν στα δεδομένα του δείγματος. Στην πράξη, κάθε μέθοδος πρόβλεψης απλά προσεγγίζει την πραγματική δομική σχέση μεταξύ εισόδου - εξόδου και πάντα περιέχει κάποιο λάθος στους υπολογισμούς της.

Το λάθος της πρόβλεψης μπορεί να οφείλεται στην απόκλιση της μεθόδου (estimator's bias) και στη διακύμανσή της (estimator's variance).

Τα νευρωνικά δίκτυα είναι ανάλογα με τα μη παραμετρικά μοντέλα παλινδρόμησης διότι δε θέτουν κανέναν, εξαρχής, περιορισμό στο πρόβλημα (π.χ γραμμικότητα) αλλά αφήνουν τα δεδομένα να κατασκευάσουν το δικό τους μοντέλο. Εδώ το μεγαλύτερο ποσοστό του συνολικού λάθους, οφείλεται στη μεγάλη διακύμανση. Ο στόχος είναι να καθοριστούν τα διαστήματα των τιμών των παραμέτρων για τις οποίες το δίκτυο δίνει σταθερά αποτελέσματα για διαφορετικά σύνολα εκπαίδευσης και ελέγχου.

3.2.2 Παράμετροι ελέγχου των μέτρων απόδοσης

Τα μέτρα απόδοσης των νευρωνικών δικτύων, είναι δυνατό να ελεγχθούν με τη βοήθεια των παρακάτω παραμέτρων:

- Την επιλογή της συνάρτησης μεταφοράς των κόμβων του δικτύου
- Το μηχανισμό ελέγχου του αλγορίθμου εκπαίδευσης (ρυθμός μάθησης)
- Την επιλογή της αρχιτεκτονικής του δικτύου
- Το χρόνο εκπαίδευσης και την αρχική επιλογή των τιμών των βαρών
- Την επιλογή της συνάρτησης κόστους

3.2.3 Συναρτήσεις μεταφοράς

Η πρώτη βασική απόφαση κατά τη σχεδίαση ενός δικτύου είναι η επιλογή της συνάρτησης μεταφοράς. Υπάρχουν πολλά είδη συναρτήσεων μεταφοράς για τους κόμβους των νευρωνικών δικτύων. Οι πρώτοι ερευνητές χρησιμοποιούσαν γραμμικές συναρτήσεις ή γραμμικές συναρτήσεις κατωφλίου. Αποδείχθηκε όμως ότι οι δυνατότητες τους ήταν περιορισμένες.

- **Σιγμοειδής ή λογιστική συνάρτηση (sigmoid transfer function)**

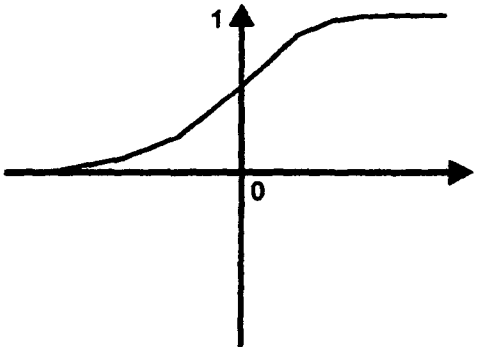
Οι σιγμοειδείς συναρτήσεις είναι οι πιο ευρεία χρησιμοποιούμενες για όλες τις διαδικασίες μάθησης και συμπεριφέρονται καλά στις περισσότερες εφαρμογές καθώς είναι μη γραμμικές και διαφορίσιμες. Υπάρχουν δύο είδη σιγμοειδών συναρτήσεων: Οι ασύμμετρες και οι συμμετρικές. Μια ευρεία χρησιμοποιούμενη ασύμμετρη συνάρτηση με πεδίο τιμών στο διάστημα [0,1] είναι η:

$$f(x)=\frac{1}{1+e^{-x}}$$

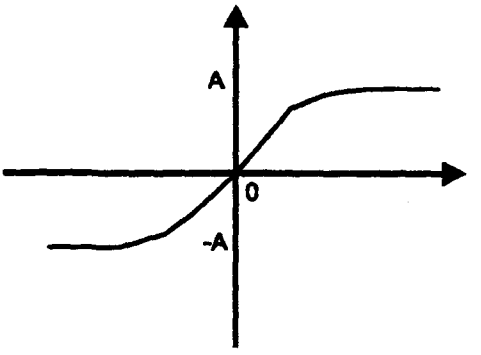
της οποίας η γραφική παράσταση δίνεται στο σχήμα 3.8α.
Τυπικό παράδειγμα συμμετρικής σιγμοειδούς συνάρτησης (σχ. 3.8β) με πεδίο τιμών στο διάστημα [-A, A], αποτελεί η:

$$f(x)=A-\frac{2\cdot A}{1+e^{2\cdot S\cdot x}}$$

όπου A το πλάτος της συνάρτησης (η συνάρτηση παρουσιάζει δυο ασύμπτωτες στο A και -A) και S μια σταθερά που καθορίζει την κλίση της, στην περιοχή κανονικής λειτουργίας.



Σχ. 3.8α Ασύμμετρη σιγμοειδής



Σχ. 3.8β Συμμετρική σιγμοειδής

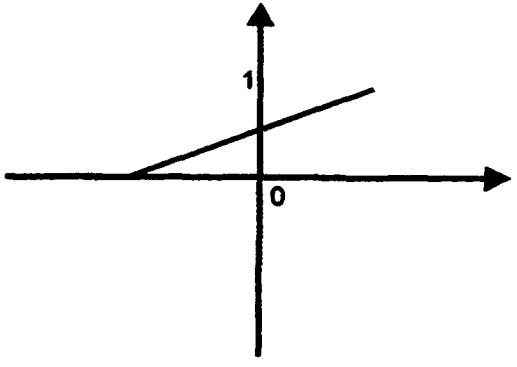
Οι συμμετρικές συναρτήσεις θεωρούνται ότι οδηγούν γρηγορότερα το δίκτυο στη σύγκλιση (convergence), αν και για πολύ μικρές τιμές των βαρών η μάθηση μπορεί να γίνει εξαιρετικά αργή. Η συνάρτηση αυτή έχει χρησιμοποιηθεί σε αρκετές πολύπλοκες εφαρμογές με A=1.7159 και S=2/3 με πολύ καλά αποτελέσματα. Για αυτές τις τιμές των παραμέτρων ισχύει f(1)=1 f(-1)=-1, με αποτέλεσμα για κανονικές συνθήκες λειτουργίας το συνολικό κέρδος να είναι γύρω στο 1, γεγονός που οδηγεί σε απλοποίηση της κατάστασης των νευρώνων και σε πιο γρήγορη μάθηση [Lecun (1989)]. Αν οι αρχικές τιμές των βαρών είναι πολύ μικρές η μάθηση γίνεται πολύ αργή. Το ίδιο συμβαίνει και για πολύ μεγάλες τιμές των βαρών καθώς τότε η σιγμοειδής λειτουργεί στο σημείο κορεσμού, με αποτέλεσμα η παράγωγός της να τείνει στο μηδέν. Για το λόγο αυτό, πριν την εκπαίδευση τα βάρη παίρνουν τυχαίες τιμές στο εύρος [-2.4/Ii, +2.4/Ii], όπου Ii ο αριθμός των εισόδων στον κόμβο που ανήκει ο σύνδεσμος, ώστε η αρχική τυπική απόκλιση του σταθμισμένου αθροίσματος για κάθε κόμβο να είναι στο ίδιο εύρος και να βρίσκεται στην κανονική περιοχή λειτουργίας της σιγμοειδούς.

Θεωρητικά, δεν έχει σημασία ποιος τύπος σιγμοειδούς συναρτήσεως (συμμετρική ή ασύμμετρη) χρησιμοποιείται. Πάντως στην πράξη, ο χρόνος εκπαίδευσης μπορεί να διαφέρει σημαντικά για διαφορετικές συναρτήσεις και δυστυχώς η επιλογή εξαρτάται από τη συγκεκριμένη εφαρμογή.

Αναφέρεται (Refenes (1991)) ότι οι συμμετρικές σιγμοειδείς συναρτήσεις είναι ικανές να βελτιώσουν την ταχύτητα της σύγκλισης σε σχέση με τις ασύμμετρες κατά δέκα φορές

- Γραμμική συνάρτηση (Linear transfer function)

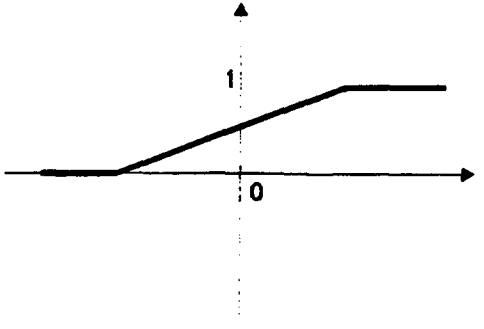
Η γραμμική συνάρτηση μεταφοράς (Σχ. 3.9) είναι χρήσιμη για ελάχιστες εφαρμογές.



Σχ. 3.9 Γραμμική συνάρτηση μεταφοράς

- Γραμμική συνάρτηση κατώφλιου (Linear threshold function)

Στη συνάρτηση αυτή (σχ. 3.10) η έξοδος είναι πολλαπλάσιο της εισόδου σε κάποιο συγκεκριμένο εύρος της. Για όλες τις υπόλοιπες τιμές της εισόδου, η έξοδος παίρνει μια μέγιστη ή ελάχιστη τιμή.

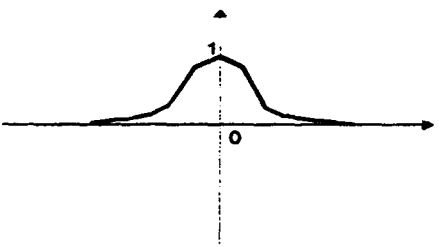


Σχ. 3.10.Γραμμική συνάρτηση μεταφοράς με κατώφλι

Λόγω της μη γραμμικότητάς της, η συμπεριφορά των δικτύων που τη χρησιμοποιούν είναι καλύτερη σε σχέση με εκείνα που χρησιμοποιούν τη γραμμική. Υπάρχουν πάντως σοβαροί περιορισμοί στις σχέσεις εισόδου - εξόδου για τις οποίες μπορούν να εκπαιδευτούν, γι' αυτό και η χρήση της είναι περιορισμένη.

- Συνάρτηση Gauss (Gaussian transfer function)

Η συνάρτηση αυτή (σχ. 3.11) είναι η σπανιότερη επιλογή συνάρτησης μεταφοράς. Δεν είναι μονότονη, όμως είναι συνεχής και διαφορίσιμη γι' αυτό και μπορεί να χρησιμοποιηθεί ως συνάρτηση μεταφοράς.

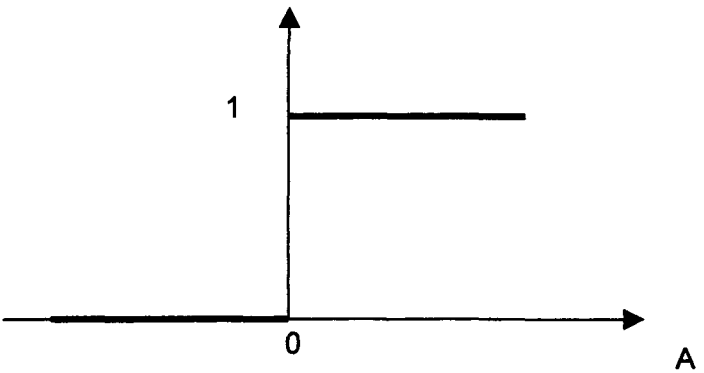


Σχ. 3.11 Συνάρτηση μεταφοράς Gauss

Η εκμάθηση με τη χρήση αυτής της συνάρτησης δεν έχει (απ' όσο γνωρίζουμε) καμιά αναλογία με τη λειτουργία των φυσικών νευρώνων, αναφέρεται όμως ότι ορισμένου τύπου προβλήματα μπορούν να λυθούν με τη χρήση λιγότερων κόμβων από αυτούς που απαιτούνται αν γινόταν χρήση ενός διαφορετικού τύπου συνάρτησης. Λόγω των ελαχίστων εφαρμογών που έχουν αναπτυχθεί με βάση αυτή τη συνάρτηση η συνέπεια των αποτελεσμάτων της δεν είναι εξακριβωμένη και για το λόγο αυτό η χρήση της περιορίζεται συνήθως μόνο σε ερευνητικό επίπεδο.

- Βηματική συνάρτηση μεταφοράς (step transfer function)

Η συνάρτηση αυτή (Σχ. 3.12) έχει δύο μόνο δυνατές τιμές. Για τιμές της



Σχ. 3.12 Βηματική συνάρτηση μεταφοράς.

ανεξάρτητης μεταβλητής μικρότερες από την τιμή κατωφλίου, η τιμή είναι πάντα μηδέν, ενώ για τιμές μεγαλύτερες από την τιμή κατωφλίου είναι πάντα μονάδα.

Το κέντρο της συνάρτησης είναι η τιμή της εισόδου για την οποία η έξοδος μεταφέρεται από την τιμή 0 στην τιμή 1 (αν το A αυξάνεται) ή από την τιμή 1 στην τιμή 0 αν το A ελαττώνεται. Εξαιτίας της ασυνέχειάς της η συνάρτηση είναι μη γραμμική. Τα νευρωνικά δίκτυα που δημιουργούνται με αυτές τις μονάδες παρουσιάζουν πιο ενδιαφέρουσα συμπεριφορά από τα αυτά που δημιουργούνται από γραμμικούς νευρώνες [Lawrence (1992)].

3.2.4 Έλεγχος της διαδικασίας μάθησης

Ο κύριος μηχανισμός ελέγχου του αλγορίθμου μάθησης είναι ο ρυθμός μάθησης λ αφού η τιμή του επηρεάζει το μέγεθος των αλλαγών στα βάρη του δικτύου. Η εύρεση της κατάλληλης τιμής του λ είναι ιδιαίτερα δύσκολη.

Η εφαρμογή μικρού ρυθμού μάθησης που συνεπάγεται μικρές αλλαγές στις τιμές των βαρών, έχει ως αποτέλεσμα τόσο τη μεγάλη αύξηση του απαιτούμενου χρόνου εκπαίδευσης, καθώς και την παγίδευση του δικτύου σε κάποιο τοπικό ελάχιστο της επιφάνειας λάθους (Σχ. 3.13). Η χρήση μιας μεγάλης τιμής του λ συνεπάγεται μεγάλες αλλαγές των βαρών, με αποτέλεσμα αυτά να ταλαντώνονται γύρω από τη σωστή τιμή τους χωρίς να μπορούν να την προσεγγίσουν.

Σχ. 3.13 Ελάχιστα της επιφάνειας λάθους

Υπάρχουν δυο τεχνικές χρήσης του ρυθμού μάθησης κατά την εκπαίδευση ενός δικτύου. Η πρώτη αφορά τη χρήση ενός ρυθμού μάθησης για ολόκληρο το δίκτυο, ενώ η δεύτερη και πιο πολύπλοκη τη χρήση ενός ρυθμού μάθησης για κάθε βάρος χωριστά (ή για κάθε επίπεδο). Στην τελευταία περίπτωση ένας εμπειρικός τρόπος χειρισμού του λ είναι ο εξής:

Αν διαδοχικές αλλαγές ενός βάρους έχουν το ίδιο πρόσημο, τότε ο ρυθμός μάθησης που σχετίζεται με αυτό το βάρος πρέπει να αυξηθεί.

Αν διαδοχικές αλλαγές ενός βάρους έχουν αντίθετο πρόσημο (δηλαδή η τιμή του βάρους ταλαντώνεται γύρω από τη βέλτιστη τιμή), τότε ο ρυθμός μάθησης που σχετίζεται με αυτό το βάρος πρέπει να μειωθεί.

Ένα μειονέκτημα της χρήσης πολλών μεταβλητών ρυθμών μάθησης είναι η αύξηση της πολυπλοκότητας του αλγορίθμου μάθησης, γιατί εκτός από το πρόβλημα της προσαρμογής των βαρών, προκύπτει και το πρόβλημα προσαρμογής των λ. Σε γενικές γραμμές πάντως, η δυνατότητα δυναμικού επηρεασμού του λ πρέπει να εκτιμηθεί ως πλεονέκτημα και όχι ως επιπλέον δυσκολία.

3.2.5 Η αρχιτεκτονική του δικτύου

Ως αρχιτεκτονική ενός νευρωνικού δικτύου ορίζουμε τον αριθμό των κόμβων που το απαρτίζουν και τον τρόπο σύνδεσης αυτών μεταξύ τους.

Όπως έχει ήδη αναφερθεί ένα νευρωνικό δίκτυο είναι ιεραρχικά οργανωμένο σε επίπεδα. Πλήρες συνδεδεμένο, θεωρείται το δίκτυο στο οποίο κάθε κόμβος ενός επιπέδου είναι συνδεδεμένος με κάθε κόμβο του ανώτερου και του

κατώτερου επιπέδου. Υπάρχουν δίκτυα, η αρχιτεκτονική των οποίων επιτρέπει τη ροή των υπολογισμών και προς τις δύο κατευθύνσεις. Στα εμπρόσθια νευρωνικά δίκτυα, η ροή των δεδομένων λαμβάνει χώρα από τα κατώτερα προς τα ανώτερα επίπεδα. Το κατώτερο επίπεδο ονομάζεται επίπεδο εισόδου, ενώ τα ενδιάμεσα κρυφά επίπεδα. Η τοπολογία του δικτύου επηρεάζει σε μεγάλο βαθμό τις νοητικές ιδιότητες του δικτύου και ειδικότερα τη σημαντικότερη ιδιότητα του, τη γενίκευση. Θεωρητικές μελέτες υποδεικνύουν, ότι η πιθανότητα καλής γενίκευσης εξαρτάται από το μέγεθος του χώρου υπόθεσης (δηλαδή, το σύνολο των δυνατών υποδικτύων), το μέγεθος του χώρου λύσεων (δηλ. το σύνολο των δικτύων που δίνουν καλή γενίκευση) και από τον αριθμό των διανυσμάτων εκπαίδευσης [Denker, Wittner (1987)]. Αν για παράδειγμα ο χώρος υπόθεσης είναι μεγάλος και ο αριθμός των προτύπων εκπαίδευσης πολύ μικρός, τότε θα υπάρχει ένας τεράστιος αριθμός δικτύων που περιγράφει σωστά αυτά τα πρότυπα, αλλά μόνο ένας πολύ μικρός αριθμός από αυτά θα ανήκει στο χώρο λύσεων, με αποτέλεσμα η πιθανότητα γενίκευσης να είναι πολύ μικρή.

Ο πιο κοινός τρόπος ελάττωσης των ελεύθερων παραμέτρων του δικτύου που οδηγεί και στη συρρίκνωση του χώρου υπόθεσης και κατά συνέπεια σε καλύτερες ιδιότητες γενίκευσης, είναι η μείωση του μεγέθους του δικτύου. Η μείωση αυτή πρέπει να γίνει μέχρι εκείνου τον σημείου που οι εναπομείναντες ελεύθεροι παράμετροι να είναι σε θέση να περιγράψουν τη συνάρτηση σύνδεσης εισόδου - εξόδου των δεδομένων.

Υπάρχουν τέλος διάφορες τεχνικές που οδηγούν στη μείωση των ελευθέρων παραμέτρων και άρα σε καλύτερη γενίκευση, χωρίς τη μείωση του μεγέθους του δικτύου [Rumelhart,et al (1986),(1988)].

3.2.6 Χρόνος εκπαίδευσης και αρχική εκτίμηση των βαρών

Ως χρόνος εκπαίδευσης, θεωρείται ο αριθμός επαναληπτικών εμφανίσεων των δεδομένων στο δίκτυο. Καθώς η διαδικασία της εκπαίδευσης εξελίσσεται και δοθέντος ότι το δίκτυο έχει επαρκή αριθμό ελευθέρων παραμέτρων (κρυφών κόμβων), το συνολικό λάθος εξόδου στα δεδομένα εκπαίδευσης συνεχώς θα μειώνεται. Για τον προσδιορισμό του χρονικού σημείου τερματισμού

της εκπαίδευσης, πριν το δίκτυο απλά απομνημονεύσει όλα τα πρότυπα και χάσει τις ιδιότητες γενίκευσης, μπορεί να εφαρμοστεί η ακόλουθη διαδικασία:

Τα διαθέσιμα δεδομένα χωρίζονται τυχαία σε δύο σύνολα, ένα σύνολο εκπαίδευσης και ένα σύνολο ελέγχου. Κατά την πρόοδο της εκπαίδευσης παρατηρείται η καμπύλη συνολικού λάθους στα δεδομένα ελέγχου (δεδομένα τα οποία το δίκτυο αντιμετωπίζει για πρώτη φορά). Η καμπύλη αυτή είναι αρχικά φθίνουσα και σε κάποια χρονική στιγμή γίνεται αύξουσα. Το σημείο καμπής αυτής της καμπύλης είναι και το σημείο τερματισμού της εκπαίδευσης.

Τέλος, ιδιαίτερη προσοχή πρέπει να δοθεί στις αρχικές τιμές των βαρών του δικτύου. Ο αλγόριθμος μάθησης "backpropagation" απαιτεί άνισες τιμές εκκίνησης. Αν όλα τα βάρη ξεκινήσουν με τις ίδιες τιμές, είναι πολύ πιθανό το δίκτυο να μη μάθει ποτέ. Για το λόγο αυτό κατά την εκκίνηση της διαδικασίας εκπαίδευσης, τα βάρη λαμβάνουν τυχαίες τιμές.

3.3 Μέθοδοι για βέλτιστη σχεδίαση της αρχιτεκτονικής του δικτύου.

3.3.1 Θεωρίες στη σχεδίαση της αρχιτεκτονικής των νευρωνικών δικτύων

Διάφορες μέθοδοι έχουν προταθεί τα τελευταία χρόνια σχετικά με το πρόβλημα της σχεδίασης και τον καθορισμό της αρχιτεκτονικής που θα εξασφαλίζει τη γρήγορη εκπαίδευση και την προσαρμογή του δικτύου στα δεδομένα εκπαίδευσης. Οι μέθοδοι αυτοί μπορούν να ταξινομηθούν σε τρεις γενικές κατηγορίες:

- Τεχνικές αναλυτικού υπολογισμού
- Τεχνικές σταδιακής ανάπτυξης (constructive techniques)
- Τεχνικές κλαδέματος (pruning techniques).

Οι μέθοδοι της πρώτης κατηγορίας στόχο έχουν να καθορίσουν εκ των προτέρων τον αριθμό των κρυφών κόμβων που απαιτούνται για την επίλυση κάποιου συγκεκριμένου προβλήματος, αναλύοντας το μέγεθος και τη διάσταση του χώρου των διανυσμάτων εισόδου με τη χρήση στατιστικών εργαλείων.

Στη δεύτερη κατηγορία περιλαμβάνονται αλγόριθμοι που επιχειρούν τη δόμηση του δικτύου σταδιακά προσθέτοντας κατά τη διάρκεια της εκπαίδευσης νέους κόμβους κατά τέτοιο τρόπο ώστε κάθε νέα προσθήκη να οδηγεί σε μείωση του σφάλματος.

Τέλος οι τεχνικές κλαδέματος αφορούν μεθόδους που λειτουργούν αντίστροφα με τις προηγούμενες και προσπαθούν να μειώσουν το μέγεθος του δικτύου αφαιρώντας κλάδους ανάλογα με τη συνεισφορά τους στην έξοδο.

Όλες αυτές οι τεχνικές ανεξαρτήτως πολυπλοκότητας, παρ' ότι εξασφαλίζουν την εκμάθηση των σχέσεων μεταξύ των δεδομένων εκπαίδευσης (convergence), δεν μπορούν να εγγυηθούν την ικανότητα γενίκευσης (generalization) του δικτύου, δηλαδή την απόκριση του σε πρωτοεμφανιζόμενα γεγονότα.

3.3.2 Τεχνικές αναλυτικού υπολογισμού

Σύμφωνα με τις απόψεις αρκετών ερευνητών ο αριθμός των κόμβων των κρυφών επιπέδων του δικτύου αποτελεί συνάρτηση του αριθμού των διανυσμάτων εκπαίδευσης. Διάφοροι κανόνες που επιχειρούν μια προσέγγιση του αριθμού αυτού βασισμένοι στη στατιστική θεωρία ταξινόμησης (classification theory) έχουν προταθεί, όπως για παράδειγμα:

- Ο αριθμός των συνολικών συνδέσεων μεταξύ των κόμβων πρέπει να είναι μικρότερος από το ένα δέκατο του συνόλου των δεδομένων εκπαίδευσης.
- Ο αριθμός των κρυφών κόμβων του δικτύου (H) είναι της τάξης του $\log_2 T$ όπου (T) ο αριθμός των δεδομένων εκπαίδευσης.

Οι παραπάνω κανόνες δε λαμβάνουν υπόψη τη διάσταση του χώρου της εισόδου και θεωρούν ότι ο αριθμός των κόμβων των κρυφών επιπέδων εξαρτάται αποκλειστικά από τον αριθμό των κόμβων εισόδου.

Οι Mirchandani και Cao απόδειξαν ότι ο d-διάστατος χώρος εισόδου είναι δυνατό να χωριστεί σε M περιοχές, οι οποίες μπορούν να συγχωνευτούν σε C ομάδες έτσι ώστε $C \leq M$. Ο ρόλος των συνδέσεων μεταξύ του κρυφού επιπέδου και των κόμβων εξόδου είναι να συνδυάσουν τις ομάδες στις περιοχές στις οποίες ανήκουν. Ο αριθμός των περιοχών M, ισούται με τον ελάχιστο αριθμό των δεδομένων εκπαίδευσης (T) που απαιτούνται για την εκπαίδευση του δικτύου, δηλ $T > M$.

Οι Mirchandani και Cao προτείνουν την σχέση (3.22) μεταξύ των M (αριθμός (#) περιοχών), H (# κρυμμένων κόμβων) και d (διάσταση χώρου εισόδου):

$$M(H,d) = \sum_{k=1}^{d} \left(\frac{H}{k}\right) \qquad (3.22)$$

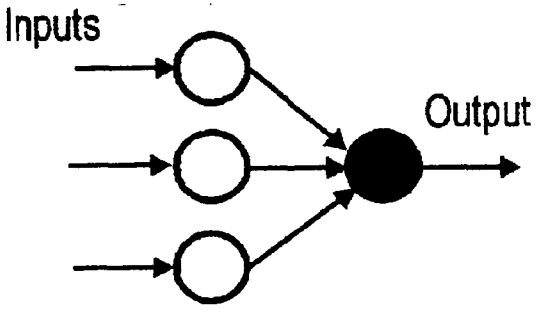
Υποστηρίζεται επίσης, ότι οι βαθμοί ελευθερίας df του δικτύου, που ορίζονται από την εξίσωση (3.23), μπορεί να καθορίσουν ένα κατώτατο όριο δεδομένων εκπαίδευσης που πρέπει να χρησιμοποιηθούν, ώστε το δίκτυο να εκπαιδευτεί αποκτώντας ικανοποιητικές ιδιότητες γενίκευσης.(Caldwell (1994))

df = (# κόμβων στρώματος εισόδου+# κόμβων στρώματος εξόδου) * (#κόμβων κρυφού στρώματος) (3.23)

Οι τεχνικές του αναλυτικού υπολογισμού έχουν το μειονέκτημα ότι προϋποθέτουν ανάλυση για την εύρεση της διάστασης και των περιοχών του χώρου εισόδου.

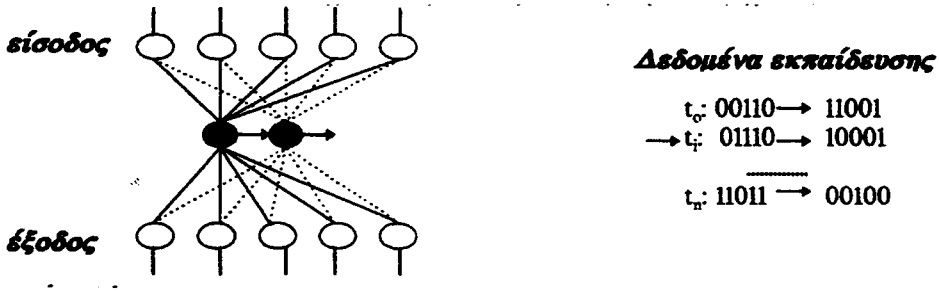
3.3.3 Τεχνικές σταδιακής ανάπτυξης

Η φιλοσοφία σχεδίασης των δικτύων βάση των τεχνικών της κατηγορίας αυτής στηρίζεται στη σταδιακή προσθήκη κόμβων κατά τη διάρκεια της εκπαίδευσης. Η ανάπτυξη του δικτύου γίνεται πάντα με την προσθήκη του ίδιου συνόλου στοιχείων, που αποτελείται από έναν κόμβο, τους κλάδους εισόδου στον κόμβο με τα βάρη τους w_i και έναν κλάδο εξόδου με βάρος w_0 (Σχ.3.14).



Σχ. 3.14 Προσθήκη κόμβου σε υπάρχον δίκτυο

Οι τρόποι σύνδεσης του νέου συνόλου στοιχείων στο δίκτυο μπορεί να γίνει με τους δύο τρόπους του σχήματος 3.15



Σχ. 3.15

Το νέο μπλοκ προσθήκης εκπαιδεύεται κατά τέτοιο τρόπο ώστε να μεγιστοποιείται ή να ελαχιστοποιείται η συσχέτιση της εξόδου του με το λάθος που εμφανίζει κάποιος συγκεκριμένος κόμβος του δικτύου. Με τον τρόπο αυτό το μπλοκ προσθήκης αναπτύσσει διορθωτικά βάρη που εξουδετερώνουν το λάθος [Lippman R.P (1987)].

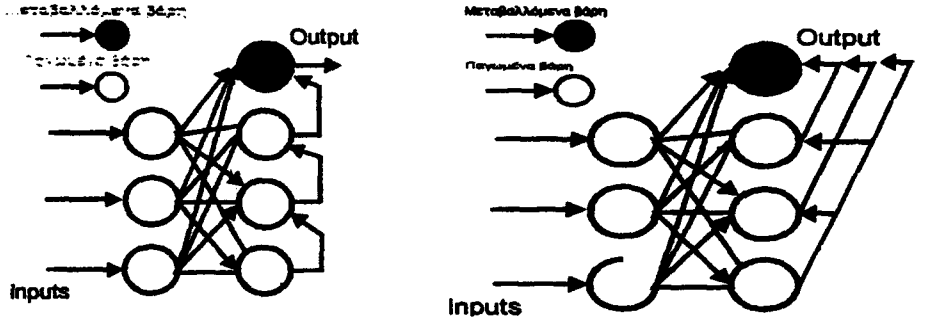
Η σύγκλιση του δικτύου με τη σταδιακή προσθήκη νέων κόμβων αποδεικνύεται μαθηματικά σύμφωνα με το παρακάτω θεώρημα:

Θεώρημα: **Έστω** $f(x) = k + \frac{c}{1+e^{Tx}}$ **η σιγμοειδής συνάρτηση. Για κάθε x, c, y, ε, k πραγματικούς με y-ε-k ≠ 0, αν f(x) = y τότε υπάρχει** $P = \frac{1}{T} \ln \frac{c}{y-\varepsilon-k} - x$, **τέτοιο ώστε f(x+P) = y-ε.**

Ένας από τους αλγόριθμους που υλοποιεί τις τεχνικές της σταδιακής ανάπτυξης είναι ο αλγόριθμος CLS (Constructive Learning by Specialization).

Αλγόριθμος CLS

Ο αλγόριθμος ξεκινά με ένα μόνο πλήρες συνδεμένο κόμβο στο κρυμμένο επίπεδο και με τυχαία βάρη στις συνάψεις του Σχ.3.16.



Σχ.3.16.

Βήμα 1. Πέρασμα του δικτύου προς τα εμπρός με είσοδο το διάνυσμα εκπαίδευσης x_i. Αν το πραγματικό διάνυσμα εξόδου διαφέρει από το επιθυμητό (10001) σε κάποια συγκεκριμένη έξοδο y_r τότε πρόσθεσε ένα νέο κόμβο $h_κ$. Σύνδεσε τον κόμβο $h_κ$ στη συγκεκριμένη έξοδο y_r με τα κατάλληλα βάρη, έτσι ώστε να εξουδετερώνεται το λάθος. Τα κατάλληλα βάρη μπορούν να υπολογιστούν αναλυτικά ως εξής:

Πριν από την προσθήκη του νέου κόμβου $h_κ$ η έξοδος O_r δίνεται από τη σχέση:

$$Y_r = f(\sum_{i=0}^{k-1} \beta_{i,0} x_{i,0}) = d_i \pm \varepsilon \qquad (3.24)$$

όπου d_i, η επιθυμητή τιμή της εξόδου αυξημένη ή ελαττωμένη κατά μια ποσότητα λάθους ε.

Μετά από την προσθήκη του νέου κόμβου $h_κ$ το λάθος ε στη συγκεκριμένη έξοδο θα εξαλειφθεί, δηλ. η Y_r θα δίνεται από τη σχέση (3.25):

$$Y_r = f(\sum_{i=0}^{k-1} \beta_{h_{i,0}} x_{h_{i,0}} + \beta_{h_{k,0}} x_{h_{k,0}}) = d_i \qquad (3.25)$$

Χρησιμοποιώντας ως συνάρτηση μεταφοράς τη σιγμοειδή συνάρτηση $f(x) = A - \frac{2A}{1+e^{2Sx}}$ και λύνοντας ως προς $\beta_{hk,0} \cdot x_{hk,0}$ προκύπτει η (3.26):

$$\beta_{h_{k,0}} \cdot x_{h_{k,0}} = \frac{\ln\left(\frac{2A}{A-d_i} - 1\right) - 2S\sum_{i=0}^{k-1} \beta_{h_{i,0}} \cdot x_{h_{i,0}}}{2S} \qquad (3.26)$$

Η εξίσωση (3.26) έχει δύο αγνώστους. Αν δώσουμε στη $X_{h_{k,0}}$ (έξοδο του κόμβου h_k) μια τιμή ώστε η έξοδος του δικτύου να είναι η επιθυμητή μπορούμε να υπολογίσουμε την απαιτούμενη τιμή του βάρους $\beta_{h_{k,0}}$ και το αντίστροφο.

Βήμα 2. Σύνδεσε το νέο κόμβο h_k σε όλες τις εισόδους. Τα βάρη των συνδέσεων πρέπει να είναι τέτοια ώστε ο νέος κόμβος να αναγνωρίζει το τρέχον διάνυσμα εκπαίδευσης x_i και κανένα από τα προηγούμενα $x_0, x_1, \ldots, x_{i-1}$.

Αυτό μπορεί να επιτευχθεί εκπαιδεύοντας το νέο κόμβο h_k, έτσι ώστε:

1. Να ενεργοποιείται στο πρότυπο x_i
2. Να απενεργοποιείται στα προηγούμενα πρότυπα $x_0, x_1, \ldots, x_{i-1}$.
3. Να ενεργοποιείται για τουλάχιστο ένα από τα επόμενα πρότυπα, $x_{i+1}, \ldots, x_n$.

Η τρίτη συνθήκη χρησιμοποιείται για να αυξήσει τις ικανότητες γενίκευσης του δικτύου [Sietsma (1991)]

3.3.4 Τεχνικές κλαδέματος (prunning techniques)

Η φιλοσοφία των τεχνικών κλαδέματος είναι ακριβώς η αντίθετη με αυτή των τεχνικών σταδιακής ανάπτυξης. Ξεκινώντας από ένα υπερμεγέθες δίκτυο, αναζητούνται τα βάρη με τη μικρότερη συμμετοχή στη λύση του προβλήματος και απομακρύνονται σταδιακά από το δίκτυο. Δύο από τις πιο διαδεδομένες τεχνικές της κατηγορίας αυτής, περιγράφονται στα παρακάτω.

- **Κλάδεμα σε δύο φάσεις (Two-stage Prunning)**

Στόχος της τεχνικής αυτής αποτελεί η εύρεση του ελαχιστοποιημένου δικτύου που μπορεί να επιτελέσει ένα συγκεκριμένο έργο, κλαδεύοντας ένα υπάρχον εκπαιδευμένο δίκτυο. Αυτό επιτυγχάνεται ως εξής: όταν το αρχικό δίκτυο εκπαιδευτεί και μπορεί να επιλύσει το πρόβλημα, αναλύονται οι έξοδοι των κόμβων του κρυφού επιπέδου ώστε να διαπιστωθεί η συμμετοχή τους στη λύση. Αν η έξοδος κάποιου συγκεκριμένου κόμβου δε μεταβάλλεται σημαντικά κατά την εναλλαγή των δεδομένων εκπαίδευσης στην είσοδο του δικτύου, τότε ο κόμβος αυτός δε συνεισφέρει στη λύση. Επίσης αν οι έξοδοι δύο κόμβων είναι συνεχώς ίσες ή αντίθετες, τότε ο ρόλος τους στη λύση είναι κοινός και ο ένας μπορεί να αφαιρεθεί. Υποστηρίζεται ότι δεν έχουμε καμιά απώλεια πληροφορίας προς το επόμενο επίπεδο λόγω της απομάκρυνσης ενός κόμβου με τα παραπάνω χαρακτηριστικά.

- **Ανάλυση Ευαισθησίας (Weight-Error Sensitivity)**

Η φιλοσοφία της τεχνικής αυτής βασίζεται στον υπολογισμό της ευαισθησίας της συνάρτησης λάθους ως προς κάθε κλάδο του δικτύου και κατόπιν στην αφαίρεση των συνδέσμων του δικτύου με τη μικρότερη ευαισθησία. Η ευαισθησία S_{ij} ενός συνδέσμου με βάρος w_{ij} μπορεί να ορισθεί ως εξής:

$$S_{ij} = E(w_{ij} = 0) - E(w_{ij} = w_{ij}^f) \qquad (3.27)$$

όπου w_{ij}^f το βάρος του συνδέσμου μετά την ολοκλήρωση της εκπαίδευσης. Το λάθος E εκφράζεται ως συνάρτηση μόνο του βάρους w του συνδέσμου την ευαισθησία του οποίου θέλουμε να υπολογίσουμε, υποθέτοντας ότι τα υπόλοιπα βάρη παραμένουν σταθερά (στις τελικές τιμές που αποκτούν μετά το πέρας της εκπαίδευσης. Η εξίσωση (3.27) μπορεί να γραφεί ως εξής:

$$S = -\frac{E(w^f) - E(0)}{w^f - 0} \cdot w^f \cong -\sum_{n=0}^{N-1} \frac{\partial E}{\partial w} \Delta w_{ij}(n) \cdot \frac{w_{ij}^f}{w_{ij}^f - w_{ij}^i} \qquad (3.28)$$

όπου N ο αριθμός των προτύπων εκπαίδευσης. Από τη συνάρτηση ευαισθησίας (3.28), οι όροι της οποίας έχουν ήδη υπολογιστεί κατά τη διαδικασία εκπαίδευσης, μπορούμε να υπολογίσουμε την ευαισθησία του κάθε συνδέσμου. Με βάση την παραπάνω ανάλυση ο αλγόριθμος απλοποίησης του δικτύου που μπορεί να χρησιμοποιηθεί είναι ο εξής:

- Ορίζεται μια παράμετρος ελάχιστης αποδεκτής ευαισθησίας Θ κοινή για όλους τους συνδέσμους.
- Υπολογίζεται η ευαισθησία (κατά τη διάρκεια της εκπαίδευσης) σύμφωνα με την εξίσωση (3.30) και κατασκευάζεται ένα αρχείο με τα μέτρα της για κάθε σύνδεσμο του δικτύου.
- Καθορίζονται οι σύνδεσμοι για τους οποίους ισχύει $S_{ij} \leq \Theta$ και αφαιρούνται από το δίκτυο.

Πλεονεκτήματα τον αλγορίθμου αυτού αποτελούν η μικρή πολυπλοκότητα του και κυρίως ο μικρός υπολογιστικός χρόνος που απαιτεί καθώς όλοι οι υπολογισμοί του γίνονται παράλληλα με τη διαδικασία εκπαίδευσης.

3.4 Η Δομή του κοινού εμπρόσθιου Τ.Ν.Δ.
(Standard Feedforward Neural Network Architecture)

- Εύρεση του απαιτούμενου αριθμού προτύπων εκπαίδευσης.

Μέσω της χρήσης ενός πακέτου ανάπτυξης νευρωνικών δικτύων είναι πολύ εύκολο να πειραματιστούμε με μια πληθώρα αρχιτεκτονικών. Μερικά μάλιστα προτείνουν μια συγκεκριμένη δομή (το «P.N.A», το οποίο χρησιμοποιείται στην παρούσα εργασία, προτείνει ένα κρυμμένο επίπεδο με δέκα κόμβους). Το πρόβλημα που προκύπτει αρχικά στην επεξεργασία ενός συνόλου δεδομένων είναι ο καθορισμός του αριθμού των προτύπων εκπαίδευσης που πρέπει να χρησιμοποιηθεί ώστε να εκπαιδευθεί το υπάρχον δίκτυο.

Ένας κανόνας που συσχετίζει τον αριθμό των προτύπων εκπαίδευσης (Training set) με τον αριθμό των βαρών του δικτύου για μια μέγιστη επιθυμητή τιμή της συνάρτησης του μέσου λάθους (Average Error) στο σύνολο των προτύπων εκπαίδευσης, δίνεται από την εξής ανισότητα:

$$M > W/E \qquad (3.31)$$

όπου M είναι ο αριθμός των προτύπων εκπαίδευσης, W ο αριθμός των βαρών του δικτύου και E η μέγιστη επιθυμητή τιμή του μέσου λάθους. Σε πολλές περιπτώσεις γίνεται χρήση της ανισότητας (3.31) για τον υπολογισμό του αριθμού των κόμβων του κρυφού επιπέδου.

- **Εύρεση του αριθμού κόμβων του κρυφού επιπέδου**

Δεν έχει ακόμη βρεθεί κάποιος γενικός κανόνας καθώς ο αριθμός αυτός φαίνεται ότι εξαρτάται σε μεγάλο βαθμό από την πολυπλοκότητα του εκάστοτε προβλήματος.

Η χρήση ενός μεγάλου αριθμού κρυφών κόμβων στο κρυμμένο επίπεδο μπορεί να ωθήσει το δίκτυο στην απομνημόνευση των προτύπων εκπαίδευσης, καταστρέφοντας τις ιδιότητες της γενίκευσης. Αντίθετα η χρήση μικρού αριθμού κρυφών κόμβων καθιστά αδύνατη την εκμάθηση των προτύπων εκπαίδευσης. Η προσέγγιση του σωστού αριθμού στην πράξη μπορεί να επιτευχθεί με μια από τις παρακάτω μεθόδους:

A) Με την εκπαίδευση πολλών δικτύων με διαφορετικό αριθμό κρυφών κόμβων και την επιλογή αυτού που επιτυγχάνει καλύτερα στα δεδομένα ελέγχου (testing set) [Siriopoulos (1990,1995)].

B) Ξεκινώντας από ένα μικρό αριθμό κρυφών κόμβων, προστίθενται νέοι κόμβοι κατά τη διάρκεια της εκπαίδευσης όταν αυτό απαιτείται. Ο αριθμός των κρυφών κόμβων κατά την εκκίνηση της διαδικασίας καθορίζεται με έναν από τους παρακάτω τρεις τρόπους:

1. Να είναι ίσος με το ημιάθροισμα των εισόδων I και των εξόδων O [(I+O) / 2].
2. Να είναι ίσος με το 5-10% του αριθμού των προτύπων εκπαίδευσης.
3. Να είναι μεταξύ του (I+O)/2 και του max(I,O)
4. Να είναι μεταξύ του I/2 και του 2N, όπου N ο αριθμός των νευρώνων των εισροών.
5. Με τη χρήση της σχέσης (3.31).

Η εφαρμογή της δεύτερης μεθόδου προϋποθέτει την παροχή από το χρησιμοποιούμενο λογισμικό της δυνατότητας προσθήκης νέων κόμβων χωρίς να επηρεάζονται οι τιμές των βαρών που έχουν δημιουργηθεί μέχρι τώρα.

Η εκπαίδευση ξεκινά με ένα μεγάλο αριθμό κρυφών κόμβων. Στην πορεία απαλείφονται οι κόμβοι που δε συνεισφέρουν σημαντικά στη λύση και η εκπαίδευση συνεχίζεται. Η μέθοδος αυτή είναι μια πρακτική εφαρμογή των τεχνικών κλαδέματος (prunning) που αναπτύχθηκαν παραπάνω. Η εμπειρία φανερώνει ότι η εκκίνηση με ένα μικρό αριθμό κρυφών κόμβων, η προσθήκη νέων και τέλος το κλάδεμα του δικτύου είναι μια διαδικασία που μπορεί να οδηγήσει στη γρήγορη και ικανοποιητική εκπαίδευση του δικτύου.

- **Χρήση πολλών επιπέδων κρυφών κόμβων.**

Δεν έχει αποδειχτεί ότι η χρήση περισσοτέρων του ενός επιπέδου επιφέρει καλύτερα αποτελέσματα. Τα δίκτυα αυτού του τύπου απαιτούν πολύ περισσότερο χρόνο εκπαίδευσης, καθώς οι σύνδεσμοι είναι περισσότεροι με αποτέλεσμα οι διορθώσεις των βαρών που απαιτούνται σε κάθε πέρασμα των δεδομένων να είναι πιο χρονοβόρες.

3.5 Μέθοδοι επεξεργασίας των δεδομένων εισόδου

Ως εργαλεία επιλογής των δεδομένων μπορούν να χρησιμοποιηθούν διάφορα στατιστικά μοντέλα, όπως η ανάλυση συσχέτισης και η παραμετρική παλινδρόμηση.

Πρέπει να καθορισθεί η φύση των ανεξάρτητων μεταβλητών εισόδου, δηλαδή αν πρόκειται για μεταβλητές πρόβλεψης (predictive variables), ή για μεταβλητές πληροφόρησης (informative variables). Ως μεταβλητή πρόβλεψης, μπορεί να θεωρηθεί αυτή που εξηγεί σε μεγάλο βαθμό τη διακύμανση της μεταβλητής εξόδου. Ως μεταβλητές πληροφόρησης μπορεί να θεωρηθούν εκείνες που μεμονωμένα δεν έχουν μεγάλη ικανότητα πρόβλεψης της μεταβλητότητας της εξόδου αλλά, όταν συνδυαστούν με άλλες μπορούν να οδηγήσουν σε καλύτερη πρόβλεψη.

3.5.1 Επιλογή μεταβλητών εισόδου και εξόδου και μέγεθος δείγματος.

Η ελαχιστοποίηση του αριθμού των μεταβλητών εισόδου έχει ένα διπλό ευεργετικό αποτέλεσμα:

Α) ελαχιστοποιείται ο χρόνος απόκρισης του συστήματος,

Β) ελαχιστοποιείται το μέγεθος της πολυπλοκότητας του δικτύου.

Υπάρχουν για την κατάλληλη επιλογή των μεταβλητών, κυρίως εισόδου, διάφορες τεχνικές, εμπειρικές, στατιστικές, ή μαθηματικές. Μια ευρύτερα γνωστή μέθοδος είναι η *ανάλυση ευαισθησίας*:

Μετά την επιλογή των μεταβλητών εισόδου και την εκπαίδευση του δικτύου, επιλέγεται μια από τις μεταβλητές και θεωρείται ίση με τη μέση τιμή της. Στη συνέχεια το δίκτυο εκπαιδεύεται από την αρχή και συγκρίνονται τα αποτελέσματα. Η διαδικασία αυτή επαναλαμβάνεται αρκετές φορές για τις διάφορες μεταβλητές. Με τον τρόπο αυτό, μπορεί να προσδιορισθεί ο βαθμός που κάθε μεταβλητή επηρεάζει το επιθυμητό αποτέλεσμα από το δίκτυο, δηλαδή την ικανότητα πρόβλεψης κάθε μεταβλητής εισόδου ως προς τη μεταβλητή εξόδου.

Μια άλλη στατιστική μέθοδος είναι η ανάλυση συσχέτισης. Μια τεχνική ελαχιστοποίησης των μεταβλητών εισόδου είναι να παραλείψουμε τις ισχυρά συσχετιζόμενες μεταβλητές.

Από τις βασικότερες χαρακτηριστικές ιδιότητες ενός ΤΝΔ είναι η ικανότητα του να γενικεύει. Για την επίτευξη αυτού του στόχου είναι χρήσιμο να έχει επιλεγεί το άριστο μέγεθος δείγματος παρατηρήσεων πάνω στο οποίο θα εκπαιδευτεί το δίκτυο. Το μέγεθος αυτό πρέπει να είναι τέτοιο ώστε να ικανοποιεί τα κριτήρια στατιστικής σημαντικότητας. Ο Caldwell (1994) προτείνει σαν κάτω όριο (το ελάχιστο δείγμα) να είναι οι βαθμοί ελευθερίας του δικτύου. Συγκεκριμένα θα πρέπει το μέγεθος των εκπαιδευόμενων στοιχείων να είναι μεγαλύτερο από τον αριθμό των βαθμών ελευθερίας όπως αυτός ορίστηκε με τη σχέση (3.23).

3.5.2 Κανονικοποίηση

Κανονικοποίηση ονομάζεται η διαδικασία καθορισμού του δυνατού εύρους τιμών που μπορεί να πάρει μια μεταβλητή. Η κανονικοποίηση πραγματοποιείται για να αποφευχθούν πιθανοί συντονισμοί των παραμέτρων του δικτύου για κάποιο εύρος τιμών εισόδου – εξόδου, καθώς και για να φέρει τις εισόδους της συνάρτησης μεταφοράς των κόμβων μέσα στο εύρος της κανονικής της λειτουργίας. Για παράδειγμα αν θεωρήσουμε τη σιγμοειδή συνάρτηση μεταφοράς με εύρος [0, 1], όταν η είσοδός της βρίσκεται έξω από την κανονική περιοχής λειτουργίας της (εκτός γραμμικής περιοχής), η έξοδός της θα τείνει ασυμπτωτικά *στο μηδέν ή στο ένα. Στην περίπτωση αυτή, η παράγωγος της σιγμοειδούς* και η τιμή ενεργοποίησης του νευρώνα τείνει στο μηδέν. Η κατάσταση αυτή είναι ανεπιθύμητη καθώς οδηγεί το δίκτυο σε μια εικονική αμεταβλητότητα που είναι γνωστή ως «παράλυση του δικτύου». Επιπρόσθετα, είναι ανεπιθύμητο να υπάρχουν μεγάλες διαφορές ανάμεσα στο εύρος των τιμών των μεταβλητών εισόδων του δικτύου, καθώς το γεγονός αυτό μπορεί να μειώσει τη σημαντικότητα μερικών στην πραγματικότητα πολύτιμων μεταβλητών. Τυπικά, οι μεταβλητές κανονικοποιούνται έτσι ώστε να έχουν μηδενικό μέσο και μοναδιαία *τυπική απόκλιση.*[Bowermann (1992)]

Ιδιαίτερη προσοχή πρέπει να δοθεί στην κανονικοποίηση της μεταβλητής της τιμής εξόδου. Αν η συνάρτηση μεταφοράς του κόμβου εξόδου είναι η σιγμοειδής με ασύμπτωτες στις τιμές μηδέν και ένα, τότε μια μέθοδος που μπορεί να μειώσει τον απαιτούμενο χρόνο εκπαίδευσης είναι η κανονικοποίηση της

μεταβλητής εξόδου στο εύρος [0.1, 0.9] αντί στο [0, 1]. Η μέθοδος αυτή μπορεί να υλοποιηθεί με τη βοήθεια ενός γραμμικού μετασχηματισμού που περιγράφεται από τις παρακάτω εξισώσεις:

$$Y^{scl}(t) = SCALE * Y(t) + OFFSET$$

$$SCALE = \frac{MAX - MIN}{Y_{max} - Y_{min}} \qquad (3.29)$$

$$OFFSET = MAX - \frac{MAX - MIN}{Y_{max} - Y_{min}} Y_{max}$$

όπου $Y^{scl}(t)$ η κανονικοποιημένη τιμή, Y(t) η αρχική τιμή της μεταβλητής, Y_{max} και Y_{min} η μέγιστη και η ελάχιστη τιμή της Y(t) και MAX, MIN τα όρια του επιθυμητού εύρους της εξόδου δηλαδή [0.1, 0.9].

3.5.3 Αντιμετώπιση του καταστροφικού θορύβου

Θόρυβος στα δεδομένα είναι μικρές αποκλίσεις της πραγματικής τιμής. Μπορούμε να επιλέξουμε να υπάρχει θόρυβος στα δεδομένα εκπαίδευσης, ελέγχου, πρόβλεψης ή σε κάποιο συνδυασμό από τα παραπάνω.

Έχει παρατηρηθεί ότι σε ορισμένες περιπτώσεις αριθμητικών δεδομένων, ο θόρυβος οδηγεί σε δίκτυα που μπορούν να γενικεύουν καλύτερα. Επίσης όταν το μέγεθος του δείγματος που έχουμε στη διάθεσή μας είναι μικρό παρατηρήθηκε ότι εκπαιδεύεται καλύτερα αν προσθέσουμε θόρυβο. Όμως η ύπαρξη θορύβου καθυστερεί την εκπαίδευση του δικτύου.

Υπάρχουν διάφορες τεχνικές ελαχιστοποίησης των παρενεργειών του θορύβου, με πιο συνηθισμένες τη χρήση κινητών μέσων και την εκθετική εξομάλυνση.

3.5.4 Αντιμετώπιση ακραίων τιμών (Statistical outliers)

Ως ακραίες τιμές θεωρούμε εκείνες τις παρατηρήσεις στο σύνολο των μεταβλητών εισόδου και εξόδου που απέχουν υπερβολικά από το μέσο του συνόλου. Επειδή προκαλούν συμπίεση της κανονικοποιημένης μεταβλητής σε ένα

πολύ μικρό εύρος, η ύπαρξή τους απαιτεί υπερβολική αριθμητική ακρίβεια κατά τη διάρκεια της εκπαίδευσης. Επιπλέον προκαλούν συμπίεση της πλειοψηφίας των παρατηρήσεων στο γραμμικό τμήμα της σιγμοειδούς.

Ο πιο συνηθισμένος τρόπος αντιμετώπισης των ακραίων τιμών είναι ο οπτικός έλεγχος των παρατηρήσεων, ή ο έλεγχος της κατανομής συχνοτήτων τους. Μια πιο επιστημονική προσέγγιση στην ανίχνευση ακραίων τιμών είναι ο υπολογισμός ενός μέτρου γνωστού ως «Mahalanobis distance» για κάθε πρότυπο εκπαίδευσης, με τη βοήθεια της παρακάτω εξίσωσης:

$$D^2(t) = \sum_{i=1}^{N} \sum_{j=1}^{\rho} \left(Zi(t) - \overline{Zi}\right) \cdot V_{ij} \cdot \left(Zj(t) - \overline{Zj}\right) \qquad (3.30)$$

όπου ρ ο αριθμός των μεταβλητών στο πρότυπο εκπαίδευσης, N ο συνολικός αριθμός των προτύπων εκπαίδευσης, $Z_i(t)$ η i μεταβλητή στο t πρότυπο, Z_i η μέση τιμή της μεταβλητής Z_i και V_{ij} είναι το στοιχείο της i γραμμής και της j στήλης του πίνακα συσχετίσεων των ρ μεταβλητών.

3.5.5 Επιλογή δεδομένων ελέγχου (Test data)

Ο έλεγχος του δικτύου που προκύπτει μετά τη διαδικασία της εκπαίδευσης γίνεται συνήθως με την παρουσίαση στο δίκτυο δεδομένων τα οποία αντιμετωπίζει για πρώτη φορά και που προήλθαν από μια διαδικασία επιλογής πριν την έναρξη της εκπαίδευσης (out-of-sample testing).

Μια μέθοδος επιλογής συνίσταται στη διαμόρφωση του συνόλου ελέγχου με την επιλογή ενός προτύπου ανά δέκα από το σύνολο εκπαίδευσης. Μια άλλη μέθοδος διαχωρίζει τυχαία το δέκα της εκατό του συνόλου των προτύπων ή, απλά διαχωρίζει το δέκα της εκατό από το τέλος του αρχείου εκπαίδευσης (τη μέθοδο αυτή εφαρμόζει το «P.N.A»).

3.6 Μέτρα απόδοσης των Τ.Ν.Δ.

Η χρήση και η σχεδίαση τεχνικών μέτρησης της απόδοσης ενός νευρωνικού δικτύου, εξυπηρετούν δυο στόχους. Ο πρώτος αφορά τη σύγκριση του δικτύου με τις υπόλοιπες στατιστικές τεχνικές πρόβλεψης [Weigend et al

(1991)], ενώ ο δεύτερος και πιο ουσιώδης, τη μέτρηση της αποτελεσματικότητας που επιτυγχάνεται μέσω της εφαρμογής του.

3.6.1 Μέτρα μέτρησης και σύγκρισης της απόδοσης των νευρωνικών δικτύων

Η χρήση και η σχεδίαση τεχνικών μέτρησης της απόδοσης ενός νευρωνικού δικτύου, αφορά τη σύγκριση του δικτύου με τις υπόλοιπες στατιστικές τεχνικές πρόβλεψης. Στην παράγραφο αυτή δίνονται μερικές από τις πιο εφαρμοσμένες τεχνικές που εξυπηρετούν τον προαναφερθέντα στόχο [Συριόπουλος (1997)].

- *Μέτρα σύγκρισης της απόδοσης*

Ως μέτρα σύγκρισης της απόδοσης των νευρωνικών δικτύων μπορούν να χρησιμοποιηθούν τεχνικές συμβατές με τις γενικότερες μεθόδους πρόβλεψης. Οι πιο γνωστές είναι οι εξής:

Α) Ο συντελεστής συσχέτισης.

Μετρά το ποσοστό της γραμμικής συσχέτισης μεταξύ της τιμής εξόδου που προβλέφθηκε και της πραγματικής τιμής εξόδου για το σύνολο των δεδομένων. Ο βαθμός συσχέτισης R δίνεται από την εξίσωση:

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (3.31)$$

όπου $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{και} \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad (3.32)$$

Συνήθως αντί της τιμής του R υπολογίζουμε την τιμή του R^2 , καθώς R^2 =1 σημαίνει τέλεια συσχέτιση μεταξύ των προβλεπομένων και των πραγματικών τιμών, ενώ R^2=0 σημαίνει μηδενική συσχέτιση.

Β) Ο συντελεστής πληροφόρησης

Ο συντελεστής πληροφόρησης T_r (t-test), είναι ένα μέτρο σύγκρισης των αποτελεσμάτων της μεθόδου πρόβλεψης που χρησιμοποιούμε, με την υπόθεση του «τυχαίου περιπάτου» (random-walk hypothesis). Ο συντελεστής T_r υπολογίζεται από την παρακάτω εξίσωση:

$$T_r = \frac{\sqrt{\sum_{t=1}^{\eta} (y_t - x_t)^2}}{\sqrt{\sum_{t=1}^{\eta} (x_t - x_{t-1})^2}} \qquad (3.3.3)$$

όπου y_t η πρόβλεψη του μοντέλου, x_t η πραγματική τιμή και x_{t-1} η πραγματική τιμή της προηγούμενης περιόδου.

Στην περίπτωση που ο συντελεστής T_r είναι μεγαλύτερος ή ίσος με τη μονάδα ($T_r \geq 1$), το μοντέλο μας είναι χειρότερο από το μοντέλο του «τυχαίου περιπάτου».

Στην αντίθετη περίπτωση το μοντέλο πραγματοποιεί καλύτερες προβλέψεις από εκείνες που στηρίζονται στην υπόθεση του «τυχαίου περιπάτου» και η απόδοση του βελτιώνεται όσο ο συντελεστής T_r τείνει στο μηδέν.

Γ) Κριτήρια «Akaike» και «Bayes»

Κατά την ανάπτυξη διαφορετικών μοντέλων με στόχο την επιλογή του βέλτιστου για μια συγκεκριμένη εφαρμογή, συχνά ο σχεδιαστής πρέπει να επιλέξει μεταξύ μοντέλων ίσης απόδοσης όσον αφορά τα παραπάνω κριτήρια. Στην περίπτωση αυτή πρέπει να επιλεγεί το μοντέλο με τις λιγότερες ελεύθερες παραμέτρους.

Το κριτήριο «Akaike» προσαρμόζει το μέσο τετραγωνικό σφάλμα έτσι ώστε να αντικατοπτρίζει την πολυπλοκότητα του μοντέλου με βάση την παρακάτω εξίσωση:

$$A = \frac{1}{\eta}\sum_{i=1}^{\eta}\left(x_i - y_i\right)^2 \cdot \left[\frac{\eta + \kappa}{\eta - \kappa}\right] \tag{3.34}$$

όπου κ οι βαθμοί ελευθερίας του μοντέλου (στην περίπτωση του νευρωνικού δικτύου, ο αριθμός των βαρών).

Το κριτήριο «Bayes» αποτελεί έναν άλλο μετασχηματισμό του μέσου τετραγωνικού σφάλματος:

$$B = \ln\left[\frac{\sum_{i=1}^{\eta}\left(x_i - y_i\right)^2}{\eta}\right] + \frac{\ln[\eta]}{\eta}\cdot\kappa \tag{3.35}$$

όπου κ ο αριθμός των βαρών του δικτύου. Η ποιότητα του μοντέλου όσον αφορά την ικανότητα πρόβλεψης, βελτιώνεται όσο μειώνεται η τιμή του συντελεστή Β.

Δ) Αλλαγή κατεύθυνσης

Σε πολλές περιπτώσεις κατά το σχεδιασμό ενός μοντέλου πρόβλεψης, δεν ενδιαφέρει τόσο ο ακριβής προσδιορισμός του επιπέδου της απόλυτης τιμής της μεταβλητής ή της μεταβολής της, αλλά απλά η πρόβλεψη της τιμής του πρόσημου της μεταβολής, δηλαδή αν η μεταβλητή κινήθηκε ανοδικά ή καθοδικά. Όπως έχει ήδη αναφερθεί αυτό είναι ένα πολύ ευκολότερο καθήκον για το νευρωνικό δίκτυο και μπορεί να το επιτελέσει με μεγαλύτερη ακρίβεια. Για τον προσδιορισμό του ποσοστού επιτυχίας της πρόβλεψης σε μια τέτοια περίπτωση χρησιμοποιείται η ακόλουθη μεταβλητή:

$$d = \frac{1}{\eta} \cdot \sum_{i}^{\eta} \alpha_i$$

$$\alpha_i = \begin{cases} 1 \text{ αν } (x_{t+1} - x_t) \cdot (y_{t+1} - x_t) > 0 \\ 0 \text{ αλλιώς} \end{cases}$$

(3.36)

Η ερμηνεία της μεταβλητής d έχει ως εξής: Για d=1 το εφαρμοζόμενο μοντέλο προβλέπει σε ποσοστό 100% την κατεύθυνση των μεταβολών της τιμής της μεταβλητής (άνοδο ή κάθοδο) ενώ, για d=0 η πρόβλεψη έχει αποτύχει σε ποσοστό 100%. Ένα μοντέλο υπάρχει μεγάλη πιθανότητα να είναι αποδοτικό για τιμές του d μεγαλύτερες από 0,8.

Η χρήση της μεταβλητής d πρέπει να γίνεται με μεγάλη προσοχή, καθώς μπορεί να οδηγήσει σε λανθασμένες εκτιμήσεις σχετικά με την ποιότητα του μοντέλου. Πιο συγκεκριμένα, είναι πολύ εύκολο ακόμη και για ένα μη βέλτιστο νευρωνικό δίκτυο, να δώσει τιμές του d κοντά στη μονάδα όταν εκπαιδευτεί σε πληθυσμό με έντονη κατεύθυνση. Για το λόγο αυτό τα αποτελέσματα πρέπει να επιβεβαιωθούν με διασταύρωσή τους σε τουλάχιστο 30 σύνολα ελέγχου και οι τιμές της μεταβλητής d που προκύπτουν από το καθένα να κανονικοποιηθούν με βάση την τυπική τους απόκλιση.

3.7 Χρήση των νευρωνικών δικτύων σε ιατρικές έρευνες

Τα νευρωνικά δίκτυα έχουν χρησιμοποιηθεί τα τελευταία χρόνια σε πολλές εργασίες που αφορούν ιατρικά θέματα. Πιο συγκεκριμένα, οι *Shen et al (1993)* μελέτησαν τους παράγοντες κινδύνου που σχετίζονται με την εμφάνιση καρδιοαγγειακών προβλημάτων. Επιχειρήθηκε η πρόβλεψη καρδιακών νοσημάτων με μεταβλητές το κάπνισμα, την αρτηριακή πίεση, το επίπεδο χοληστερίνης στο αίμα, την ηλικία και του φύλου. Αντί του δείκτη Gensini, ως μέτρο αξιολόγησης της σοβαρότητας της νόσου χρησιμοποιήθηκε ο δείκτης Dundee Rank Factor Score. Το Νευρωνικό δίκτυο εκπαιδεύτηκε με ένα μικρό δείγμα μεγέθους 32 και επετεύχθη ευαισθησία 67%. Στη συνέχεια, οι Shen,

et al (1994) σύγκριναν τα αποτελέσματα αυτά, με τα αποτελέσματα της στατιστικής μεθόδου Dundee Coronary Risk Disc και απέδειξαν την αποτελεσματικότητα των νευρωνικών δικτύων τα οποία είχαν ακρίβεια πρόβλεψης 89%.

Οι Amendolia et al (1993) ανέπτυξαν ένα νευρωνικό δίκτυο το οποίο με βάση μη επεμβατικές (non invasive) πληροφορίες όπως ιστορικό του ασθενούς και λιπίδια επιτυγχάνουν την ταξινόμηση ασθενών με καρδιοαγγειακά προβλήματα με επίπεδο ακρίβειας 87%.

Οι Anderson et al (1993) ανέπτυξαν ένα νευρωνικό σύστημα για την πρόβλεψη της εξέλιξης της στένωσης στην αρτηρία LAD. Το μοντέλο έχει 27 παραμέτρους και μπορεί να χειρίζεται missing data. Το σύστημα έχει 77 % επιτυχία στην πρόβλεψη της εξέλιξης της στένωσης στη LAD.

Οι Lapuerta et al (1994) εκπαίδευσαν ένα νευρωνικό δίκτυο για την πρόβλεψη του χρόνου εμφάνισης καρδιοαγγειακών προβλημάτων με βάση το "serum lipids profile ". Το δίκτυο συγκρίθηκε με τα αποτελέσματα από το μοντέλο παλινδρόμησης του Cox. Αποδείχθηκε ότι το νευρωνικό δίκτυο προβλέπει καλύτερα τα κλινικά αποτελέσματα από το μοντέλο του Cox (66% ακρίβεια σε σχέση με το 56%).

Οι Jain et al (1995) έκαναν μια συγκριτική μελέτη των expert systems, fuzzy logic και νευρωνικών δικτύων που χρησιμοποιούνται στον εντοπισμό καρδιακών παθήσεων. Χρησιμοποίησαν δείγμα μεγέθους 30 και τρεις εξαρτημένες μεταβλητές σχετικές με την ακουστική μέθοδο πρόβλεψης καρδιακών νοσημάτων. Στη μελέτη αυτή στηρίχθηκε ο ισχυρισμός ότι τα εμπειρικά συστήματα είναι λιγότερο αποτελεσματικά σε σχέση με τις άλλες τεχνικές.

Ο Akay (1995) ανέπτυξε ένα ασαφές (fuzzy) νευρωνικό σύστημα οποίο στηριζόμενο σε «ανώδυνες» παρατηρήσεις αλλά και στο διαστολικό θόρυβο της καρδιάς, ταξινομεί ασθενείς με καρδιοαναπνευστικά προβλήματα. Το σύστημα έχει ακρίβεια 85,5%.

Οι Sztandera et al (1996) εκπαίδευσαν νευρωνικά συστήματα με την βοήθεια αλγορίθμων από την θεωρία ασαφών (fuzzy) συνόλων με σκοπό την πρόβλεψη καρδιοαγγειακών νοσημάτων. Τα αποτελέσματα είναι εντυπωσιακά αφού το 100% του δείγματος ελέγχου ταξινομήθηκε σωστά.

Οι Bologna et al (1997) χρησιμοποιώντας μη επεμβατικές πληροφορίες σύγκριναν διάφορες μεθοδολογίες πρόβλεψης των καρδιοαγγειακών

νοσημάτων, όπως τη γραμμική διαχωριστική ανάλυση, τον C4.5 αλγόριθμο και το MLP μοντέλο. Όπως σημειώνουν, τα αποτελέσματα των νευρωνικών συστημάτων είναι δύσκολο να ερμηνευτούν όταν κάποιες από τις μεταβλητές που εισάγουμε είναι συνεχείς. Το μοντέλο τους το ονόμασαν IMLP (Interpretable Multy Layer Perception) και έχει την ίδια διαχωριστική ικανότητα με το MLP μοντέλο, αλλά και το πλεονέκτημα της εύκολης ερμηνείας των αποτελεσμάτων όταν χρησιμοποιήσουμε συνεχείς μεταβλητές.

Τέλος οι Ohno-Machado και Musen (1997) παρατήρησαν ότι τα sequential νευρωνικά συστήματα είναι αποτελεσματικότερα σε σχέση με τα standard νευρωνικά συστήματα.

3.8 Εφαρμογή των Τ.Ν.Δ. στο πιλοτικό δείγμα και αξιολόγησή της ικανότητας γενίκευσης στο συνολικό δείγμα

Ο αλγόριθμος "back propagation" όπως και κάθε μέθοδος μη γραμμικής αριστοποίησης, μπορεί να επηρεαστεί από το σύνολο τιμών που δίδεται αρχικά στο σύστημα ώστε να ξεκινήσει η διαδικασία αριστοποίησης. Έτσι επαναλαμβάνουμε την διαδικασία εκπαίδευσης του δικτύου με τυχαία βάρη τα οποία επιλέγονται από μια τυπική κανονική κατανομή. Η διαδικασία επαναλαμβάνεται 10 φορές για να μπορέσουμε να ελαττώσουμε τον κίνδυνο επηρεασμού των αποτελεσμάτων της αριστοποίησης από τις αρχικές τιμές. Το ζητούμενο είναι να βρεθεί κάποιο ελάχιστο, ιδεατά το ολικό, της τετραγωνικής ρίζας του μέσου σφάλματος τετραγώνου (RMSE) σε χώρο $n \times m \times r$ παραμέτρων.
Είναι δυνατό, αλλά υπολογιστικά ασύμφορο, να βρεθεί το ολικό ελάχιστο RMSE με τη μέθοδο των μη γραμμικά ελαχίστων τετραγώνων.

Η εφαρμογή των νευρωνικών δικτύων στο πιλοτικό δείγμα των 160 ατόμων έγινε σύμφωνα με τις γενικά αποδεκτές αρχές σχεδίασης και υλοποίησης ενός τυπικού εμπρόσθιου νευρωνικού δικτύου (Standard Feedforward Neural Network). Χρησιμοποιήθηκε το πρόγραμμα PROFILE Neural Applications (P.N.A)

- **Σχεδίαση των Δικτύων**

Με σκοπό να επιτευχθεί ο καλύτερος διαχωρισμός του πιλοτικού δείγματος σχεδιάσθηκαν και επεξεργάστηκαν 10 νευρωνικά δίκτυα τα οποία διαφοροποιούνται ως προς τον αριθμό των παραμέτρων (weights), τον αριθμό των κύκλων εκπαίδευσης (runs), των συναρτήσεων μεταφοράς (transfer functions), και του αριθμού των κρυφών επιπέδων (hidden layers).

Χρησιμοποιήθηκαν ως ανεξάρτητες μεταβλητές (inputs) οι παράγοντες κινδύνου και το κλάσμα εξώθησης, όπως και στις προαναφερθείσες διαχωριστικές τεχνικές. Αναλυτικά χρησιμοποιήθηκαν όλοι οι παράγοντες κινδύνου (11 μεταβλητές): φύλο (SEX), ηλικία (AGE), λιπίδια (LIPIDS), αρτηριακή πίεση (HBP), κάπνισμα (SMOKE), διαβήτης (DIABETES), οικογενειακό ιστορικό (FAMILY), παχυσαρκία (OBESITY), καθιστική ζωή (SEDENTARY), τύπος Α (TYPE A) και κλάσμα εξώθησης (EF), ή εναλλακτικά 7 μεταβλητές: λιπίδια, αρτηριακή πίεση, κάπνισμα, οικογενειακό ιστορικό, καθιστική ζωή, τύπος Α, κλάσμα εξώθησης.

Η εξαρτημένη μεταβλητή είναι όπως και προηγουμένως η DCODE. Χρησιμοποιήθηκαν τα αποτελέσματα των συσχετίσεων και στοιχεία στατιστικής ανάλυσης τα οποία αναφέρονται στα προηγούμενα κεφάλαια. Αξίζει να σημειωθεί ότι ο έλεγχος του δικτύου γινόταν μετά από κάθε 10 επαναλήψεις εκπαίδευσης.

- **Εκπαίδευση των δικτύων**

Όλες οι παραπάνω μεταβλητές εισήχθησαν στο υποσύστημα του P.N.A , το Brain Maker, που είναι ο πυρήνας του Νευρωνικού δικτύου και αναλαμβάνει την εκπαίδευση καθώς και την εκτέλεση του δικτύου. Τα δίκτυα εκπαιδεύτηκαν με τα 160 δεδομένα του πιλοτικού δείγματος. Στη συνέχεια παρατίθεται ο πίνακας περιγραφής των 10 τεχνητών νευρωνικών δικτύων που σχεδιάστηκαν και επεξεργάστηκαν.

ΠΙΝΑΚΑΣ 3.2: ΠΕΡΙΓΡΑΦΗ ΤΝΔ

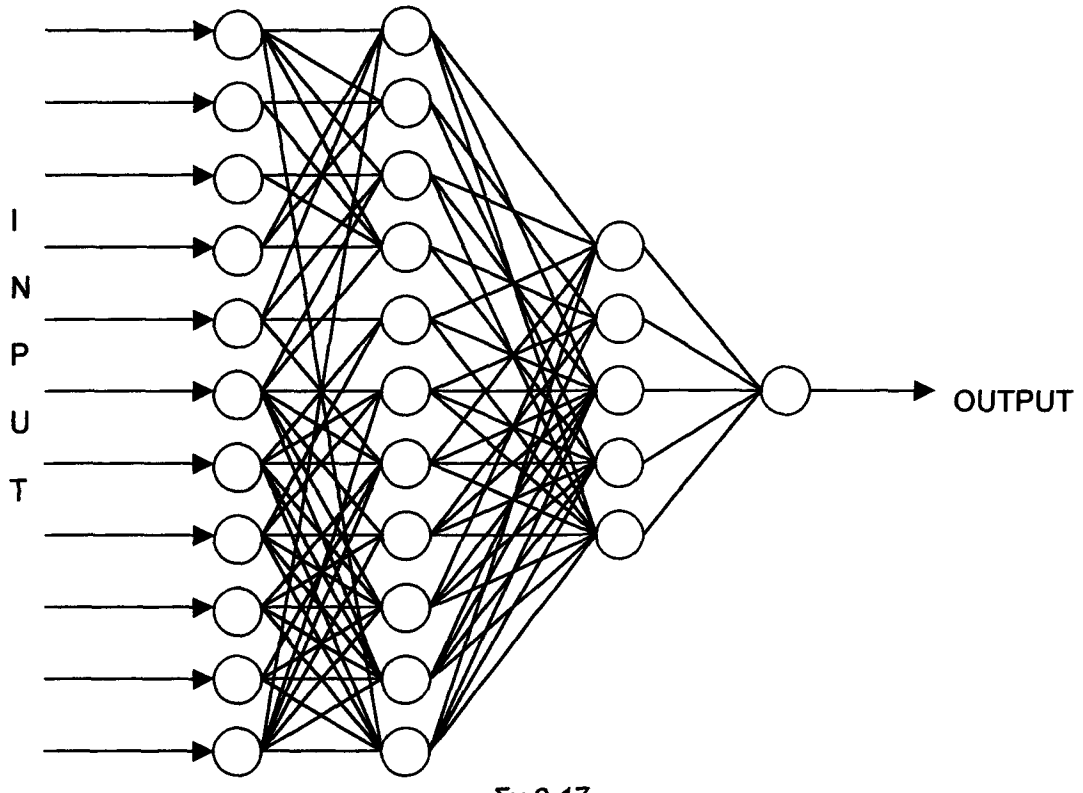
Α/Α ΤΝΔ	Υπόδειγμα	Συναρτήσεις μεταφοράς	Παράμετροι.	Κύκλοι (Runs)
1	11x11x1	Σιγμοειδής	132	200
2	11x11x1	Σιγμοειδής	132	500
3	11x11x1	Σιγμοειδής	132	1,000
4	11x5x1	Σιγμοειδής	60	200
5	11x11x 5x1	Σιγμοειδής	192	200
6	11x11x1	Βηματική	132	350
7	11x11x1	Κατωφλίου	132	340
8	11x5x1	Κατωφλίου	60	1000
9	7x7x1	Σιγμοειδής	56	215
10	7x7x1	Κατωφλίου	56	160

Ο αριθμός των νευρώνων και των κρυφών επιπέδων που χρησιμοποιήθηκε σε κάθε ΤΝΔ παρουσιάζεται στη στήλη "Υπόδειγμα" (n × m × r) όπου n = το πλήθος των κόμβων εισόδου, m = το πλήθος των κόμβων του κρυφού επιπέδου, r = το πλήθος των κόμβων του επιπέδου εκροής (εξόδου). Ο αριθμός των συνδέσμων μεταξύ των νευρώνων που χρησιμοποιήθηκε για να μεταφερθεί και επεξεργαστεί η πληροφορία από το επίπεδο εισόδου (input layer) στο κρυφό ή στα κρυφά επίπεδα (hidden layer) παρουσιάζεται στη στήλη "Παράμετροι".

Η επεξεργασία των δεδομένων ανάμεσα στο επίπεδο εισόδου και στο κρυφό επίπεδο έγινε με τις συναρτήσεις που αναφέρονται στη στήλη "Συναρτήσεις μεταφοράς" (στη βιβλιογραφία συναντάται και ως «συναρτήσεις ενεργοποίησης» [Limin Fu (1994)]).

Από το κρυφό επίπεδο μέχρι το επίπεδο εξόδου (output layer), η μεταφορά του αποτελέσματος της επεξεργασίας και η ταξινόμηση των αποτελεσμάτων έγινε μέσω ενός κόμβου με τη βοήθεια των συναρτήσεων μεταφοράς που επιλέχθηκαν να συμπίπτουν με τις συναρτήσεις που χρησιμοποιήθηκαν προηγουμένως. Οι συναρτήσεις μεταφοράς μπορεί να είναι στοχαστικές συναρτήσεις, συνήθως όμως είναι προσδιοριστικές.

Το σχήμα 3.17 παριστά την αρχιτεκτονική του δικτύου 11×11×5×1.



Σχ.3.17

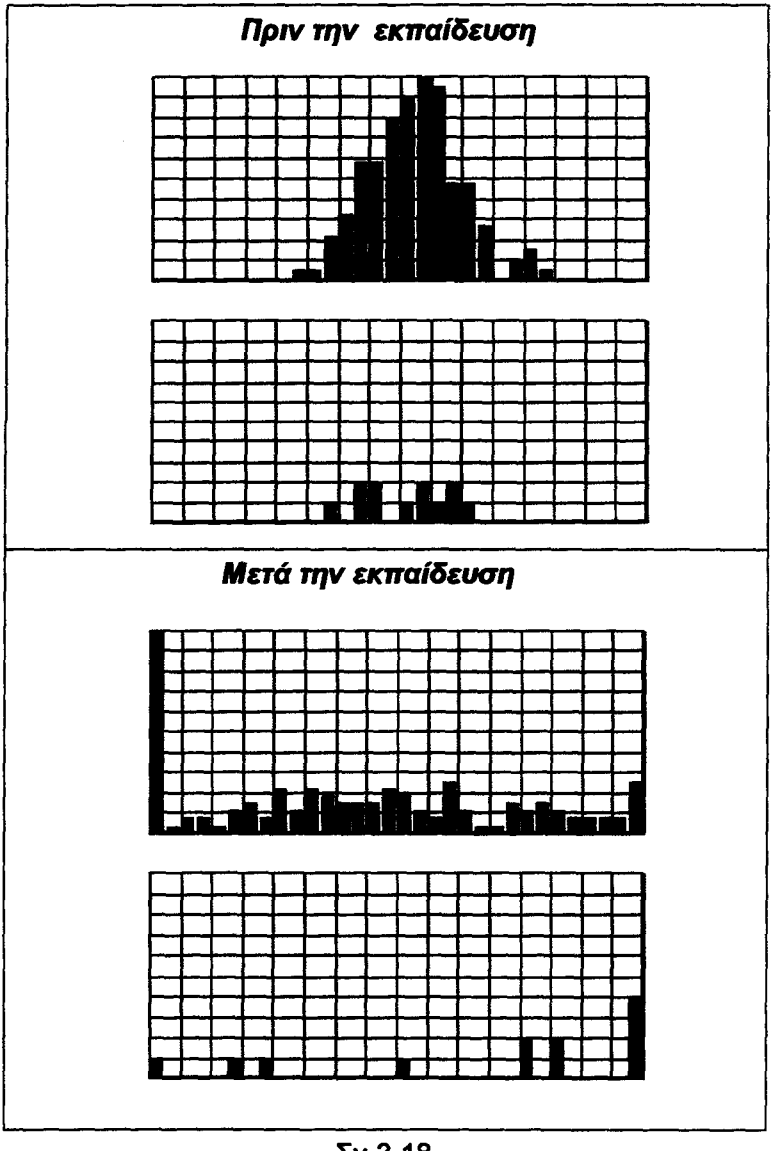
Το δίκτυο 11×11×5×1 αποτελείται από ένα επίπεδο εισόδου από κόμβους που λαμβάνουν εισροές από εξωτερικές πηγές, δύο κρυφά επίπεδα και το επίπεδο εξόδου που κατευθύνει τα σήματα εξόδου.

Αν θεωρήσουμε ένα δίκτυο με ένα κρυφό επίπεδο $m + 1=11$ κόμβων, $n + 1=12$ πλήθος κόμβων εισόδου (μαζί με το σταθερό όρο) και 1 κόμβο στο επίπεδο εκροής τότε ο αριθμός των προς εκτίμηση παραμέτρων είναι ίσος με $(n+1) \cdot (m+1)= 11 \cdot 12=132$. Επειδή το δίκτυο του σχήματος 3.17 έχει δύο κρυφά επίπεδα τότε ο αριθμός των παραμέτρων αυξάνεται σε:

$11 \cdot 12 + 5 \cdot 12=132 + 60 = 192$. [Συριόπουλος (1998)].

Η εξέλιξη της εκπαίδευσης του δικτύου καθίσταται εφικτή με τη χρήση ιστογραμμάτων των βαρών. Στη συνέχεια παραθέτουμε τα ιστογράμματα κατά την εξέλιξη της εκπαίδευσης του δικτύου 11×11×1.

Κατανομή βαρών ΤΝΔ 11x11x1 πριν και μετά την εκπαίδευση



Σχ.3.18

Παρατηρείται ότι εμφανίζονται δύο ιστογράμματα στο στάδιο πριν και στο στάδιο μετά την εκπαίδευση. Το πρώτο από τα δύο ιστογράμματα και στα δύο στάδια απεικονίζει την κατανομή των βαρών των συνδέσμων σε εύρος

[-8,8] μεταξύ του επιπέδου εισόδου και των κρυφών επιπέδων και το δεύτερο απεικονίζει την κατανομή των βαρών μεταξύ των κρυφών επιπέδων και του επιπέδου εξόδου. Τα ιστογράμματα αυτά είναι σημαντικά για την αξιολόγηση της κατάστασης του δικτύου κατά την εξέλιξη της εκπαίδευσης και η ερμηνεία τους είναι η ακόλουθη:

α) Αν τα βάρη είναι συγκεντρωμένα στο κέντρο του ιστογράμματος (κατανέμονται κανονικά γύρω από το μηδέν) το δίκτυο έχει πολλές δυνατότητες βελτίωσης στο υπόλοιπο της εκπαίδευσης.

β) Αν ένας μεγάλος αριθμός συγκεντρώνεται στα δύο άκρα (τιμές περίπου ίσες με +8 ή –8) και το δίκτυο έχει πολλά δεδομένα εκπαίδευσης να μάθει, η συνέχιση της εκπαίδευσης είναι πιθανότατα μάταιη.

Στο δίκτυο 11×11×1 (Σχ. 3.18) στο αρχικό στάδιο της εκπαίδευσης παρατηρείται ότι τα βάρη βρίσκονται συγκεντρωμένα στο κέντρο του διαγράμματος σε σχήμα κανονικής κατανομής απ' όπου συμπεραίνουμε ότι το δίκτυο έχει ακόμα περιθώρια να εκπαιδευθεί. Όσο η εκπαίδευση εξελίσσεται, τα βάρη κατανέμονται ομοιόμορφα σε όλο το εύρος. Στο δεύτερο διάγραμμα μετά την εκπαίδευση παρατηρούμε ότι τα βάρη έχουν μετακινηθεί προς τα άκρα οπότε οι πιθανότητες για αποτελεσματική περαιτέρω εκπαίδευση μειώνονται άρα το δίκτυο έχει πιθανότατα ολοκληρώσει την εκπαίδευσή του.

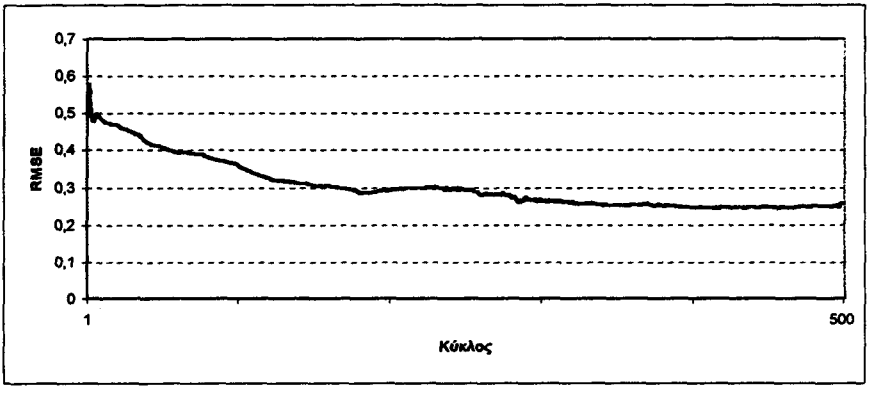
Ένα άλλο γράφημα που δίνει σημαντικές πληροφορίες για την εξέλιξη της *εκπαίδευσης του δικτύου και αποτελεί μέτρο αποτελεσματικότητας του δικτύ*ου, είναι αυτό που παριστάνει τη μεταβολή του μέτρου της τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος (RMSE) $RMSE = \sqrt{\frac{\sum_{j,i}^{n}(y_{ic} - d_{j,c})^2}{n}}$

Όπου $y_{j,c}$ είναι η τιμή εξόδου του κόμβου j για το πρότυπο εκπαίδευσης c και $d_{j,c}$ η επιθυμητή τιμή.

Θεωρούμε ότι η εκπαίδευση προχωρά ομαλά όταν από κύκλο σε κύκλο οι τιμές του σφάλματος συγκεντρώνονται στα άκρα του ιστογράμματος και η τιμή του RMSE μειώνεται τείνοντας στο μηδέν.

Στη συνέχεια παρατίθεται το γράφημα RMSE για το δίκτυο11×11×1.

Εξέλιξη RMSE ανά κύκλο εκπαίδευσης για ΤΝΔ 11x11x1



Σχ. 3.19

Παρατηρείται ότι σε 500 κύκλους εκπαίδευσης η τιμή του RMSE προσεγγίζει το 0.26.

Κατά την εκπαίδευση του δικτύου η επιτρεπτή ανοχή λάθους (Training tolerance) στα δεδομένα εκπαίδευσης καθορίστηκε στο 0.1. Η τιμή ανοχής καθορίζει πόσο ακριβής πρέπει να είναι η πρόβλεψη του δικτύου στα δεδομένα εκπαίδευσης ως ποσοστό του εύρους της επιθυμητής εξόδου για να θεωρηθεί σωστή. Στη συγκεκριμένη περίπτωση η τιμή ανοχής 0.1 δηλαδή 10% σημαίνει ότι κάθε πρόβλεψη στο εύρος [90 ,100] θα θεωρηθεί σωστή. Αν μια πρόβλεψη κριθεί σωστή ο αλγόριθμος δεν προχωρεί στη μεταβολή των βαρών του δικτύου.

Στα ήδη εκπαιδευμένα δίκτυα εφαρμόστηκαν επίσης τεχνικές κλαδέματος με σκοπό να οδηγηθούν σε καλύτερη γενίκευση. Για κάθε επίπεδο ορίστηκε μια τιμή στο εύρος [0, 8]. Όλοι οι σύνδεσμοι ενός επιπέδου με τιμές μικρότερες από αυτή που επελέγη τέθηκαν ίσοι με το μηδέν. Μετά το κλάδεμα του δικτύου η εκπαίδευση συνεχίστηκε μέχρι αυτό να φθάσει σε σύγκλιση.

- **Απόδοση των δικτύων**

Όπως αναφέρθηκε παραπάνω, κατά τη διάρκεια της εκπαίδευσης παρακολουθήθηκε η εξέλιξη της τιμής του RMSE, τα ιστογράμματα, τα διαγράμματα, καθώς και το ποσοστό σωστής ταξινόμησης των δεδομένων εκπαίδευσης. Η εκπαίδευση σταμάτησε για εκείνο τον αριθμό των κύκλων όπου η τιμή του

RMSE ήταν η μικρότερη που παρατηρήθηκε και το επίπεδο ανοχής 0,1 [Συριόπουλος (1996)]

Ο πίνακας που ακολουθεί παρουσιάζει το ποσοστό σωστής ταξινόμησης των δεδομένων εκπαίδευσης (πιλοτικό δείγμα) καθώς και την τιμή του RMSE ως προς τον αριθμό των κύκλων κατά τον οποίο το δίκτυο κρίθηκε εκπαιδευμένο.

ΠΙΝΑΚΑΣ 3.3: ΣΥΓΚΡΙΣΗ ΤΝΔ

Α/Α ΤΝΔ	Υπόδειγμα	Συνάρτηση μεταφοράς	Παράμετροι.	Κύκλοι	Αποτελεσματικότητα %	RMSE
1	**11x11x1**	**Σιγμοειδής**	**132**	**200**	**91.20**	**0.2619**
2	11x11x1	Σιγμοειδής	132	500	94.37	0.2574
3	11x11x1	Σιγμοειδής	132	1,000	94.37	0.2463
4	11x5x1	Σιγμοειδής	60	200	85.62	0.3441
5	11x11x 5x1	Σιγμοειδής	192	200	90.62	0.2883
6	11x11x1	Βηματική	132	350	88.12	0.3446
7	**11x11x1**	**Κατωφλίου**	**132**	**340**	**90.62**	**0.2866**
8	11x5x1	Κατωφλίου	60	1000	82.50	0.3651
9	**7x7x1**	**Σιγμοειδής**	**56**	**215**	**88.80**	**0.2996**
10	7x7x1	Κατωφλίου	56	160	87.50	0.3565

Καταλληλότερα κρίνονται εκείνα τα ΤΝΔ που συνδυάζουν μικρό αριθμό κύκλων, μικρό αριθμό παραμέτρων, μικρή τιμή RMSE, και μεγάλη αποτελεσματικότητα. Η σπουδαιότερη όμως ιδιότητα των ΤΝΔ είναι η ικανότητα γενίκευσης σε άγνωστα δεδομένα. Στη συνέχεια ελέγχθηκε η διαχωριστική ικανότητα των 10 τεχνητών νευρωνικών δικτύων (ΤΝΔ) στο συνολικό δείγμα (660 άτομα), με σκοπό να συγκριθεί η αποτελεσματικότητα των ΤΝΔ μεταξύ τους, καθώς και με την αντίστοιχη των διαχωριστικών τεχνικών. Η σύγκριση των ΤΝΔ με τις διαχωριστικές τεχνικές παρουσιάζεται στα συμπεράσματα της έρευνας (Κεφ.5).

Παρατίθεται ο πίνακας αποτελεσματικότητας(%) των ΤΝΔ στο συνολικό δείγμα

ΠΙΝΑΚΑΣ 3.4

Α/Α ΤΝΔ	1	2	3	4	5	6	7	8	9	10
Αποτελεσματικότητα %	**86,10**	83,79	83,79	85,00	80,61	79,70	**85,00**	84,09	**86,97**	83.33

Παρατηρείται ότι τα δίκτυα που εμφανίζουν τη μεγαλύτερη αποτελεσματικότητα στο συνολικό δείγμα των 660 ατόμων είναι το πρώτο ,το έβδομο και το ένατο, δηλαδή τα υποδείγματα 11x11x1 ,11x11x1 και 7x7x1. Στη συνέχεια παρατίθεται ο πίνακας αξιολόγησης των συγκεκριμένων ΤΝΔ.

ΠΙΝΑΚΑΣ 3.5: Αξιολόγηση ΤΝΔ

	Αξιολόγηση	**ΤΝΔ 1° 11x11x1 Σιγμοειδής**	**ΤΝΔ 9° 7x7x1 Σιγμοειδής**	**ΤΝΔ 7° 11x11x1 Κατωφλίου**
Στοιχεία του ΤΝΔ	Παράμετροι	132	56	132
	Κύκλοι	200	215	340
Απόδοση των ΤΝΔ στο δείγμα εκπαίδευσης	Ειδικότητα (specificity)	**92,5%**	88,8%	90,0%
	Ευαισθησία (sensitivity)	90,0%	87,5%	**91,3%**
	Αποτελεσματικότητα n=160 (Efficiency)	**91.20%**	88,13%	90,63%
	Δείκτης Youden	**82.50%**	76,25%	81,25%
Ικανότητα γενίκευσης των ΤΝΔ	Αποτελεσματικότητα n=660	86,10%	**86,97%**	85,00%
Κριτήρια εκτίμησης τάξης υποδειγμάτων	Τετραγωνική ρίζα μέσου Σφάλματος τετραγώνου RMSE	**0.2620**	0.2997	0.2867
	Μέσο απόλυτο σφάλμα MAE	0.1230	0.1601	**0.1104**
	R^2	**0.7314**	0.6434	0.6852

Τα υποδείγματα 1 και 9 χρησιμοποιούν τη σιγμοειδή συνάρτηση ενώ το υπόδειγμα 7 χρησιμοποιεί τη συνάρτηση κατωφλίου. Επίσης τα υποδείγματα αυτά, εμφανίζουν την καλύτερη συμπεριφορά και στο δείγμα εκπαίδευσης (πιλοτικό δείγμα) σύμφωνα με τα κριτήρια που προαναφέρθηκαν. Το πρώτο ΤΝΔ εμφανίζει τη μεγαλύτερη αποτελεσματικότητα, τη μεγαλύτερη τιμή δείκτη Youden, την καλύτερη τιμή RMSE και R^2 και πολύ καλή ικανότητα γενίκευσης. Για τα υποδείγματα επτά και εννέα, σύμφωνα με τον πίνακα αξιολόγησης, παρατηρούμε ότι έχουν μεγάλη ικανότητα σωστής ταξινόμησης στο συνολικό δείγμα. Το υπόδειγμα τέσσερα επίσης έχει αποτελεσματικότητα 85% στο δείγμα των 660 ατόμων, αλλά η τιμή RMSE στο πιλοτικό δείγμα είναι μεγάλη. (Κόκλα et al, (2000)). Σύμφωνα με τα παραπάνω ισχυριζόμαστε ότι τα δίκτυα ένα, επτά και εννέα έχουν τη μεγαλύτερη δυνατότητα γενίκευσης και εμφανίζουν την καλύτερη εκτίμηση τάξης υποδείγματος. Αν θα έπρεπε να προταθεί ένα από τα τρία ΤΝΔ αυτό θα ήταν το πρώτο.

ΚΕΦΑΛΑΙΟ 4

ΕΝΑ ΜΑΘΗΜΑΤΙΚΟ ΜΟΝΤΕΛΟ ΓΙΑ ΤΟ ΔΕΙΚΤΗ GENSINI

4.1 Ο Δείκτης Gensini

Στην προσπάθεια να αξιολογηθεί και να ποσοτικοποιηθεί η βαρύτητα της στεφανιαίας νόσου, μια σειρά δείκτες έχουν κατά καιρούς προταθεί, [Seltzer A. (1982)]. Ένας από τους σπουδαιότερους δείκτες είναι αυτός του Gensini [Gensini (1967), (1975), (1980), (1983)]. Για τον υπολογισμό του λαμβάνονται υπόψη :

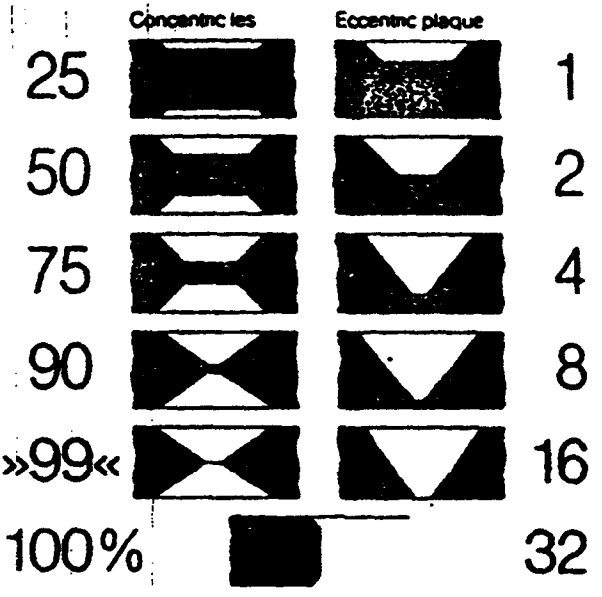
1) η γεωμετρικά αυξανόμενη κρισιμότητα της νόσου σε αντιστοιχία με το ποσοστό στένωσης της αρτηρίας,
2) η αρτηρία όπου εμφανίζεται η στένωση,
3) η θέση της στένωσης ,
4) η επίδραση των παράπλευρων αρτηριών,
5) το μέγεθος και η ποιότητα των περιφερικών αγγείων,
6) η κατάσταση της λειτουργίας του μυοκαρδίου και
7) τα αθροιστικά αποτελέσματα πολλαπλών στενώσεων.

Σύμφωνα με τον Gensini (1983) ο δείκτης αυτός έχει τα εξής πλεονεκτήματα:

1) Δίνει μια στρωματοποίηση των ασθενών σύμφωνα με τη σοβαρότητα της ασθένειάς τους. Έτσι μπορούν να ταξινομηθούν ασθενείς με ανάλογο βαθμό σοβαρότητας της ασθένειας.
2) Επιτρέπει τη χρήση λογισμικού στην αποθήκευση, επεξεργασία και ανάλυση των δεδομένων που προέρχονται από την εξέταση των ασθενών.

Ο υπολογισμός του βασίζεται σε δύο παράγοντες όπως προκύπτει από τα σχήματα 4.1 και 4.2 [Gensini (1983)].

α) στο βαθμό κρισιμότητας (Severity Score) ο οποίος αντιστοιχεί στο ποσοστό στένωσης της αρτηρίας, όπως προκύπτει από το Σχ.4.1



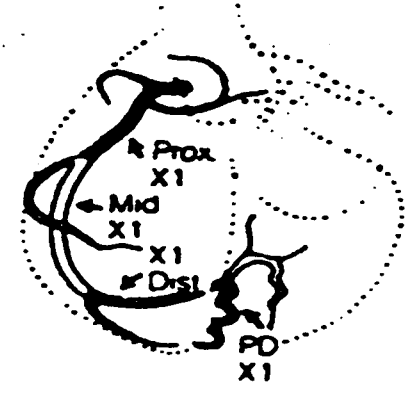
Σχ. 4.1

και παρουσιάζεται στον πίνακα 4.1

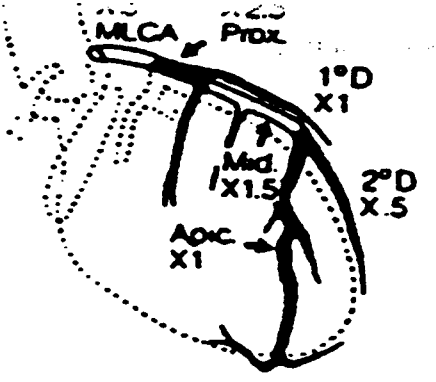
ΠΙΝΑΚΑΣ 4.1

Ποσοστό στένωσης (%)	Βαθμός Κρισιμότητας
25	1
50	2
75	4
90	8
99	16
100	32

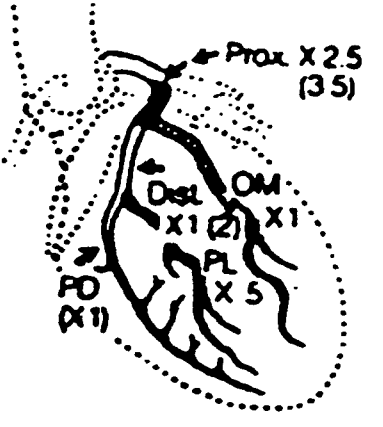
β) στο **συντελεστή** w_{ij} $i=1,...,10$, $j=1,2,3$, που η τιμή του εξαρτάται από την θέση i της στένωσης, καθώς και την αρτηρία j. Οι συντελεστές αυτοί ορίζονται στο σχήμα 4.2 [Gensini (1983)], και συγκεντρώνονται στον πίνακα 4.2 που ακολουθεί.



(α) RCA



(β)LAD



(γ) LCX

Σχ. 4.2

Στον πίνακα 4.2 καταγράφονται οι συντελεστές (ή βάρη) του δείκτη Gensini [Gensini (1983)]. Σε κάθε θέση i της αρτηρίας j αντιστοιχεί ένας συντελεστής w_{ij}, η τιμή του οποίου προσδιορίζεται ανάλογα με την βαρύτητα που προσδίδει στη νόσο η εμφάνιση στένωσης στη συγκεκριμένη θέση i της αρτηρίας j.

ΠΙΝΑΚΑΣ 4.2: Οι συντελεστές του δείκτη Gensini (w_{ij})

Αρτηρία Θέση στένωσης	j=1(RCA)	j=2(LAD)	j=3(LCX)
i=1(PROX)	$w_{11}=1$	$w_{12}=2,5$	$w_{13}=2,5$
i=2(MID)	$w_{21}=1$	$w_{22}=1,5$	
i=3(DIST)	$w_{31}=1$		$w_{33}=1$
i=4(PD)	$w_{41}=1$		$w_{43}=1$
i=5(MLCA)		$w_{52}=5$	
i=6(1oD)		$w_{62}=1$	
i=7(2oD)		$w_{72}=0,5$	
i=8(APIC)		$w_{82}=1$	
i=9(OM)			$w_{93}=1$
i=10(PL)			$w_{103}=0,5$

Ο δείκτης Gensini για κάθε στένωση ορίζεται ως το γινόμενο του βαθμού κρισιμότητας επί τον αντίστοιχο συντελεστή w_{ij}. Σε περίπτωση πολλαπλών στενώσεων, το άθροισμα των επί μέρους δεικτών, οδηγεί στον τελικό δείκτη.

Ο υπολογισμός του δείκτη, εμφανίζει μερικά μειονεκτήματα:

α) Για τις τιμές έμφραξης τις ενδιάμεσες αυτών που αναφέρονται στον Πίνακα 4.1, η αντίστοιχη τιμή κρισιμότητας υπολογίζεται εμπειρικά δηλ. κατ' εκτίμηση των ιατρών.

β) Η τιμή του δείκτη που προκύπτει από πολλές μη σημαντικές στενώσεις μπορεί να είναι συγκρίσιμη με την τιμή που αντιστοιχεί σε μια μόνο σημαντική στένωση. Με άλλα λόγια, μια τιμή του δείκτη Gensini μπορεί να αντιστοιχεί είτε σε πολλές μικρές στενώσεις είτε σε μια μεγάλη.

Την τελευταία δεκαετία έγιναν μια σειρά έρευνες, οι οποίες είχαν ως αντικείμενο τη συσχέτιση του δείκτη Gensini με παράγοντες που επηρεάζουν την καρδιοαγγειακή νόσο, αλλά και τη μελέτη παραγόντων κινδύνου σε ασθενείς των οποίων η βαρύτητα της νόσου είχε ταξινομηθεί με βάση τον δείκτη Gensini. Ενδεικτικά αναφέρουμε τις ακόλουθες:

Οι Tzung-Dau Wang et al (1998) μελέτησαν τη σχέση του επιπέδου χοληστερίνης και του δείκτη Gensini. Το δείγμα χωρίστηκε σε τρεις ομάδες ανάλογα με επίπεδο της χοληστερίνης. Την πρώτη ομάδα αποτελούσαν άτομα με επίπεδο χοληστερίνης μικρότερο του 200mg/dl, τη δεύτερη μεταξύ 200 και 240 mg/dl και την τρίτη μεγαλύτερο του 240mg/dl. Η τρίτη ομάδα είχε υψηλότερο δείκτη Gensini σε σχέση με τις άλλες δύο ομάδες αν και η διαφορά αυτή δεν ήταν στατιστικώς σημαντική.

Οι Krämec B. et al (1986) μελέτησαν τη σχέση του δείκτη Gensini με το κλάσμα εξώθησης EF, τα QT-διαστήματα, καθώς επίσης και τον αριθμό φραγμένων αρτηριών.

Οι Chia-Lun Chao et al (1996), βρήκαν ότι υπάρχει γραμμική σχέση μεταξύ του δείκτη Gensini και του TI-defect severity score (r=0,7343).

Οι Krishnaswami et al (1996), έδειξαν ότι οι διαβητικοί έχουν υψηλότερο μέσο δείκτη Gensini σε σχέση με τους μη διαβητικούς.

Οι Shunji Kasaoka et al (1997), απέδειξαν ότι ο δείκτης Gensini είναι στατιστικώς υψηλότερος στην ομάδα ασθενών με επίπεδο χοληστερίνης μεγαλύτερο του 240mg/dl σε σχέση με την ομάδα ασθενών με χοληστερίνη μικρότερη των 200mg/dl.

Οι Garfagnini et al (1995), απέδειξαν ότι για τις δύο ομάδες οι οποίες διαχωρίζονται με βάση το δείκτη Gensini (≤ 18, >18) υπάρχει στατιστικώς σημαντική διαφορά των δύο μέσων τιμών των επιπέδων της απολιποπρωτεΐνης A1, καθώς και του λόγου απολιποπρωτεΐνη A1/απολιποπρωτεΐνη B.

Οι Kyriakidis et al (1994), έδειξαν ότι ο δείκτης Gensini είναι στατιστικώς υψηλότερος στους άνδρες απ' ότι στις γυναίκες, γεγονός που δείχνει σοβαρότερη και πιο εκτεταμένη καρδιαγγειακή νόσο στους άνδρες.

Οι Takayanagi et al (1991), χώρισαν το δείγμα σε δύο ομάδες και παρατήρησαν ότι η ομάδα των ασθενών που εμφάνισαν ST-segment depression με την είσοδό τους στο νοσοκομείο (≥1mm in V2-V6) είχαν στατιστικώς μεγαλύτερο δείκτη σε σχέση με την ομάδα η οποία δεν εμφάνισε ST-depression(<1mm).

Στο κεφάλαιο αυτό προτείνεται ένας νέος δείκτης (DSN) ο οποίος αποδίδει ένα βαθμό κρισιμότητας για κάθε ποσοστό στένωσης βάσει μιας μαθηματικής συνάρτησης και όχι κατ' εκτίμηση των ιατρών.

4.2 Ο εναλλακτικός δείκτης DSN

Για τον υπολογισμό του δείκτη DSN απαιτείται ο προσδιορισμός του δείκτη κρισιμότητας ds (i,y,j) για κάθε αρτηρία (j=1(RCA) , j=2(LAD), j=3(LCX)). Στο μοντέλο που προτείνεται, ο δείκτης κρισιμότητας ds (i,y,j) υπολογίζεται από μια μαθηματική συνάρτηση για τις τιμές των i,y,j που παρατίθενται στους πίνακες 4.1, 4.2. Έγινε προσπάθεια με τη συνάρτηση αυτή να αποδίδονται με μεγάλη ακρίβεια οι προβλεπόμενες τιμές του δείκτη Gensini.

Ο προσδιορισμός του δείκτη κρισιμότητας ds(i,y,j) επιχειρείται σε δύο στάδια: α) Ο προσδιορισμός μιας συνάρτησης f(i,,j), μέσω της οποίας σε κάθε θέση i της αρτηρίας j να αποδίδεται ο αντίστοιχος συντελεστής (w_{ij}) όπως ορίζεται από τον Gensini

β) Η εύρεση μιας συνεχούς συνάρτησης g(y), με μεταβλητή το κλάσμα έμφραξης y όπου $0,25 \leq y \leq 1,00$

Στο πρώτο στάδιο θα χρησιμοποιηθεί πολυώνυμο παρεμβολής Lagrange διερχόμενο από τα σημεία (i, w_{ij}) για j=1,2,3 όπως ορίζονται στον πίνακα 4.2.

Γενικότερα το πρόβλημα της παρεμβολής συνίσταται στην εύρεση ενός πολυωνύμου το οποίο να διέρχεται από δοθέντα ζεύγη τιμών $(x_i, y_i), i = 0,\ldots,n$.

Δηλαδή αν $P(x)$ είναι το πολυώνυμο παρεμβολής τότε $P(x_i) = y_i$ για i=0,1,...n. Συγκεκριμένα ο τύπος του πολυωνύμου είναι:

$$P(x) = \sum_{i=0}^{n} L_i(x) y_i \qquad (4.1)$$

Τα πολυώνυμα $L_i(x)$ για i=0,1,...n ονομάζονται συντελεστές Lagrange, είναι πολυώνυμα n βαθμού και ορίζονται ως εξής:

$$L_i(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})\cdot(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} \qquad (4.2)$$

Σε συνοπτική μορφή

$$L_i(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x-x_i}{x_j-x_i} \qquad (4.3)$$

Όπως είναι προφανές

$$L_i(x_j) = \begin{cases} 1 \text{ για } i = j \\ 0 \text{ για } i \neq j \end{cases} \quad (4.4)$$

Πράγματι για i=j ισχύει ότι:

$$L_i(x) = \frac{(x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})\cdot(x_i - x_{i+1})\ldots(x_i - x_n)}{(x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})\cdot(x_i - x_{i+1})\ldots(x_i - x_n)} = 1 \quad (4.5)$$

και ομοίως για i≠ j με j=k ισχύει ότι:

$$L_j(x_\kappa) = \frac{(x_\kappa - x_0)(x_\kappa - x_1)\ldots(x_0 - x_\kappa)\ldots(x_\kappa - x_n)}{(x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})\cdot(x_i - x_{i+1})\ldots(x_i - x_n)} = 0 \quad (4.6)$$

[Αλεξανδρόπουλος et al (1995)]

Για την εύρεση των πολυωνύμων παρεμβολής Lagrange, ο πίνακας 4.2 μετασχηματίζεται ως εξής:

ΠΙΝΑΚΑΣ 4.3

Αρτηρία(j)	J=1(RCA)	J=2(LAD)	J=3(LCX)
	(i, $w_{i,j}$) = (i, $P_j(i)$)		
Θέση έμφραξης(i)	(i, $P_1(i)$)	(i, $P_2(i)$)	(i, $P_3(i)$)
i=1(PROX)	(1 . 1)	(1. 2.5)	(1 .2.5)
i=2(MID)	(2 . 1)	(2. 1.5)	
i=3(DIST)	(3 . 1)		(3 . 1)
i=4(PD)	(4 . 1)		(4. 1)
i=5(MLCA)		(5 . 5)	
i=6(1oD)		(6 . 1)	
i=7(2oD)		(7 . 0.5)	
i=8(APIC)		(8 . 1)	
i=9(OM)			(9 .1)
i=10(PL)			(10 . 0.5)

Εφαρμόζοντας τον τύπο (4.1) σε κάθε στήλη του πίνακα 4.3 παίρνουμε τα αντίστοιχα πολυώνυμα παρεμβολής. Πιο συγκεκριμένα, για την πρώτη στήλη το πολυώνυμο που διέρχεται από τα σημεία: (1,1), (2,1), (3,1), (4,1) είναι το $P_1(i) = 1$.

Για τη δεύτερη στήλη, το πολυώνυμο που περνά από τα σημεία: (1, 2.5), (2, 1.5), (5, 5), (6, 1), (7, 0.5) και (8,1) είναι το

$P_2(i) = 38\text{-}67{,}2488i + 41.0375i^2 - 10{,}4036i^3 + 1{,}1625i^4 - 0{,}047619i^5$.

Για την τρίτη στήλη το πολυώνυμο που περνά από τα σημεία: (1, 2.5), (3, 1), (4, 1), (9, 1), (10, 0.5) είναι το

$P_3(i) = 4{,}60714 - 2{,}7371i + 0{,}695602i^2 - 0{,}067791i^3 + 0{,}00214947i^4$.

Για να προσδιορισθεί η συνάρτηση f(i,j) απαιτείται να βρεθεί το πολυώνυμο παρεμβολής Lagrange, το οποίο αντιστοιχεί τις τιμές j=1,2,3, στα αντίστοιχα πολυώνυμα $P_1(i)$, $P_2(i)$, $P_3(i)$ δηλαδή παρεμβάλλεται στα σημεία:

(1, $P_1(i)$), (2, $P_2(i)$) και (3, $P_3(i)$)

Το ζητούμενο πολυώνυμο είναι:

$$P(j) = 3P_1(i) - 3P_2(i) + P_3(i) + \left(-\frac{5P_1(i)}{2} + 4P_2(i) - \frac{3P_3(i)}{2}\right)j + \left(\frac{P_1(i)}{2} - P_2(i) + \frac{P_3(i)}{2}\right)j^2$$

Αντικαθιστώντας τα πολυώνυμα $P_j(i)$ στην παραπάνω έκφραση βρίσκουμε την τελική μορφή της συνάρτησης η οποία είναι :

$$\begin{aligned} f(i,j) = {} & \left(-106{,}393 + 199{,}009i - 122{,}417i^2 + 31{,}1429i^3 - 3{,}48535i^4 + 0{,}142857i^5\right) \\ & + \left(142{,}589 - 264{,}89i + 163{,}107i^2 - 41{,}5126i^3 - 4{,}64678i^4 - 0{,}190476i^5\right)j \\ & + \left(-35{,}1964 + 65{,}8803i - 40{,}6897i^2 + 10{,}3697i^3 - 1{,}16143^4 + 0{,}047619i^5\right)j^2\Big) \end{aligned} \qquad (4.7)$$

Στο δεύτερο στάδιο επιχειρείται η μοντελοποίηση του βαθμού κρισιμότητας σε σχέση με το κλάσμα έμφραξης (y). Η εύρεση μιας συνεχούς συνάρτησης g(y) η οποία να αποδίδει το βαθμό κρισιμότητας της νόσου, επιτυγχάνεται με την χρήση μη γραμμικού μοντέλου παλινδρόμησης.

Τα σημεία που χρησιμοποιούνται για τον προσδιορισμό του μοντέλου σύμφωνα με τον πίνακα 4.1 είναι τα εξής:

ΠΙΝΑΚΑΣ 4.4

Κλάσμα στένωσης (y)	0,25	0,50	0,75	0,90	0,99	1,00
Βαθμός κρισιμότητας g(y)	1	2	4	8	16	32

Στην ανάπτυξη του μοντέλου δε χρησιμοποιήθηκαν τα δύο ακραία σημεία (0.25 ,1) και (1, 32). Όπως έχει ήδη ορισθεί, ένα άτομο χαρακτηρίζεται ως στεφανίαιος ασθενής όταν το ποσοστό έμφραξης μιας αρτηρίας είναι μεγαλύτερο ή ίσο του 70% ($y \geq 0,7$), άρα η τιμή έμφραξης y=0,25 κατατάσσει το άτομο στους υγιείς. Η ακραία τιμή y=1 σημαίνει πλήρη έμφραξη της αρτηρίας, που δεν είναι αντιστρέψιμη κατάσταση ως εκ τούτου αποδόθηκε ο βαθμός κρισιμότητας g(y) = 32 όπως και στο δείκτη Gensini. Επιπλέον, δεν μπορούν να ληφθούν τιμές του y μεταξύ 0,99 και 1,00 με τις μεθόδους που χρησιμοποιούνται.

Το μοντέλο το οποίο θεωρείται το πλέον κατάλληλο είναι της μορφής

$$g(y) = \text{Exp}(\text{Exp}(b+ay)) \qquad (4.8)$$

Ο λόγος του διπλού εκθετικού αποδίδεται στο μεγάλο ρυθμό μεταβολής του βαθμού σοβαρότητας σε σχέση με το ποσοστό έμφραξης. Τα αποτελέσματα της ανάλυσης παλινδρόμησης παρουσιάζονται στον πίνακα 4.5 που ακολουθεί.

ΠΙΝΑΚΑΣ 4.5:Ανάλυση διασποράς

	Βαθμοί Ελευθερίας	**Άθροισμα Τετραγώνων**	**Μέσο Τετραγωνικό σφάλμα**
Διασπορά	1	1,08356	1,0835511
Κατάλοιπα	2	0,000066	0,0003282

ΠΙΝΑΚΑΣ 4.6: Υπολογισμός παραμέτρων της εξίσωσης (4.7)

Μεταβλητές	**συντελεστές**	**Τυπικό σφάλμα**	**t-test**	**Σημαντικότητα (α)**
y	2,8052	0,0488	57,457	0,0003
Σταθερός όρος	-1,77405	0,03938	-45,048	0,0005

Παρατηρούμε ότι η μεταβλητή y είναι στατιστικώς σημαντική και ότι το μοντέλο ερμηνεύει το 99,97% της μεταβλητότητας της εξαρτημένης μεταβλητής (R^2=0.9997).

Η συνάρτηση παλινδρόμησης δίδεται από τη σχέση

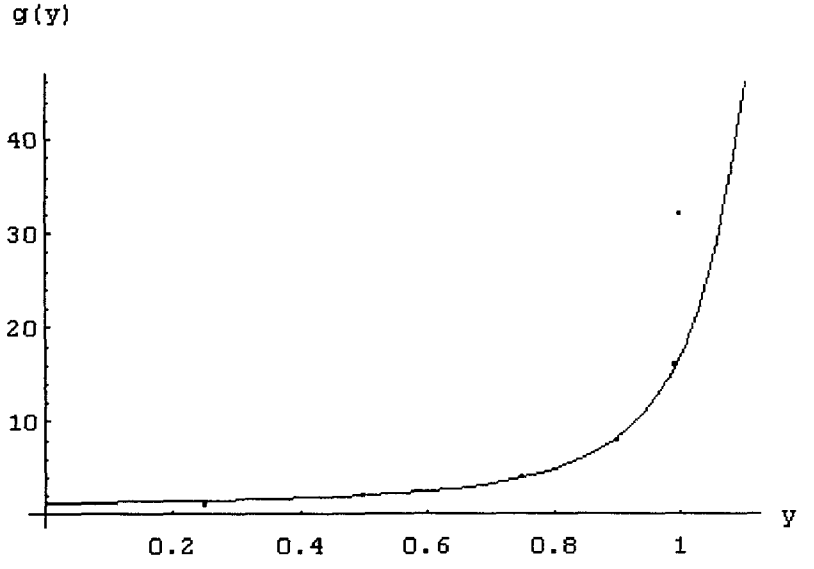
$$g(y) = Exp(Exp(-1,77405+2,8052y)) \qquad (4.9)$$

Το μοντέλο που προκύπτει με τη χρήση του ακραίου σημείου (1, 32) ερμηνεύει το 84,89% της μεταβλητότητας του y.

Στον πίνακα (4.7) που ακολουθεί παρατηρούμε την απόκλιση του δείκτη Gensini και των τιμών της g(y). Το μοντέλο συμπεριφέρεται αρκετά καλά για $y \geq 0,25$ ενώ για την ακραία τιμή y = 1 υπάρχει μεγάλη απόκλιση, όπως φαίνεται και στο σχήμα 4.3.

ΠΙΝΑΚΑΣ 4.7: Συγκριτικός πίνακας τιμών του δείκτη Gensini και των τιμών της g(y)

y	τιμή του Gensini	g(y)
0,25	1	1,407
0,50	2	1,993
0,75	4	4,018
0,90	8	8,317
0,99	16	15,283
1	32	16,515



Σχ. 4.3

Αξίζει να σημειωθεί ότι επιχειρήθηκε και η εφαρμογή εκθετικού μοντέλου της μορφής: $g(y) = Exp\,(Exp\,(\,c+by+ay^2\,)\,)$ (4.10)

Από τον υπολογισμό των παραμέτρων της εξίσωσης (4.10) προκύπτει το παρακάτω μοντέλο:

$$g(y) = Exp\,(Exp\,(\,-1{,}604+2{,}312y+0{,}334y^2\,)\,)$$

Η ανάλυση παλινδρόμησης αναδεικνύει το δευτεροβάθμιο παράγοντα του μοντέλου μη στατιστικά σημαντικό όπως προκύπτει από τις τιμές του επιπέδου σημαντικότητας που καταγράφονται κατωτέρω:

ΠΙΝΑΚΑΣ 4.8

Μεταβλητές	**Μη Κανονικοποιημένοι συντελεστές**		**t-test**	**Επίπεδο σημαντικότητας (α)**
		Τυπικό σφάλμα		
y	2,312	0,563	4,107	0,152
y^2	0,334	0,380	0,880	0,540
Σταθερός	-1,604	0,197	-8,134	0,078

Το επίπεδο σημαντικότητας α = 0,018 του F-test ενισχύει την άποψη για την ύπαρξη πολυσυγγραμμικότητας μεταξύ των y^2 και y κατά συνέπεια το μοντέλο (4.9) κρίνεται καταλληλότερο του (4.10).

Λόγω της σημαντικής απόκλισης που εμφανίζει το μοντέλο (4.9) για y=1 από την προβλεπόμενη τιμή Gensini, ο τελικός δείκτης ds(i,y,j) που προτείνεται, ορίζεται ως εξής :

$$ds(i,y,j)=\begin{cases} f(i,j).g(y) & ,1\le i\le 10,\ 0{,}25\le y\le 0{,}99,\quad 1\le j\le 3 \\ 32.f(i,j) & ,1\le i\le 10,\ y=1,\quad 1\le j\le 3 \end{cases} \qquad (4.11)$$

Έτσι για $1\le i\le 10,\ 0{,}25\le y\le 0{,}99,\quad 1\le j\le 3$ έχουμε:

$$\begin{aligned} ds(i,y,j) &= f(i,j)g(y) \\ &= (-106{,}393+199{,}009i-122{,}417i^2+31{,}1429i^3-3{,}48535i^4+0{,}142857i^5 \\ &+(142{,}589-264{,}89i+163{,}107i^2-41{,}5126i^3-4{,}64678i^4-0{,}190476i^5)j \\ &+(-35{,}1964+65{,}8803i-40{,}6897i^2+10{,}3697i^3-1{,}16143i^4+0{,}047619i^5)j^2)\square \\ &\square Exp[Exp(-1{,}77405+2{,}8052y)] \end{aligned}$$

Ενώ για $1\le i\le 10,\ y=1,\quad 1\le j\le 3$

$$\begin{aligned} ds(i,y,j) &= 32f(i,j) \\ &= 32(-106{,}393+199{,}009i-122{,}417i^2+31{,}1429i^3-3{,}48535i^4+0{,}142857i^5 \\ &+(142{,}589-264{,}89i+163{,}107i^2-41{,}5126i^3-4{,}64678i^4-0{,}190476i^5)j \\ &+(-35{,}1964+65{,}8803i-40{,}6897i^2+10{,}3697i^3-1{,}16143i^4+0{,}047619i^5)j^2) \end{aligned}$$

Η συνάρτηση ds(i,y,j) είναι συνεχής για τιμές του κλάσματος στένωσης (y) από 0,25 ως 0,99 ενώ εμφανίζει ασυνέχεια για την τιμή y = 1. Επιτυγχάνεται έτσι ο υπολογισμός ενός δείκτη κρισιμότητας της νόσου ds(i,y,j) ο οποίος δεν περιέχει εμπειρικές εκτιμήσεις για το βαθμό κρισιμότητας της νόσου. Επιπλέον είναι προφανής η εύκολη χρήση του μέσω υπολογιστή [Kokla et al (2000)].

Στην περίπτωση που υπάρχουν περισσότερες από μια εμφράξεις, ο τελικός δείκτης για κάθε αρτηρία (j) θα δίνεται από το άθροισμα των επιμέρους τιμών της παραπάνω συνάρτησης $ds(i,j,y)=ds_j(i, y)$

Έστω ds_1, ds_2, ds_3, οι επιμέρους δείκτες που αντιστοιχούν στις τρεις αρτηρίες (RCA, LAD ,LCX). Ακολουθώντας το σκεπτικό του Gensini, ο τελικός δείκτης DSN θα είναι το άθροισμα των επιμέρους δεικτών. Δηλαδή

$$DSN= ds_1+ds_2+ ds_3. \qquad (4.12)$$

ΜΕΛΕΤΗ ΤΩΝ ΠΑΡΑΓΟΝΤΩΝ ΚΙΝΔΥΝΟΥ ΣΤΟΥΣ ΑΣΘΕΝΕΙΣ

Στη συνέχεια μελετάται η συσχέτιση των παραγόντων κινδύνου στο σύνολο των ασθενών. Μέσω παραγοντικής ανάλυσης στο δείγμα ασθενών ανδρών επιχειρείται ο εντοπισμός παραγόντων που ενοχοποιούνται για την εκδήλωση της νόσου.

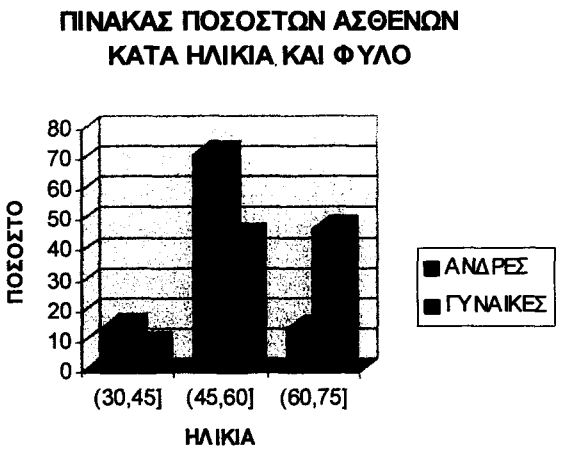
Υπολογίζονται επίσης οι τιμές του νεοεισαγώμενου DSN και τα στατιστικά του δείκτη αυτού στο δείγμα ασθενών ανδρών στο οποίο η καταγραφή των στοιχείων είναι πλήρης, σε αντίθεση με το δείγμα γυναικών όπου σε πολλές περιπτώσεις δεν καταγράφεται το ποσοστό της στένωσης των αρτηριών. Μελετάται επίσης η σχέση του DSN με το δείκτη Gensini καθώς και με τους παράγοντες κινδύνου, και επιχειρείται η χρήση του για την πρόβλεψη του εμφράγματος και της παράπλευρης κυκλοφορίας.

4.3 Μετρήσεις στο συνολικό αρχείο ασθενών

Το συνολικό αρχείο των 580 ασθενών αποτελείται από 416 (72%) άνδρες και 164(28%) γυναίκες. Ο αριθμός των ασθενών, ανδρών – γυναικών, ανά ομάδα ηλικιών φαίνεται στον πίνακα 4.9 και στο ραβδόγραμμα του σχήματος 4.4.

ΠΙΝΑΚΑΣ 4.9: Κατανομή συχνοτήτων ασθενών κατά φύλο και ηλικία.

	ΑΝΔΡΕΣ		ΓΥΝΑΙΚΕΣ	
ΗΛΙΚΙΑ	ΣΥΧΝΟΤΗΤΑ	ΠΟΣΟΣΤΟ(%)	ΣΥΧΝΟΤΗΤΑ	ΠΟΣΟΣΤΟ(%)
(30, 45]	61	14,60	14	8,50
(45,60]	296	71,15	73	44,50
(60,75]	59	14,25	77	47.00
Σύνολο	416		164	



Σχ. 4.4

Στο σχήμα 4.4 και στον πίνακα 4.9, παρατηρείται ότι, το ποσοστό των ασθενών γυναικών αυξάνεται σημαντικά μετά την ηλικία των 45 ετών, ενώ η αύξηση μεταξύ των ομάδων (45,60] και (60,75] εμφανίζεται στατιστικά μη σημαντική. Στους άνδρες το ποσοστό των ασθενών εμφανίζει αύξηση από την ηλικία (30, 45] στην ηλικία (45, 60] ενώ στη συνέχεια παρατηρείται μείωση στο ποσοστό ασθενών στην τρίτη ομάδα ηλικίας (60, 75].

Ο έλεγχος διαφοράς ποσοστών μεταξύ των ομάδων του ιδίου φύλου που έγινε με το χ^2 – test (χ_*^2), αποκαλύπτει ότι στον πληθυσμό των ασθενών γυναικών τα ποσοστά που εμφανίζονται στις δύο τελευταίες ομάδες ηλικιών είναι ίσα [$\chi^2_{(0.05)}$ =3.84 και χ_*^2=0.106, άρα $\chi_*^2 < \chi^2_{(0.05)}$]. Ομοίως για τον πληθυσμό των ανδρών ασθενών η εφαρμογή του χ^2 – test, δείχνει ότι τα ποσοστά των ομάδων (30, 45] και (60, 75] δε διαφοροποιούνται [$\chi^2_{(0.05)}$ =3.84 και χ_*^2=0.033].

Ο έλεγχος της ισότητας των ποσοστών των ασθενών ανά ομάδα ηλικίας στα δύο φύλα έδειξε ότι στις ίδιες ομάδες ηλικιών τα ποσοστά ανδρών-γυναικών διαφοροποιούνται.

Στο συνολικό αρχείο ασθενών μελετήθηκαν οι αλληλεξαρτήσεις των παραγόντων κινδύνου. Οι τιμές του επιπέδου σημαντικότητας (α) των μεταβλητών βρέθηκαν με τη βοήθεια του χ^2-test και παρουσιάζονται στον πίνακα 4.10. Καταγράφηκαν οι τιμές του επιπέδου σημαντικότητας μόνο στις μεταβλητές που εμφανίζουν εξάρτηση.

ΠΙΝΑΚΑΣ 4.10 :Τιμές του επίπεδου σημαντικότητας (α) μεταξύ των παραγόντων ΚΙΝΔΥΝΟΥ

	HBP	SMOKE	FAMILY	TYPE A	SEDENTARY	DIABETES	OBESITY
AG	0,001	0,000	0,023	0,000	0,027	0,000	
LIPIDS				0,000	0,027	0,005	
SMOKE	0,000			0,050	0,000		0,036
OBESITY	0,005				0,024	0,009	
SEDEN-	0,018						
TYPE A	0,000				0,000		

Παρατηρείται ότι οι μεταβλητές που εμφανίζονται στατιστικώς εξαρτημένες στους ασθενείς είναι οι ίδιες που προέκυψαν από τη μελέτη του συνολικού δείγματος, πρέπει μόνο να προσθέσουμε τη συσχέτιση της μεταβλητής καθιστική ζωή (SEDENTARY) με τα λιπίδια (LIPIDS) και της παχυσαρκίας (OBESITY) με το κάπνισμα (SMOKE), όπως επίσης και του διαβήτη (DIABETES) με τις ομάδες ηλικιών (AG).

Παρατηρήθηκε επίσης ότι υπάρχει στατιστικώς σημαντική συσχέτιση μεταξύ των μεταβλητών:

VD και AG	α = 0,020
VD και DIABETES	α = 0,018
EFTG και DIABETES	α = 0,000
EFTG και TYPE A	α = 0,001

Ελέγχθηκε επίσης η εξάρτηση μεταξύ του φύλου και του αριθμού των φραγμένων αρτηριών. Παρατηρήθηκε ότι οι δύο μεταβλητές εμφανίζονται εξαρτημένες (α=0,020), η κατανομή του αριθμού των φραγμένων αρτηριών ως προς τα δύο φύλα παρουσιάζεται στον πίνακα που ακολουθεί:

ΠΙΝΑΚΑΣ 4.11

	Αριθμός φραγμένων αρτηριών VD			
Φύλο	1	2	3	Σύνολο
0	75	47	42	164
1	139	153	124	416
Σύνολο	214	200	166	580

Στις γυναίκες το ποσοστό των ασθενών με νόσο μιας αρτηρίας εμφανίζεται αυξημένο, ενώ στους άνδρες το μεγαλύτερο ποσοστό των ασθενών πάσχει από νόσο δύο αρτηριών.

Όπως και στο συνολικό αρχείο, ελέγχθηκε η εξάρτηση μεταξύ και των παραγόντων κινδύνου και του φύλου. Τα αποτελέσματα του ελέγχου παρατίθενται στον ακόλουθο πίνακα.

ΠΙΝΑΚΑΣ 4.12: Τιμές του επιπέδου σημαντικότητας (α) του χ^2 -test των παραγόντων κινδύνου με το φύλο

Παράγοντες κινδύνου	Επίπεδο Σημαντικότητας (α)
AG	0,000
DIABETES	0,000
HBP	0,000
SMOKE	0,000
SEDENTARY	0,000
TYPE A	0,000
VD	0.020

Όσον αφορά την εξάρτηση του φύλου του ασθενούς με το διαβήτη (πίνακας 4.12), τα ποσοστά των διαβητικών ασθενών κατά φύλο φαίνονται στον πίνακα 4.13.

ΠΙΝΑΚΑΣ 4.13

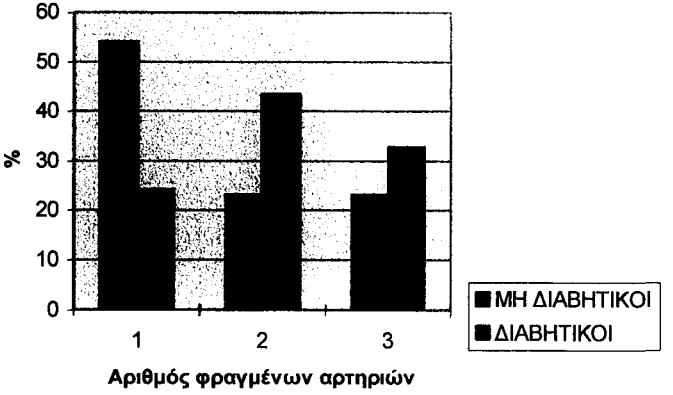
	Άνδρες	Γυναίκες
Διαβητικοί	48(11.5%)	46 (28 %)

Ο έλεγχος της ισότητας των ποσοστών διαβητικών ασθενών στα δύο φύλα έδειξε στατιστικά σημαντικά αυξημένα ποσοστά διαβητικών γυναικών. Στον πίνακα 4.14 καθώς και στο ραβδόγραμμα του σχήματος 4.5 που ακολουθεί, παρουσιάζονται τα ποσοστά των ασθενών γυναικών με 1,2 και 3 φραγμένες αρτηρίες σε σχέση με την παρουσία διαβήτη.

ΠΙΝΑΚΑΣ 4.14

VD / DIABETES	1	2	3	ΣΥΝΟΛΟ
0	64 (54%)	27 (23%)	27 (23%)	118 (100%)
1	11 (24%)	20 (43.4%)	15 (32.6%)	46 (100%)

Ποσοστά ασθενών γυναικών με 1,2,3, φραγμένες αρτηρίες σε σχέση με την παρουσία διαβήτη.



Σχ. 4.5

Παρατηρούμε ότι το ποσοστό των μη διαβητικών γυναικών με μια φραγμένη αρτηρία είναι υπερδιπλάσιο σε σχέση με αυτό των διαβητικών. Μια προσεκτική ματιά στο διάγραμμα αποκαλύπτει ότι το ποσοστό των μη διαβητικών με δύο φραγμένες αρτηρίες είναι το ίδιο με αυτό των μη διαβητικών με τρεις .

Η κατάσταση αντιστρέφεται στην περίπτωση των δύο φραγμένων αρτηριών. Το ποσοστό των διαβητικών γυναικών με νόσο δύο αρτηριών εμφανίζεται αυξημένο έναντι των ασθενών γυναικών που δεν πάσχουν από διαβήτη (α=0,009). Δε συμβαίνει όμως το ίδιο και με τα ποσοστά των ασθενών με τρεις φραγμένες αρτηρίες (α=0,112).

Το συμπέρασμα που μπορεί να εξαχθεί από την ανάλυση του συγκεκριμένου δείγματος είναι ότι η ύπαρξη διαβήτη επιδεινώνει τη νόσο δύο αγγείων.

Ανάλογη μελέτη έγινε και για τον επίσης σημαντικό παράγοντα της υπέρτασης. Ο αριθμός και τα ποσοστά υπερτασικών για τα δύο φύλα δίδονται στον πίνακα 4.15.

ΠΙΝΑΚΑΣ 4.15: Κατανομή συχνοτήτων των υπερτασικών ως προς το φύλο στο σύνολο των ασθενών

SEX	Άνδρες	Γυναίκες
	113 (27,2 %)	81(49,4%)
Σύνολο	416	164

Στο συγκεκριμένο δείγμα είναι εμφανές ότι το ποσοστό των υπερτασικών γυναικών είναι μεγαλύτερο από το αντίστοιχο ποσοστό των υπερτασικών ανδρών. Στο συμπέρασμα αυτό καταλήξαμε με τη χρήση z-test.

Η τιμή που προέκυψε από τον έλεγχο (z=5.103) συγκρίθηκε με την τιμή z*=1,65 για επίπεδο σημαντικότητας α=0,05.

4.4 Μετρήσεις στο αρχείο των ασθενών ανδρών

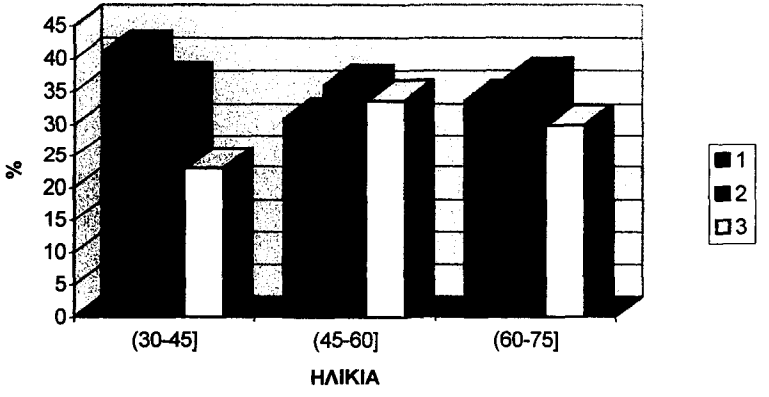
Η κατανομή του αριθμού των φραγμένων αρτηριών στους ασθενείς άνδρες σε σχέση με την ηλικία φαίνεται στον Πίνακα 4.16 και απεικονίζεται στο Σχήμα 4.6.

ΠΙΝΑΚΑΣ 4.16

ΗΛΙΚΙΑ / VD	(30-45]	(45-60]	(60-75]	ΣΥΝΟΛΟ
1	25	91	23	139
2	22	106	25	153
3	14	99	11	124
ΣΥΝΟΛΟ	61	296	59	416

Χρησιμοποιήθηκε το χ^2-test για να ελεγχθεί η εξάρτηση του αριθμού των φραγμένων αρτηριών από τις ομάδες υλικών. Τα αποτελέσματα του χ^2 -test είναι τα ακόλουθα: $\chi^2 = 7,55$, βαθμοί ελευθερίας 4, σε επίπεδο σημαντικότητας $\alpha = 0,109$. Σύμφωνα με τα παραπάνω συμπεραίνουμε ότι ο αριθμός των φραγμένων αρτηριών των ασθενών ανδρών δεν εξαρτάται από την ηλικία.

Κατανομή του αριθμού των φραγμένων αρτηριών ανά ηλικία



Σχ. 4.6

Το σχήμα 4.6 απεικονίζει τον αριθμό των ασθενών με νόσο μιας, δύο ή τριών αρτηριών ως προς το σύνολο των ασθενών ανά ομάδα ηλικίας.

Παρατηρήθηκε ότι η νόσος μονήρους στελέχους εμφανίστηκε σε 3 άνδρες (2 είχαν ηλικία 45-60 έτη και ένας ηλικία στο διάστημα 60-75 ετών).

Στο αρχείο ασθενών ανδρών ορίζουμε τέσσερις νέες δυαδικές μεταβλητές, "ακινησία (AKINESIA)", "ανεύρυσμα (ANEVRISMA)", "παράπλευρη κυκλοφορία (CCC)", και "έμφραγμα (MI)", ως εξής:

$$(\text{AKINESIA})_i = \begin{cases} 0 \text{ όταν κανένα τμήμα της καρδιάς δεν αδρανεί για το ιοστό άτομο} \\ 1 \text{ όταν ένα τμήμα της καρδιάς αδρανεί για το ιοστό άτομο} \end{cases}$$

$$(\text{ANEVRISMA})_i = \begin{cases} 0 \text{ όταν δεν υπάρχει ανεύρυσμα για το ιοστό άτομο} \\ 1 \text{ όταν υπάρχει ανεύρυσμα για το ιοστό άτομο} \end{cases}$$

$$(\text{CCC}) = \begin{cases} 0 \text{ όταν δεν υπάρχει παράπλευρο τριχοειδές αγγείο για το ιοστό άτομο} \\ 1 \text{ όταν υπάρχει παράπλευρο τριχοειδές αγγείο για το ιοστό άτομο} \end{cases}$$

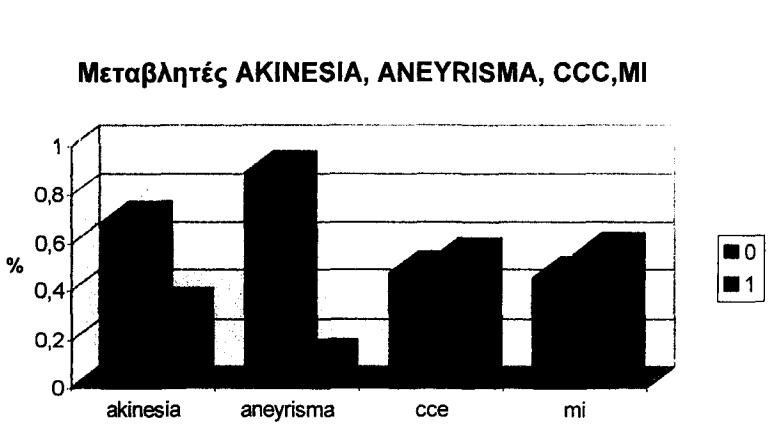
$$(\text{MI})_i = \begin{cases} 0 \text{ όταν δεν υπάρχει καταγραφή εμφράγματος στο ιοστό άτομο} \\ 1 \text{ όταν υπάρχει καταγραφή εμφράγματος στο ιοστό άτομο} \end{cases}$$

Στον πίνακα 4.15 παρουσιάζεται η κατανομή συχνοτήτων των τεσσάρων νέων μεταβλητών και στο σχήμα 4.7 τα αντίστοιχα συγκριτικά ραβδογράμματα. Η τιμή 0 των μεταβλητών δηλώνει την απουσία της πάθησης στο σύνολο των 416 ατόμων.

ΠΙΝΑΚΑΣ 4.17: Κατανομή συχνοτήτων των μεταβλητών AKINESIA, ANEVRISMA, CCC, MI

	Akinesia	ANEVRISMA	CCC	MI
0	283 (68%)	389 (88,7%)	196 (47,1%)	187 (45%)
1	133 (32%)	47 (11,3%0	220 (52,9%)	229 (55%)

Στον παραπάνω πίνακα καθώς και στο σχήμα 4.7 που ακολουθεί, γίνεται εμφανές ότι το ποσοστό των ασθενών από στεφανιαία νόσο που παρουσιάζουν έμφραγμα ή παράπλευρη κυκλοφορία είναι αυξημένο.



Σχ. 4.7

Ο έλεγχος συσχέτισης των παραπάνω μεταβλητών με τους παράγοντες κινδύνου, με τις ομάδες ηλικιών και με τη δυαδική μεταβλητή EFTG έγινε με τη χρήση του χ^2-test και οι τιμές του επιπέδου σημαντικότητας (α) καταγράφονται στον πίνακα 4.18. Στατιστικώς σημαντική συσχέτιση αντιστοιχεί σε επίπεδα μικρότερα του 0.05.

ΠΙΝΑΚΑΣ 4.18

	AKINESIA	ANEVRISMA	CCC	MI
AG	0,267	0,957	0,234	0,184
EFTG	**0,000**	**0,000**	**0,005**	**0,000**
LIPIDS	0,402	0,696	**0,000**	0,496
SMOKE	0,591	0,082	**0,000**	**0,029**
HBP	0,975	0,436	0,620	0,169
DIABETES	0,586	0,082	**0,000**	0,627
OBESITY	0,322	0,077	0,689	0,378
FAMILY	0,186	0,271	0,198	**0,007**
SEDENTARY	0,106	**0,004**	0,265	0,677
TYPE A	0,379	0,617	**0,000**	**0,029**

Η εξάρτηση της μεταβλητής EF από τις μεταβλητές AKINESIA, ANEVRISMA, CCC, MI ήταν αναμενόμενη, απλώς μπορούμε να ισχυρισθούμε

ότι επιβεβαιώνεται η ορθότητα στατιστικών ευρημάτων στο χώρο της καρδιολογίας [Selzer (1982)]. Μπορούμε επίσης να ισχυρισθούμε ότι στεφανιαίοι άνδρες ασθενείς υπερλιπιδαιμικοί, καπνιστές, διαβητικοί και τύπου Α έχουν μεγάλη πιθανότητα να εμφανίσουν παράπλευρη κυκλοφορία όπως και οι καπνιστές, που έχουν οικογενειακό ιστορικό, και οι τύπου Α έχουν κίνδυνο εμφράγματος.

4.5 Παραγοντική ανάλυση του συνόλου των ασθενών ανδρών

Με τη χρήση της παραγοντικής ανάλυσης επιχειρείται και στο σύνολο των ασθενών ανδρών όπως και στο πιλοτικό δείγμα η διερεύνηση των σχέσεων και αλληλεπιδράσεων μεταξύ των ανεξάρτητων μεταβλητών και η εύρεση εκείνων των παραγόντων που δρουν επιβαρυντικά. Για την εύρεση των κύριων συνιστωσών στο δείγμα των ασθενών ανδρών θα χρησιμοποιηθεί η ίδια τεχνική όπως και στο πιλοτικό δείγμα.

Ο πίνακας 4.19 δείχνει ότι υπάρχουν πέντε κύριες συνιστώσες αν χρησιμοποιήσουμε ως κριτήριο τις ιδιοτιμές εκείνες που είναι μεγαλύτερες του ένα ($\lambda > 1$), ή οκτώ κύριες συνιστώσες με το κριτήριο $\lambda > 0,7$ που προτείνει η Jolliffee.

ΠΙΝΑΚΑΣ 4.19: Συνολική διασπορά η οποία εξηγείται από τους παράγοντες

	Αρχικές ιδιοτιμές			Άθροισμα των φορτίων στο τετράγωνο		
Παράγοντες	Ιδιοτιμές	% της διασποράς	Αθροιστική %	Ιδιοτιμές	% της διασποράς	Αθροιστική %
1	1,623	16,229	16,229	1,623	16,229	16,229
2	1,278	12,783	29,012	1,278	12,783	29,012
3	1,170	11,698	40,711	1,170	11,698	40,711
4	1,081	10,806	51,517	1,081	10,806	51,517
5	1,051	10,510	62,026	1,051	10,510	62,026
6	0,953	9,533	71,559			
7	0,862	8,623	80,182			
8	0,726	7,262	87,443			
9	0,679	6,795	94,238			
10	0,576	5,762	100,000			

Η ανάλυση αυτή σε πέντε κύριες συνιστώσες ερμηνεύει το 62.03% της συνολικής μεταβλητότητας των αρχικών μεταβλητών ή αντίστοιχη σε οκτώ κύριες συνιστώσες ερμηνεύει το 87,4% αντίστοιχα. Ο εμπειρικός κανόνας που χρησιμοποιείται εδώ είναι ότι μια μεταβλητή, αντιστοιχεί στον παράγοντα εκείνο, για τον οποίο τα φορτία στον πίνακα 4.20 είναι μεγαλύτερα κατ' απόλυτο τιμή του 0,5 [Sharma (1996)].

ΠΙΝΑΚΑΣ 4.20: Φορτία μετά από περιστροφή

	ΠΑΡΑΓΟΝΤΕΣ				
	1	2	3	4	5
DIABETES	**0.579**	0.152	0.334	0.120	-0.131
EF	**-0.704**	-0.086	0.231	0.345	0.0094
AGE	0.124	**0.814**	0.0911	0.167	0.0197
TYPE A	**0.610**	-0.278	-0.0366	0.101	0.180
SMOKE	0.253	**-0.607**	0.211	0.212	0.173
HBP	-0.226	0.312	**0.553**	-0.145	0.432
OBESITY	0.049	-0.140	**0.837**	0.007	-0.196
FAMILY	0.0089	-0.0315	0.0434	**-0.922**	0.00036
LIPIDS	0.33	-0.148	0.0074	0.165	**0.459**
SEDENTARY	0.0415	0.0324	0.121	0.059	**-0.774**

Μέθοδος: Ανάλυση κατά παράγοντες
Μέθοδος περιστροφής:Varimax.

Από τον πίνακα 4.20 μπορούμε να θεωρήσουμε ότι υπάρχουν πέντε παράγοντες οι οποίοι περιέχουν τις εξής μεταβλητές:

Ο πρώτος παράγοντας:	DIABETES, EF, TYPE A (προδιάθεση)
ο δεύτερος:	AGE, SMOKE (συνήθειες)
ο τρίτος:	HBP, OBESITY (διατροφή)
ο τέταρτος:	FAMILY (ιστορικό)
ο πέμπτος:	LIPIDS, SEDENTARY (τρόπος ζωής).

Ο πίνακας 4.21 που ακολουθεί δίνει το ποσοστό της μεταβλητότητας της κάθε αρχικής μεταβλητής που ερμηνεύεται από τους πέντε αυτούς παράγοντες.

ΠΙΝΑΚΑΣ 4.21

Μεταβλητές	**cummunalities**
DIABETES	0.501
EF	0.676
AGE	0.715
TYPE A	0.493
SMOKE	0.552
HBP	0.662
OBESITY	0.76
FAMILY	0.854
LIPIDS	0.369
LIPIDS	0.620

Παρατηρούμε ότι τα λιπίδια έχουν το χαμηλότερο ποσοστό ερμηνείας ενώ το υψηλότερο είναι αυτό της μεταβλητής OBESITY (παχυσαρκία).

4.6 Εφαρμογή του δείκτη DSN στο δείγμα ασθενών ανδρών.

Ο δείκτης Gensini, σύμφωνα με ιατρικές αναφορές, αποτελεί έναν πολύ σημαντικό δείκτη κρισιμότητας της νόσου στην καρδιολογία. Στον προτεινόμενο εναλλακτικό DSN, θα επιχειρηθεί η εύρεση στατιστικών χαρακτηριστικών που προκύπτουν από την εφαρμογή του στο δείγμα των 416 ασθενών ανδρών.

Η τιμή του δείκτη DSN υπολογίσθηκε για κάθε άνδρα ασθενή σύμφωνα με την καταγραφή των εμφράξεων των αρτηριών που έγινε από τους ιατρούς του Ιπποκράτειου Νοσοκομείου, επίσης για το ίδιο δείγμα υπάρχει καταγεγραμμένη η αντίστοιχη τιμή του δείκτη Gensini.

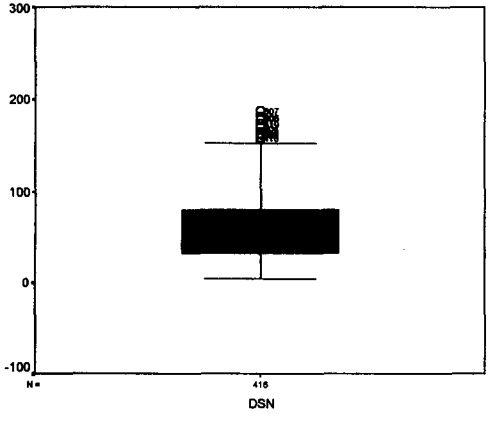
4.6.1 Τα στατιστικά του δείκτη DSN

Τα στατιστικά του δείκτη DSN που προκύπτουν από το δείγμα των 416 ατόμων δίνονται στον πίνακα 4.22

ΠΙΝΑΚΑΣ 4.22

Μέση τιμή	60,08
Τυπικό σφάλμα	1,80
Τυπική απόκλιση	36,72
Mln	4,02
max	186,35
ασυμμετρία	0,9358
κύρτωση	0,653

Το θηκόγραμμα (box-plot) του δείκτη DSN παρουσιάζεται στο σχήμα 4.8.



Σχ. 4.8: Θηκόγραμμα του δείκτη DSN

Υπολογίσθηκε η συσχέτιση των δύο δεικτών στο δείγμα των ασθενών ανδρών και παρατηρήθηκε ότι ο δείκτης DSN σχετίζεται γραμμικά με τον κλασσικό δείκτη Gensini με δείκτη προσδιορισμού R^2=0,977. Η γραμμική σχέση παλινδρόμησης περιγράφεται από την σχέση:

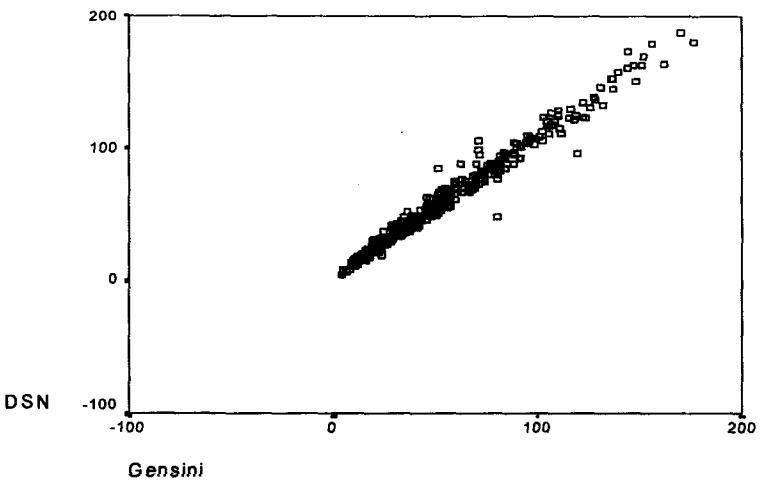
$$DSN=3,182+1,055\cdot Gensini \qquad (4.13)$$

σύμφωνα με τον πίνακα 4.23 που ακολουθεί:

ΠΙΝΑΚΑΣ 4.23

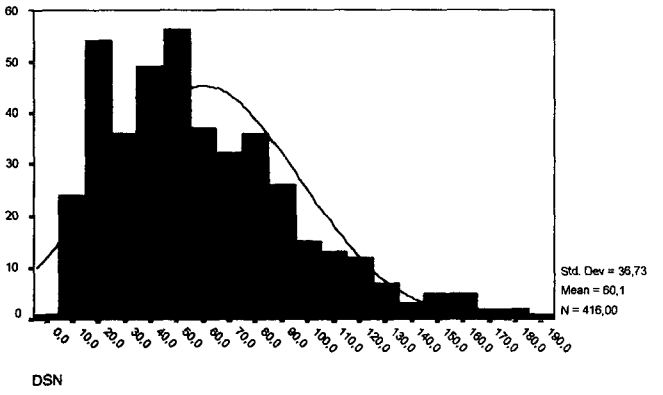
	Συντελεστές	Τυπικό σφάλμα	t	α
Σταθερά	3,182	0,51	6,243	0,000
Gensini	1,055	0,008	132,37	0,000

Το διάγραμμα διασποράς της συνάρτησης (4.13) φαίνεται στο σχήμα 4.9.



Σχ. 4.9

Στη συνέχεια θα εξεταστεί αν ο δείκτης DSN ακολουθεί την κανονική κατανομή. Το ιστόγραμμα του δείκτη DSN το οποίο δίδεται στο σχήμα 4.10 είναι μια ένδειξη ότι ο δείκτης δεν ακολουθεί την κανονική κατανομή.



Σχ. 4.10

Εφαρμόζοντας το test των Kolmogorov-Smirnov βρίσκουμε :

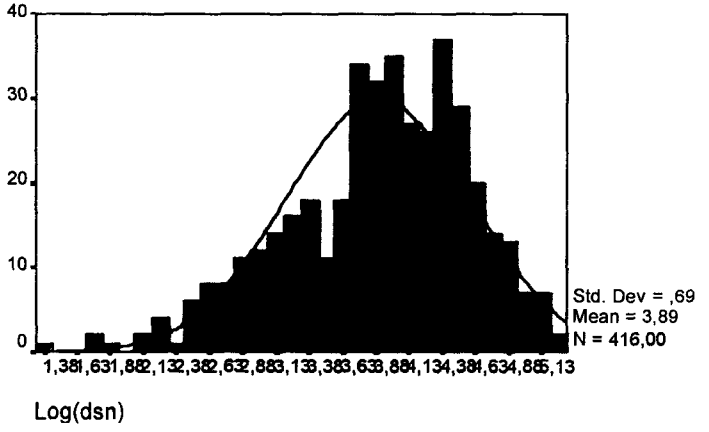
D=0,092,	Kolmogorov-Smirnov Z=1,87	ασυμπτωτικό α=0,002

Για να ελεγχθεί η υπόθεση ότι ο DSN ακολουθεί lognormal κατανομή, ελέγχεται η κανονικότητα του λογαρίθμου των παρατηρήσεων του DSN.

Τα αποτελέσματα του test των Kolmogorov-Smirnov είναι τα εξής:

D=0,070,	Kolmogorov-Smirnov Z=1,42	ασυμπτωτικό α = 0,035

Είναι προφανές ότι δεν μπορούμε να απορρίψουμε την υπόθεση ότι ο δείκτης DSN ακολουθεί την lognormal κατανομή (σχήμα 4.11) σε επίπεδο σημαντικότητας α = 0,01.



Σχ. 4.11: Ιστόγραμμα του λογαρίθμου του δείκτη DSN.

Λόγω του μεγάλου δείγματος το Κεντρικό Οριακό Θεώρημα επιτρέπει τον υπολογισμό διαστημάτων εμπιστοσύνης για τη μέση τιμή του δείκτη. Όλα τα διαστήματα εμπιστοσύνης είναι της μορφής

$$\bar{x} \pm z_{0,025} \frac{s}{\sqrt{n}}.$$

Ο Πίνακας 4.24 παρουσιάζει τα βασικά αριθμητικά χαρακτηριστικά του δείκτη DSN σε σχέση με τις μεταβλητές MI, ANEVRISMA, AKINESIA, CCC,EFTG

ΠΙΝΑΚΑΣ 4.24

	Δείκτης DSN		
	μέση τιμή	τυπική απόκλιση	διάστημα εμπιστοσύνης
MI=0	40,28	25,66	(36,58, 43,99)
MI=1	76,25	36,52	(71,49, 81,00)
ANEVRISMA	58,25	36,02	(54,59, 61,96)
ANEVRISMA	74,25	39,44	(62,67, 85,84)
AKINESIA =0	53,04	33,71	(49,09, 56,98)
AKINESIA =1	75,07	38,36	(68,47, 81,67)
CCC=0	46,72	32,72	(42,11, 51,33)
CCC=1	71,99	36,06	(67,19, 76,78)
EFTG=0	73,53	38,58	(66,26, 80,79)
EFTG=1	55,07	34,60	(51,16, 58,98)

Ενδιαφέρον παρουσιάζει το γεγονός ότι τα δύο διαστήματα εμπιστοσύνης ανά μεταβλητή δεν επικαλύπτονται. Έτσι αυτές οι μεταβλητές μπορούν να χρησιμοποιηθούν για την πρόβλεψη του δείκτη DSN. Επίσης οι δύο μέσες τιμές ανά μεταβλητή που αντιστοιχούν στις τιμές 0 και 1, είναι διαφορετικές. Πράγματι, για κάθε μεταβλητή και με τη χρήση του Z-test, λόγω του μεγέθους του δείγματος, απορρίπτεται η μηδενική υπόθεση

$$H_0 \ \mu_0 = \mu_1$$

έναντι της εναλλακτικής

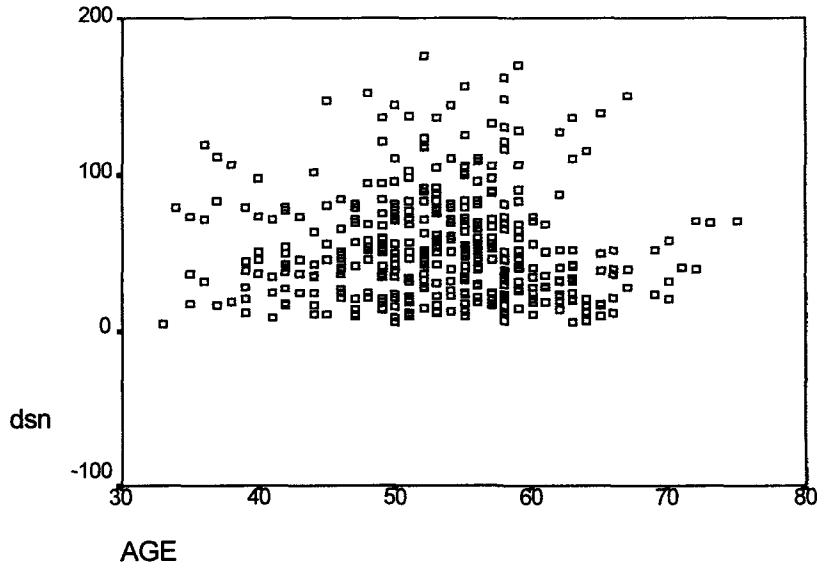
$$H_\alpha\ \mu_0 \neq \mu_1.$$

Παρουσιάζει επίσης ενδιαφέρον το γεγονός ότι δε συμβαίνει το ίδιο με τους παράγοντες κινδύνου. Από τον πίνακα που ακολουθεί είναι φανερό ότι για τις μεταβλητές SMOKE, OBESITY, LIPIDS (Garfagnini A (1995)), οι δύο ομάδες 0 και 1, έχουν τον ίδιο μέσο δείκτη DSN. Η μόνη εξαίρεση είναι η μεταβλητή FAMILY για την οποία οι δύο ομάδες 0 και 1, έχουν διαφορετικούς μέσους. Ο έλεγχος των μέσων έγινε με τη χρήση του t-test. Ο έλεγχος της ισότητας των διασπορών με τη βοήθεια του F-test έδειξε ότι σε επίπεδο σημαντικότητας α=0,05 οι δύο ομάδες που αντιστοιχούν στις τιμές 0 και 1 και για όλες τις μεταβλητές του πίνακα 4.25 είναι ίσες.

ΠΙΝΑΚΑΣ 4.25

	Δείκτης DSN		
	μέση τιμή	τυπική απόκλιση	επίπεδο σημαντικότητας(α)
SMOKE=0	67,61	35,03	
SMOKE=1	58,76	36,09	0,08
OBESITY=0	59,55	36,96	
OBESITY=1	60,98	36,42	0,70
LIPIDS=0	56,62	37,8	
LIPIDS=1	60,49	35,8	0,8
LIPIDS=0	61,63	36,28	
LIPIDS=1	58,36	37,23	0,36
FAMILY=0	56,79	36,29	
FAMILY=1	66,10	36,87	0,013

Η ηλικία και ο δείκτης DSN δε σχετίζονται γραμμικά μεταξύ τους. Αυτό φαίνεται και από την τιμή του συντελεστή συσχέτισης Pearson (r=0,002), ο οποίος δεν είναι στατιστικά σημαντικός. Το διάγραμμα διασποράς του DSN σε σχέση με την ηλικία φαίνεται στο σχήμα 4.12.



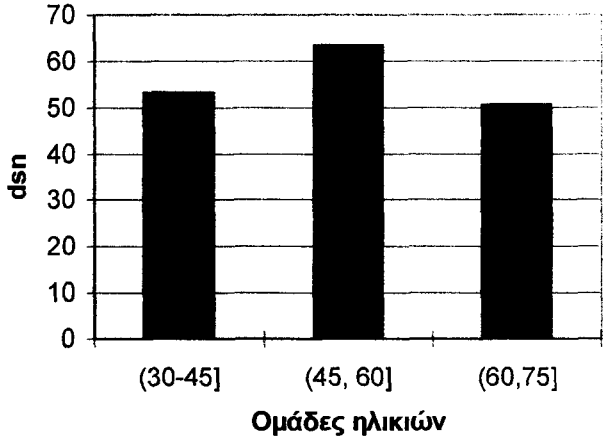
Σχ. 4.12

Αξίζει να σημειωθεί η σχέση των ομάδων ηλικιών και του δείκτη DSN. Από τον πίνακα (4.26) παρατηρείται ότι οι διασπορές στις τρεις ομάδες ηλικιών είναι ίσες, όπως προκύπτει από τον έλεγχο που έγινε με το χ^2 test. Δε συμβαίνει όμως το ίδιο με τους μέσους των τριών ομάδων, που όπως φαίνεται από τον πίνακα (4.26) δεν είναι ίσοι. Η ομάδα ηλικίας (45, 60] είναι αυτή που εμφανίζει τον υψηλότερο μέσο, ενώ η τρίτη ηλικία (60, 75] εμφανίζει το χαμηλότερο μέσο του δείκτη.

ΠΙΝΑΚΑΣ 4.26

Ηλικία	μέσος	τυπική απόκλιση	95% διάστημα εμπιστοσύνης
(30,45]	53,11	31,79	(44.97, 61.25)
(45,60]	63,42	37,08	(59.18, 67.67)
(60,75]	50,52	37,54	(40.74, 60.30)

Το ραβδόγραμμα των μέσων των ομάδων ηλικιών σε σχέση με το δείκτη DSN απεικονίζεται στο σχήμα 4.13.



Σχ. 4.13

Η ανάλυση διασποράς για τον έλεγχο της ισότητας των μέσων των τριών ομάδων ηλικιών συνοψίζεται στους πίνακες 4.27 και 4.28 που ακολουθούν.

ΠΙΝΑΚΑΣ 4.27: Έλεγχος ομοιογένειας των τριών διασπορών

		Βαθμοί ελευθερίας		
DSN	Levene Statistic	df_1	df_2	επίπεδο σημαντικότητας (α)
	1,01	2	413	0,362

ΠΙΝΑΚΑΣ 4.28: Ανάλυση διασποράς του DSN στις τρεις ομάδες ηλικιών.

Μεταβλητότητα	**Άθροισμα τετραγώνων**	**βαθμοί ελευθερίας**	**μέσο άθροισμα τετραγώνων**	**F**	**α**
Μεταξύ των ομάδων	11664,66	2	5832,33	4,394	0,013
Μέσα στις ομάδες	548133,80	413	1327,2		
Σύνολο	559798,4	415			

4.7 Ο δείκτης DSN ως μέσον πρόβλεψης

Ο δείκτης DSN θα μπορούσε να χρησιμεύσει και ως μέσο πρόβλεψης εμφάνισης του καρδιακού εμφράγματος (MI) και της παραπλεύρου κυκλοφορίας (CCC).

- Πρόβλεψη καρδιακού εμφράγματος

Με τη βοήθεια της λογιστικής ανάλυσης (Πίνακας 4.29) βρίσκουμε ότι το λογιστικό μοντέλο για την πρόβλεψη της μεταβλητής MI [Takayanagi (1990)] σε σχέση με το δείκτη DSN είναι το εξής:

$$\hat{p} = \frac{e^{-2,0029+0,0398dsn}}{1+e^{-2,0029+0,0398dsn}} \qquad (4.14)$$

ΠΙΝΑΚΑΣ 4.29

Μεταβλητές	b	τυπικό σφάλμα	Wald	βαθμοί ελευθερίας	α
DSN	0.0398	0.0045	77.25	1	0.000
Σταθερά	-2.0029	0.2592	59.71	1	0.000

Τα ποσοστά σωστής ταξινόμησης των ανδρών με καρδιακό έμφραγμα σύμφωνα με τις τιμές του δείκτη DSN παρουσιάζονται στον πίνακα 4.30.

ΠΙΝΑΚΑΣ 4.30

		Προβλεπόμενη ταξινόμηση		Ποσοστό σωστής ταξινόμησης
Παρατηρούμενη ταξινόμηση		0	1	
MI	0	**135**	52	72,19%
	1	62	**167**	72,93%
Ποσοστό σωστής ταξινόμησης		68,58%	76,25%	

Ειδικότητα=(135/187)x100=72,19%
Ευαισθησία=(167/229)x100=72,93%
Δείκτης Youden = (145,12 – 100)% = 45,12%

Η αποτελεσματικότητα του μοντέλου είναι 72,6%

Η ταξινόμηση έγινε χρησιμοποιώντας την κριτική τιμή 0,5 (δηλαδή την πιθανότητα να είναι κάποιος υγιής ή ασθενής.). Η αντίστοιχη κριτική τιμή του δείκτη DSN βρίσκεται αντικαθιστώντας στην (4.14) p=0,5, και λύνοντας ως προς DSN. Τελικά βρίσκουμε ότι η κριτική τιμή για την πρόβλεψη του MI είναι:

$$DSN= - \frac{-2,0029}{0,0398} = 50,32$$

Από τον Πίνακα 4.28 εξάγεται το συμπέρασμα ότι για DSN ≤ 50,3 το MI είναι 0 με πιθανότητα 68,58% ενώ όταν DSN > 50 το MI είναι 1 με πιθανότητα 76,25%.

Η αντίστοιχη κριτική τιμή του δείκτη Gensini βρίσκεται με τη χρήση του τύπου (4.13). Έτσι

$$50,32 = 3,181 + 1,055 \cdot \text{Gensini},$$

από την οποία βρίσκουμε Gensini = 44,68.

Το λογιστικό μοντέλο για την πρόβλεψη του εμφράγματος με χρήση των τιμών Gensini έχει ικανότητα σωστής συνολικής ταξινόμησης 72,84% έναντι του ποσοστού του DSN 72,60%. Η διαφορά μεταξύ των δύο ποσοστών δε θεωρείται στατιστικά σημαντική.

- Πρόβλεψη εμφάνισης παράπλευρης κυκλοφορίας (CCC).

Με τη βοήθεια της λογιστικής ανάλυσης βρίσκουμε ότι το λογιστικό μοντέλο για την πρόβλεψη της μεταβλητής CCC(W.R Webb et al (1973)) είναι το εξής:

ΠΙΝΑΚΑΣ 4.31

Μεταβλητές	b	τυπικό σφάλμα	Wald	Βαθμοί ελευθερίας	α
DSN	0.0224	0.0034	43.12	1	0.000
Σταθερά	-1.181	0.2157	29.97	1	0.000

Ο τύπος είναι: $$\hat{p} = \frac{e^{-1,181+0,0224dsn}}{1+e^{-1,181+0,0224dsn}}$$

Τα ποσοστά πρόβλεψης δίδονται από τον πίνακα 4.32 που ακολουθεί:

ΠΙΝΑΚΑΣ 4.32

		Προβλεπόμενη ταξινόμηση		Ποσοστό σωστής ταξινόμησης
Παρατηρούμενη ταξινόμηση		0	1	
CCC	0	**132**	64	67,35%
	1	76	**144**	65,45%
Ποσοστό σωστής ταξινόμησης		63,46%	69,23%	

Ειδικότητα=(132/196)x100=67,35%
Ευαισθησία=(144/220)x100=65,45%

Δείκτης Youden = (132,80 – 100)% = 32,80%
Συνολικό ποσοστό σωστής ταξινόμησης=66,35%

Η κριτική τιμή του δείκτη DSN για την πρόβλεψη του CCC είναι:

$$DSN= -\frac{-1,1810}{0,0224} = 52,72.$$

Επομένως για DSN≤ 52,7 η CCC παίρνει την τιμή 0 με πιθανότητα 63,46% ενώ όταν DSN > 52,7 η CCC παίρνει την τιμή 1 με πιθανότητα 69,23%.

Η αντίστοιχη κριτική τιμή του Gensini βρίσκεται πάλι από τη σχέση (4.13) και είναι Gensini=46,95.

Το λογιστικό μοντέλο για την εμφάνιση παράπλευρης κυκλοφορίας με χρήση των τιμών Gensini εμφανίζει μειωμένο ποσοστό σωστής συνολικής ταξινόμησης 63,94% έναντι του ποσοστού του DSN 66,35% .

ΚΕΦΑΛΑΙΟ 5

ΣΥΜΠΕΡΑΣΜΑΤΑ

5.1. Γενικές παρατηρήσεις

Οι κύριοι στόχοι της έρευνας αυτής ήταν οι εξής:

► Ο προσδιορισμός των καταλληλότερων διαχωριστικών τεχνικών ή ΤΝΔ που να διαχωρίζουν με τη μεγαλύτερη δυνατή ακρίβεια έναν πληθυσμό ατόμων που εμφάνισαν συμπτώματα καρδιοπάθειας σε υγιείς και ασθενείς, χωρίς τη βοήθεια επεμβατικών μεθόδων. Για το σκοπό αυτό χρησιμοποιήθηκε δείγμα 660 ατόμων από το Ιπποκράτειο Γενικό Νοσοκομείο Αθηνών και

► η αξιολόγηση της σοβαρότητας της νόσου σε άτομα που έχουν ήδη υποστεί στεφανιογραφία.

Πριν προχωρήσουμε σε συνοπτική παρουσίαση των ευρημάτων της έρευνας, αξίζει να σημειωθούν τα εξής:

► Το δείγμα που χρησιμοποιήθηκε περιείχε άτομα που στην πλειοψηφία τους ήταν κάτοικοι του Λεκανοπεδίου Αττικής και είχαν χαρακτηριστεί από τους γιατρούς ως ασθενείς (≈ 85%).

► Το ερωτηματολόγιο που συμπληρώθηκε από τους θεράποντες ιατρούς περιείχε στοιχεία βασισμένα στην ιατρική γνώση των αρχών της δεκαετίας του 1990. Από τότε έχουν εμφανιστεί στη βιβλιογραφία [Tuomi (1994), Zodpey *etal* (1994), Townsend (1995), Rosenberg(1995), Kimmel *etal* (1995), Rees (1996), Denollet *etal* (1996), Godsland *etal* (1998), Menotti *etal* (1999), Poppius *etal* (1999)] μια σειρά από παράγοντες κινδύνου που συμπληρώνουν τη γνώση μας, χωρίς όμως να αναιρούν τη συμβολή των μέχρι τότε γνωστών παραγόντων κινδύνου στην εμφάνιση της νόσου.

► Μερικοί παράγοντες κινδύνου έχουν αναλυθεί σε επιμέρους χαρακτηριστικά, όπως για παράδειγμα ο τύπος Α (TYPE A). Ο συγκεκριμένος παράγοντας κινδύνου έχει αντικατασταθεί από άλλους, Hostility, Anger, Aggressiveness [Mendes (1988), Bernardo (1993), Hayano *etal* (1997)]. Αντίστοιχη ανάλυση έχει αναφερθεί και για τον παράγοντα FAMILY [Silberberg *etal* (1999)].

Είναι σαφές ότι οι έρευνες αυτές αναγνωρίζουν την επίδραση των συγκεκριμένων παραγόντων και προσπαθούν να μελετήσουν τη συμβολή των επιμέρους χαρακτηριστικών τους στην εμφάνιση της νόσου.

Μια πρώτη ανάλυση στο συνολικό δείγμα αποκαλύπτει ότι η κρίσιμη ηλικία για την εμφάνιση της νόσου είναι μεταξύ 45 και 60 έτη. Η αρτηρία όπου υπάρχουν τα περισσότερα εμφράγματα είναι η LAD στη θέση PROXIMAL, ενώ δεν προκαλεί εντύπωση το γεγονός ότι η μεγάλη πλειοψηφία όσων προσήλθαν για εξέταση στο Ιπποκράτειο ήταν καπνιστές. Αξίζει να σημειωθεί ότι στο συγκεκριμένο δείγμα οι άνδρες είναι 11,962 φορές πιθανότεροι στεφανιαίοι ασθενείς από ότι οι γυναίκες.

Με βάση τα στοιχεία του συνολικού δείγματος η εκδήλωση της νόσου [μεταβλητή DCODE] εμφανίζεται να εξαρτάται από τους παράγοντες :

- *Υπερλιπιδαιμία* (LIPIDS)
- Κάπνισμα (SMOKE)
- Ιστορικό (FAMILY)
- Φύλο (SEX)
- Τρόπος ζωής (SEDENTARY και TYPE A)

Είναι αυτονόητο ότι η εκδήλωση και η έκταση της νόσου εξαρτάται, εκτός από τους παράγοντες που προαναφέρθηκαν, από το κλάσμα εξώθησης (EF) καθώς και από τον αριθμό φραγμένων αρτηριών (VD).

Ο αριθμός των φραγμένων αρτηριών (VD) εξαρτάται, σύμφωνα με τα στοιχεία του συνολικού δείγματος, από το κάπνισμα, το φύλο και το διαβήτη.

Το κλάσμα εξώθησης επηρεάζεται, σύμφωνα με τα στοιχεία του συνολικού δείγματος από το κάπνισμα και το διαβήτη.

Όσον αφορά τις αλληλεξαρτήσεις μεταξύ των μεταβλητών, μπορούν να παρατηρηθούν τα ακόλουθα:

Στο συνολικό δείγμα, κάθε μία από τις μεταβλητές AGE, TYPE A, HBP (υπέρταση) εμφανίζεται συσχετισμένη με έξι άλλες, όπως φαίνεται στον πίνακα 1.7 του αντίστοιχου κεφαλαίου, ενώ οι μεταβλητές SMOKE και SEDENTARY συσχετίζονται με πέντε άλλες μεταβλητές. Δεν είναι τυχαίο ότι τέσσερις από αυτές περιγράφουν κυρίως τον τρόπο ζωής του ατόμου και όπως έχει ήδη αναφερθεί, παίζουν σημαντικό ρόλο στην εκδήλωση της νόσου.

Η εξάρτηση της εκδήλωσης της νόσου από τους παράγοντες κινδύνου, διαφοροποιείται όταν εξετάζονται χωριστά τα αρχεία ανδρών και γυναικών.

Έτσι στο συγκεκριμένο δείγμα, στους άνδρες η εκδήλωση της νόσου φαίνεται να εξαρτάται αποκλειστικά από το κάπνισμα, ενώ στις γυναίκες από το κάπνισμα, την υπερλιπιδαιμία, το διαβήτη, το ιστορικό και τον τρόπο ζωής.

Σημειώνεται εδώ η εμφάνιση του διαβήτη που επηρεάζει την εμφάνιση της νόσου στις γυναίκες, και ο οποίος δεν εμφανίζεται στους καθοριστικούς για τη νόσο παράγοντες όταν μελετάται στο συνολικό δείγμα. Παρόμοια ευρήματα αναφέρονται στην εργασία των Kyriakidis etal (1995).

Στο σύνολο των ανδρών το κλάσμα εξώθησης εξαρτάται από το διαβήτη, το οικογενειακό ιστορικό και τον χαρακτήρα (TYPE A), ενώ στις γυναίκες από το κάπνισμα και το διαβήτη. Αξίζει να τονιστεί το γεγονός ότι, στο συγκεκριμένο δείγμα, όσοι άνδρες είχαν κλάσμα εξώθησης μικρότερο του 0.4 ήταν ασθενείς. Δεν παρατηρήθηκε αντίστοιχο ποσοστό στις γυναίκες.

Υπάρχει έντονη διαφοροποίηση ανάμεσα στα δύο φύλα σε ότι αφορά την ηλικία όπου εκδηλώθηκε η νόσος. Οι περισσότεροι άνδρες ασθενείς βρίσκονται ηλικιακά στην περιοχή (45-60] ετών, ενώ η συντριπτική πλειοψηφία των γυναικών στην περιοχή (45-75] χρόνων. Είναι σημαντικό το ότι ο αριθμός των ασθενών ανδρών μειώνεται σημαντικά μετά το 60^{o} έτος, ενώ ο αντίστοιχος αριθμός για τις γυναίκες παραμένει ο ίδιος με αυτόν των ηλικιών (45-60]. Ο αριθμός των φραγμένων αρτηριών στους μεν άνδρες δεν παρουσιάζει εξάρτηση από κάποιο παράγοντα, ενώ στις γυναίκες εξαρτάται από το κάπνισμα, το διαβήτη και την υπερλιπιδαιμία.

- Στο δείγμα των ανδρών η μεταβλητή SMOKE εμφανίζει συσχέτιση με πέντε άλλες μεταβλητές (πίνακας1.11). Το γεγονός αυτό επιβεβαιώνει αυτό που όλοι ξέρουμε για τον επιβλαβή ρόλο του καπνίσματος στην υγεία του ατόμου. Οι μεταβλητές AGE, SEDENTARY, DIABETES και LIPIDS παρουσιάζουν συσχετίσεις με τρεις μόνο μεταβλητές η καθεμιά.
- Στο δείγμα των γυναικών η εικόνα είναι τελείως διαφορετική. Οι μεταβλητές με τις περισσότερες αλληλεξαρτήσεις είναι η TYPE A (με πέντε) και η LIPIDS (με τρεις). Αντιθέτως, οι μεταβλητές SMOKE και SEDENTARY εμφανίζουν αλληλεξάρτηση μόνον με τη μεταβλητή TYPE A. (πίνακας 1.18)

Οι παραπάνω διαπιστώσεις καθιστούν σαφές ότι διαφορετικοί παράγοντες επηρεάζουν την εκδήλωση της νόσου στα δύο φύλα, γεγονός που επιβεβαιώνεται και από την Ιατρική Επιστήμη.

5.2. Διαχωριστικές τεχνικές-Τεχνητά νευρωνικά δίκτυα

Η συμφωνία των ευρημάτων που προέκυψαν από την στατιστική ανάλυση των συσχετίσεων με την μέχρι σήμερα ιατρική γνώση, οδήγησε στην περαιτέρω επεξεργασία του δείγματος με τη βοήθεια της παραγοντικής ανάλυσης και των διαχωριστικών τεχνικών που εφαρμόστηκαν πρώτα σε πιλοτικό δείγμα αποτελούμενο από 160 άτομα – 80 ασθενή και 80 υγιή. Τα συμπεράσματα συνοψίζονται ως εξής:

- Κατά την παραγοντική ανάλυση, οι μεταβλητές χωρίστηκαν σε 5 κατηγορίες που περιγράφουν αντίστοιχα τον τρόπο ζωής του ατόμου (AGE, TYPE A, SEDENTARY, SMOKE, SEX), την προδιάθεση (DIABETES, EF), την κατάσταση υγείας (DIABETES, LIPIDS, HBP), το ιστορικό (FAMILY) και την παχυσαρκία (OBESITY). Οι πέντε αυτές ομάδες παραγόντων ερμηνεύουν το 66% του συνόλου της διακύμανσης με λ>1. Αξίζει να τονιστεί το γεγονός ότι η εκδήλωση της νόσου εξαρτάται από 4 μεταβλητές της πρώτης ομάδας (TYPE A, SEDENTARY, SMOKE, SEX), 1 της τρίτης (LIPIDS) και αυτόν της τέταρτης (FAMILY).
- Η διαχωριστική ανάλυση κατέληξε στην εύρεση 3 διαχωριστικών συναρτήσεων (2.13, 2.14, 2.15) με 11, 7 και 4 μεταβλητές αντίστοιχα. Η αποτελεσματικότητα στο διαχωρισμό του πιλοτικού δείγματος σε ασθενείς και υγιείς ήταν για τις δύο πρώτες συναρτήσεις 82.5% με δείκτη Youden 65%, και 83.1% με δείκτη Youden 66.3% για την τρίτη. Επειδή οι δύο πρώτες συναρτήσεις δίνουν τα ίδια περίπου αποτελέσματα, η σύγκριση για το συνολικό δείγμα έγινε μεταξύ της δεύτερης και της τρίτης. Η αποτελεσματικότητα και ο δείκτης Youden ήταν 80.4%, 69.1% και 80.9%, 67.5% αντίστοιχα.

Αν και η αποτελεσματικότητα της διαχωριστικής συνάρτησης (2.15) στο συνολικό δείγμα εμφανίζεται ελαφρώς αυξημένη, η διαφορά αυτή δε θεωρείται στατιστικώς σημαντική (z-test=0,23 έναντι z*=1,65 για επίπεδο σημαντικότητας α=0,05), ενώ η αυξημένη τιμή του δείκτη Youden της διαχωριστικής συνάρτησης 2.14 την καθιστά καταλληλότερη .

- Η όλη διαδικασία ανέδειξε την μεταβλητή EF σαν ισχυρό διαχωριστικό παράγοντα.
- Εφαρμόστηκαν 4 μοντέλα λογιστικής παλινδρόμησης στο πιλοτικό δείγμα.

- Η διαδικασία αυτή ανέδειξε ως καλύτερο το μοντέλο (2.24) που χρησιμοποιεί 1 μεταβλητή και 8 αλληλεπιδράσεις (αποτελεσματικότητα 90%, δείκτης Youden 80%) και πληρεί το κριτήριο της οικονομικότητας

Από τα παραπάνω είναι προφανές ότι το μοντέλο (2.24) της λογιστικής παλινδρόμησης υπερτερεί ως προς τα μοντέλα διαχωριστικής ανάλυσης που εφαρμόστηκαν στο πιλοτικό δείγμα, σε όλα τα κριτήρια αξιολόγησης.

Η ικανότητα σωστής ταξινόμησης των μοντέλων αυτών σε άγνωστα δεδομένα, θα προκύψει με την εφαρμογή τους στο συνολικό δείγμα των 660 ατόμων (Πίνακας Σ1). Εδώ πρέπει να τονιστεί ότι έχουν αναφερθεί [Morise AP (1996)] στατιστικώς σημαντικά σφάλματα σε περιπτώσεις που τα λογιστικά μοντέλα εφαρμόζονται σε δείγματα με διαφορετικά ιδιαίτερα χαρακτηριστικά από αυτά που χρησιμοποιήθηκαν για τον προσδιορισμό των μοντέλων.

ΠΙΝΑΚΑΣ Σ1: Αξιολόγηση των μοντέλων ως προς τη διαχωριστική τους ικανότητα στο συνολικό δείγμα

	Διαχωριστική ανάλυση		**Λογιστική παλινδρόμηση**
Συναρτήσεις	2.14	2.15	2.24
Αριθμός μεταβλητών	7	4	9
Ειδικότητα	**90%**	87,5%	88,8%
Ευαισθησία	79,1%	80%	**84,3%**
Αποτελεσματικότητα	80,4%	80,9%	**84,9%**
Δείκτης Youden	69,1%	67,5%	**73,1%**

Παρατηρείται ότι η αποτελεσματικότητα των μοντέλων στο συνολικό δείγμα εμφανίζεται αυξημένη στο λογιστικό μοντέλο. Με το δεδομένο ότι η ειδικότητα εκφράζει το ποσοστό των υγιών που έχουν ταξινομηθεί σωστά και η ευαισθησία το ποσοστό των σωστά ταξινομημένων ασθενών, ο δείκτης Youden άμεσα εξαρτώμενος από τα προαναφερθέντα μεγέθη, αποτελεί ένα αξιόπιστο κριτήριο καλής ταξινόμησης στην ιατρική στατιστική.

Άρα η επιλογή του διαχωριστικού μοντέλου θα γίνει με κριτήριο του δείκτη Youden.

Στον πίνακα αξιολόγησης (Σ1) το μοντέλο 2.24 το οποίο και προτείνεται, έχει το μεγαλύτερο δείκτη Youden, τη μεγαλύτερη αποτελεσματικότητα και σχετικά μικρό αριθμό μεταβλητών.

Το μοντέλο λογιστικής παλινδρόμησης (2.24) είναι το ακόλουθο:

$$\ln(p/1-p) = 23{,}020 - 32{,}738.X_1 + 3{,}143.INT_1 - 4{,}682.INT_2 + 3{,}522.INT_3 - 4{,}389.INT_4 - 2{,}151.INT_5 - 1{,}934.INT_6 - 5{,}060.INT_7 + 4{,}188.INT_8$$

οι μεταβλητές που χρησιμοποιήθηκαν είναι :

X_1 = EF
INT_1= FAMILY*LIPIDS
INT_2=LIPIDS* SEDENTARY
INT_3=DIABETES*HBP
INT_4=HBP*TYPE A
INT_5=SEDENTARY*SMOKE
INT_6=SEDENTARY*DIABETES
INT_7=FAMILY*OBESITY
INT_8=OBESITY *SEDENTARY

Στη συνέχεια χρησιμοποιήθηκαν τεχνητά νευρωνικά δίκτυα για το διαχωρισμό του συνολικού δείγματος με σκοπό τη σύγκριση των διαχωριστικών τεχνικών με τα ΤΝΔ. Τα τεχνητά νευρωνικά δίκτυα διαφέρουν από τις κλασσικές διαχωριστικές μεθόδους της πολυμεταβλητής ανάλυσης γιατί κυρίως χρησιμοποιούν παραδείγματα παρά κανόνες για να αναγνωρίσουν patterns που μπορεί να είναι αόριστα ή ελαφρά αντιφατικά. Αν και τα τελευταία χρόνια η εισαγωγή των ΤΝΔ έχει λύσει πολλά προβλήματα στον επιστημονικό, επιχειρηματικό και βιομηχανικό χώρο και έχουν γίνει αντικείμενο μελέτης σε πολλούς ερευνητικούς τομείς, στην Ελλάδα τα ΤΝΔ δεν χρησιμοποιούνται ακόμα στην ιατρική στατιστική .

Κατασκευάστηκαν και εφαρμόστηκαν δέκα ΤΝΔ στο πιλοτικό δείγμα το οποίο χρησιμοποιήθηκε ως δείγμα εκπαίδευσης, εκ των οποίων αναδείχθηκαν τρία. Τα δίκτυα αυτά αξιολογήθηκαν ως προς την ικανότητα διαχωρισμού του πιλοτικού δείγματος και έγινε εκτίμηση τάξης υποδείγματος.

Η ίδια αξιολόγηση έγινε και για τα μοντέλα 2.15 της διαχωριστικής ανάλυσης και 2.24 της λογιστικής παλινδρόμησης τα οποία εμφάνιζαν την μεγαλύτερη αποτελεσματικότητα και τον υψηλότερο δείκτη Youden στο πιλοτικό δείγμα.

Ο πίνακας Σ.2 παρουσιάζει τη σύγκριση των ΤΝΔ και των διαχωριστικών μοντέλων.

ΠΙΝΑΚΑΣ Σ.2

	ΤΝΔ			Δ ιχωριστική ανάλυση	Λογιστική παλινδρόμηση
Υπόδειγμα	11x11x1	7x7x1	11x11x1	2.15	2.24
Συναρτήσεις	Σιγμοειδής		Κατωφλίου	Δ ιχωριστική	Λογιστική
Παράμετροι	132	56	132	4	9
Ειδικότητα (%)	**92,5**	88,8	90,0	87,5	90
Ευαισθησία (%)	90,0	87,5	**91,3**	78,8	90
Youden (%)	**82,5**	81,3	81,3	66,3	80
Αποτελεσματικότητα(%)	**91,2**	88,1	90,6	83,1	90
RMSE	**0,2620**	0,2997	0,2867	0,418	0,314
MAE	0,123	0,160	**0,110**	0,175	0,129
RSQ	**0.731**	0,643	0,685	0,427	0,573

Όπως φαίνεται στον πίνακα, Σ.2 τα ΤΝΔ έχουν καλύτερη απόδοση έναντι των διαχωριστικών τεχνικών στο πιλοτικό δείγμα. Το ΤΝΔ (11x11x1 - υπόδειγμα 1), προτείνεται ως το καλύτερο και ως προς την ικανότητα σωστής ταξινόμησης των δεδομένων εκπαίδευσης αλλά και ως προς την εκτίμηση τάξης υποδείγματος.

Η ικανότητα γενίκευσης των ΤΝΔ και των διαχωριστικών μοντέλων που προτείνονται, ελέγχθηκε στο συνολικό αρχείο των 660 ατόμων. Τα αποτελέσματα του ελέγχου παρουσιάζονται στον Πίνακα Σ.3 από όπου συμπεραίνουμε ότι τα υποδείγματα 11x11x1 και 7x7x1 των ΤΝΔ έχουν μεγαλύτερη ικανότητα γενίκευσης.

ΠΙΝΑΚΑΣ Σ.3 Ικανότητα γενίκευσης

	ΤΝ ι			Δ αχωριστική ανάλυση	Λογιστική ι αλινδρόμηση
Υπόδειγμα	11x11x1	7x7x1	11x11x1	2.15	2.24
Αποτελεσματικότητα(%)	86,1	86,97	85	80,9	84,9

Η ικανότητα γενίκευσης συνδυαζόμενη με την απόδοση στα δεδομένα εκπαίδευσης (πιλοτικό δείγμα) και την εκτίμηση τάξης υποδείγματος αναδεικνύει το ΤΝΔ (11x11x1) ως το καλύτερο μέσον πρόβλεψης στεφανιαίας νόσου. Το ΤΝΔ (11x11x1) έχει ένα κρυφό στρώμα 11 κόμβων, m+1=12 κόμβους στο επίπεδο εισόδου (μαζί με το σταθερό όρο) και κατά συνέπεια 132 προς εκτίμηση παραμέτρους, έχει εκτελέσει μικρό αριθμό κύκλων εκπαίδευσης (runs=200) για να φθάσει στην επιθυμητή απόδοση και, όπως αναφέρεται στον Πίνακα 3.3, χρησιμοποιεί τη σιγμοειδή συνάρτηση και έχει ικανοποιητικό RMSE.

Συγκριτικές έρευνες για την απόδοση των μαθηματικών μοντέλων πρόγνωσης [Selker *et.al* (1995)], κατέληξαν στο συμπέρασμα ότι, ανεξαρτήτως του αριθμού των μεταβλητών που χρησιμοποιούνται, τα ΤΝΔ έχουν καλύτερη διαχωριστική ικανότητα σε σχέση με τη λογιστική παλινδρόμηση και τα δένδρα ταξινόμησης (classification trees), υστερούν όμως ως προς τη διαφορά μεταξύ της αναμενόμενης πιθανότητας και αυτής που εξάγεται από το δείγμα (calibration). Η επιλογή μεταξύ των διαφόρων μεθόδων φαίνεται να βασίζεται περισσότερο στις ανάγκες της συγκεκριμένης εφαρμογής, παρά στο κατά πόσο κάποια μέθοδος είναι εν δυνάμει ισχυρότερη των άλλων.

5.3. Δείκτης Gensini

Ο δείκτης Gensini είναι ένας από τους σπουδαιότερους δείκτες στην καρδιολογία μέσω του οποίου αξιολογείται και ποσοτικοποιείται η βαρύτητα της νόσου. Ο υπολογισμός του βασίζεται σε δύο παράγοντες: α) στο βαθμό κρισιμότητας ο οποίος αντιστοιχεί στο ποσοστό έμφραξης (y) της αρτηρίας και β) στο συντελεστή w_{ij} i=1,...,10 , j=1,2,3, που η τιμή του εξαρτάται από τη θέση i της στένωσης καθώς και την αρτηρία j (πίνακας 4.2). Οι βαθμοί κρισιμότητας σύμφωνα με τον πίνακα 4.1 αντιστοιχούν μόνο σε έξι συγκεκριμένα ποσοστά έμφραξης (y) και κατά συνέπεια για τιμές έμφραξης ενδιάμεσες αυτών που αναφέρονται στον πίνακα, η αντίστοιχη τιμή κρισιμότητας δίδεται εμπειρικά.

Στην παρούσα εργασία προτείνεται ένας εναλλακτικός δείκτης (DSN) ο οποίος υπολογίζεται από μια μαθηματική συνάρτηση ds (i,y,j), στην οποία το ποσοστό έμφραξης (y) μπορεί να κυμανθεί στο διάστημα [0,25 ,0,99] και οι προβλεπόμενες τιμές του δείκτη Gensini να αποδίδονται με μεγάλη ακρίβεια.

Για τιμή έμφραξης y =1(100%) διατηρείται η τιμή του δείκτη Gensini, λόγω της σημαντικής απόκλισης που εμφανίζει το μοντέλο σε σχέση με την τιμή του δείκτη. Επειδή ο δείκτης DSN πρέπει να λαμβάνει υπόψη το βαθμό κρισιμότητας αλλά και το συντελεστή w_{ij}, όπως αυτός ορίζεται από τον Gensini, για τον προσδιορισμό της συνάρτησης χρησιμοποιήθηκαν για μεν τη μοντελοποίηση του βαθμού κρισιμότητας μη γραμμικό στατιστικό μοντέλο, για δε τον προσδιορισμό του συντελεστή w_{ij}, πολυώνυμο παρεμβολής Lagrange.

Το μοντέλο το οποίο θεωρείται το πλέον κατάλληλο για τον υπολογισμό του βαθμού κρισιμότητας είναι :

$$g(y)=\mathrm{Exp}(\mathrm{Exp}(-1{,}77405+2{,}8052y))$$

Ο λόγος του διπλού εκθετικού είναι ο μεγάλος ρυθμός μεταβολής του βαθμού σοβαρότητας σε σχέση με το ποσοστό έμφραξης. Η μεταβλητή είναι στατιστικώς σημαντική και το μοντέλο ερμηνεύει το 99,97% της μεταβλητότητας της εξαρτημένης μεταβλητής όταν $0{,}25 < y \leq 0{,}99$

Το πολυώνυμο παρεμβολής Lagrange το οποίο υπολογίζει τον συντελεστή w_{ij} είναι το ακόλουθο:

$$f(i,j)=\left(-106{,}393+199{,}009i-122{,}417i^2+31{,}1429i^3-3{,}48535i^4+0{,}142857i^5\right)$$
$$+\left(142{,}589-264{,}89i+163{,}107i^2-41{,}5126i^3-4.64678i^4-0{,}190476i^5\right)j$$
$$+\left(-35{,}1964+65{,}8803i-40{,}6897i^2+10{,}3697i^3-1{,}16143i^4+0{,}047619i^5\right)j^2\Big)$$

Σύμφωνα με τα παραπάνω ο δείκτης ds(i,y,j) που προτείνεται για μια έμφραξη ορίζεται ως εξής :

$$ds_j(i,y)=\begin{cases} f(i,j)\cdot g(y), 1\le i\le 10, 0{,}25\le y\le 0{,}99,\ 1\le j\le 3 \\ 32\cdot f(i,j), 1\le i\le 10,\ y=1,\ 1\le j\le 3 \end{cases}$$

Ο τελικός δείκτης DSN θα προκύψει ως άθροισμα των δεικτών ds_1, ds_2, ds_3 που αντιστοιχούν στις τρεις αρτηρίες (RCA, LAD, LCX).

Στη συνέχεια μελετήθηκε η συσχέτιση των παραγόντων κινδύνου στο σύνολο των ασθενών. Από τη μελέτη προέκυψε ότι οι μεταβλητές που εμφανίζονται στατιστικώς εξαρτημένες είναι οι ίδιες όπως και στο συνολικό αρχείο. Πρέπει επίσης να προστεθεί η συσχέτιση της μεταβλητής Sedentary με τα λιπίδια (LIPIDS) και τη παχυσαρκία (OBESITY), της παχυσαρκίας με το κάπνισμα (SMOKE), όπως επίσης και του διαβήτη (DIABETES) με τις ομάδες ηλικών (AG).

Αξίζει να σημειωθεί ότι παρά το γεγονός ότι ο διαβήτης δεν αξιολογείται ως καθοριστικός παράγοντας για την εκδήλωση της νόσου στο συνολικό αρχείο, στο αρχείο ασθενών φαίνεται να δρα ως επιβαρυντικός παράγοντας στην κρισιμότητα της νόσου, διότι εμφανίζει εξάρτηση με τον αριθμό των φραγμένων αρτηριών και με την τιμή του κλάσματος εξώθησης. Ιδιαίτερα στο αρχείο ασθενών γυναικών το ποσοστό των διαβητικών γυναικών με νόσο δύο αγγείων εμφανίζεται αυξημένο. Μπορούμε επίσης να ισχυρισθούμε ότι η υπέρταση επιβαρύνει την κρισιμότητα της νόσου στις γυναίκες ασθενείς δοθέντος ότι το ποσοστό των υπερτασικών ασθενών γυναικών εμφανίζεται αυξημένο έναντι του αντίστοιχου των ανδρών.

Στο αρχείο ασθενών ανδρών, όπου ορίσθηκαν τέσσερις νέες μεταβλητές όπως ακινησία, ανεύρυσμα, παράπλευρη κυκλοφορία και έμφραγμα, πρέπει να σημειωθεί ότι στεφανιαίοι άνδρες ασθενείς υπερλιπιδαιμικοί, καπνιστές, διαβητικοί και τύπου προσωπικότητας Α έχουν αυξημένη πιθανότητα να εμφανίσουν παράπλευρη κυκλοφορία ενώ οι καπνιστές, τύπου Α και οι έχοντες

οικογενειακό ιστορικό έχουν κίνδυνο εμφράγματος. Η καθιστική ζωή επίσης ενισχύει την εμφάνιση ανευρύσματος.

Για τον προσδιορισμό των στατιστικών του δείκτη DSN υπολογίστηκαν οι τιμές του στο αρχείο ασθενών ανδρών όπου αντίστοιχα υπήρχε η καταγραφή των τιμών του δείκτη Gensini στα ιατρικά δελτία. Παρατηρήθηκε ότι οι δύο δείκτες σχετίζονται γραμμικά με δείκτη προσδιορισμού R^2 =0,977 και η ευθεία παλινδρόμησης περιγράφεται από την σχέση:

$$DSN = 3,182 + 1,055 \cdot (\text{Gensini index})$$

Θα μπορούσαμε επίσης να ισχυρισθούμε ότι ο δείκτης DSN ακολουθεί την lognormal κατανομή σε επίπεδο σημαντικότητας α = 0.01.

Ερευνήθηκε επίσης η σχέση του DSN με τους παράγοντες κινδύνου και τις τέσσερις νέες μεταβλητές στο αρχείο ασθενών ανδρών. Η ηλικία και ο δείκτης *DSN δε σχετίζονται γραμμικά μεταξύ τους. Για τους δίτιμους παράγοντες κι*νδύνου κάπνισμα, παχυσαρκία, καθιστική ζωή, υπερλιπιδαιμία που δρουν ως επιβαρυντικοί στο αρχείο ασθενών ανδρών η μέση τιμή του δείκτη DSN είναι η ίδια στις ομάδες 0 και 1 με εξαίρεση τη μεταβλητή οικογενειακό ιστορικό για την οποία οι δύο ομάδες έχουν διαφορετικούς μέσους. Επιβεβαιώνεται επίσης *από την υψηλή μέση τιμή του δείκτη* DSN *ότι η ομάδα ηλικίας* (45,60] *των ανδρών είναι υψηλού κινδύνου για στεφανιαίους ασθενείς.*

Ο δείκτης DSN θα μπορούσε επίσης να χρησιμεύσει ως μέσον πρόβλεψης. Με τη βοήθεια της λογιστικής παλινδρόμησης εξάγεται το συμπέρασμα ότι για τιμή του DSN ≤ 50,3 η τιμή της μεταβλητής MI (έμφραγμα) είναι μηδέν με πιθανότητα 0,6858 δηλ. 68,58%, ενώ όταν DSN>50 ένας στεφανιαίος ασθενής έχει πιθανότητα να εμφανίσει έμφραγμα (MI = 1) 0,7625 δηλ. 76,25%. Η αντίστοιχη κριτική τιμή του δείκτη Gensini για το συγκεκριμένο δείγμα υπολογίστηκε 44,68. Ομοίως υπολογίστηκε η κριτική τιμή του δείκτη DSN για την πρόβλεψη της παράπλευρης κυκλοφορίας. Επομένως για DSN ≤ 52,7 η CCC (παράπλευρη κυκλοφορία) παίρνει την τιμή 0 με πιθανότητα 63,46% ενώ όταν DSN > 52,7 η CCC παίρνει την τιμή 1 με πιθανότητα 69,23%.

5.4. Προτάσεις για περαιτέρω έρευνα

Κατά τη διάρκεια της έρευνας έγινε σαφές ότι ένας από τους βασικούς παράγοντες που επηρεάζουν την αξιοπιστία των αποτελεσμάτων είναι και το ερωτηματολόγιο το οποίο συμπληρώνεται από το θεράποντα ιατρό. Το ερωτηματολόγιο αυτό περιέχει στοιχεία που προέρχονται από εργαστηριακές αναλύσεις, από πιθανή στεφανιογραφία, αλλά και από ερωτήσεις που υποβάλλονται στον ασθενή και περιέχουν υποκειμενικού χαρακτήρα απαντήσεις και κρίσεις του ιατρού. Όλα τα παραπάνω στοιχεία περικλείουν αβεβαιότητα, την οποία ο συντάκτης του ερωτηματολογίου πρέπει να ελαχιστοποιήσει, ώστε τα στοιχεία που χρησιμοποιούνται στη στατιστική ανάλυση να είναι όσο το δυνατόν ακριβέστερα. Έρευνες έδειξαν ότι στα ιστορικά που ελήφθησαν από ασθενείς, ορισμένοι δείκτες εμφανίζονται υπερεκτιμημένοι σε σχέση με τα αποτελέσματα μετέπειτα στεφανιογραφίας [Derby (1990)]. Το γεγονός αυτό, που εμφανίζεται ιδιαίτερα σε άτομα που γνωρίζουν ότι έχουν συμπτώματα καρδιοπάθειας, δημιουργεί φαινόμενα μεροληψίας (bias), που μπορεί να επηρεάσουν τις μελέτες που βασίζονται σε παρόμοια ερωτηματολόγια. Επιπλέον, η ύπαρξη κατηγορικών μεταβλητών στη θέση συνεχών (OBESITY 0-1 αντί BODY INDEX, SMOKE 0-1 αντί συνολικού αριθμού τσιγάρων [Wang (1994)], LIPIDS 0-1 αντί της ακριβούς μέτρησης, DIABETES 0-1 αντί της αντίστοιχης τιμής κλπ) πιθανόν να στερούν από τη στατιστική ανάλυση τη δυνατότητα εξαγωγής ακριβέστερων συμπερασμάτων. Είναι λοπόν απαραίτητο τα ερωτηματολόγια να παρέχουν τις ακριβέστερες, κατά το δυνατόν, πληροφορίες που θα βοηθήσουν το έργο της μετέπειτα ανάλυσης.

Πρέπει επίσης τα ερωτηματολόγια που δίνονται για ανάλυση να περιέχουν στοιχεία για όλους τους μέχρι σήμερα γνωστούς παράγοντες κινδύνου και να αντιπροσωπεύουν μεγάλο δείγμα από διάφορες περιοχές της χώρας. Χαρακτηριστικά αναφέρεται ότι η θνησιμότητα από καρδιοπάθεια συνδέεται με τις διατροφικές συνήθειες και κυμαίνεται από 25 ανά 1000 κατοίκους στην Κρήτη έως 268/1000 στην Ανατολική Φινλανδία [Menotti (1999)].

Η ύπαρξη μεγάλου δείγματος θα βοηθούσε σημαντικά τη βελτίωση της εφαρμογής των ΤΝΔ. Η δυνατότητα εκπαίδευσης των ΤΝΔ με μεγάλο αριθμό δεδομένων, αν και χρονοβόρος, θα βελτίωνε σημαντικά τη διαχωριστική τους

ικανότητα και τα άλλα στατιστικά χαρακτηριστικά (δείκτης Youden, calibration κλπ). Θα ήταν ενδιαφέρον να μελετηθεί η διαχωριστική ικανότητα των διαφόρων μεθόδων συναρτήσει του μεγέθους του δείγματος που χρησιμοποιούν για την εξαγωγή των συντελεστών των μοντέλων στα οποία καταλήγουν.

Με τη χρήση των ΤΝΔ θα μπορούσε επίσης να επιχειρηθεί η πρόβλεψη των τιμών του δείκτη Gensini και των τιμών του DSN. Η ανάπτυξη ασαφών νευρωνικών δικτύων σε διαφορετικά δείγματα με δεδομένα που δεν προέρχονται από επεμβατικές μεθόδους, θα μπορούσε να οδηγήσει σε πρόβλεψη κριτικών τιμών του DSN σε σχέση με παθήσεις στεφανιαίας νόσου και θα μπορούσε να αποτελέσει αντικείμενο επόμενης μελέτης.

Ένα άλλο πιθανό αντικείμενο έρευνας θα ήταν η συνδυασμένη δράση των διαχωριστικών τεχνικών και ΤΝΔ με την αναλογιστική επιστήμη με σκοπό τον προσδιορισμό του κόστους νοσηλείας και κατ' επέκταση των ασφαλίστρων ζωής σε άτομα που ανήκουν σε ομάδες υψηλού κινδύνου.

ΠΑΡΑΡΤΗΜΑ Α

ΜΑΘΗΜΑΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΤΩΝ Τ.Ν.Δ.

Στατιστική Εκμάθηση

Η στατιστική εκμάθηση βασίζεται στις πιθανότητες ή στις στατιστικές πληροφορίες που προσλαμβάνει το δίκτυο κατά την διάρκεια της εκπαίδευσης του. Ένα αντιπροσωπευτικό παράδειγμα είναι η εκπαίδευση κατά Bayes, η οποία εκτιμά την απαιτούμενη συνάρτηση πυκνότητας πιθανότητας για μια διαδικασία απόφασης (Limin Fu (1994)). Ανάλογες μέθοδοι εκμάθησης συμπεριλαμβάνουν γραμμικές διαχωριστικές διαδικασίες για ταξινόμηση καθώς και μεταβατικές πιθανοθεωρητικές διαδικασίες.

Στην αναγνώριση προτύπου και στα συστήματα ελέγχου, μια απόφαση έχει σχέση με την κατανομή δεδομένων εισόδου σε μια προκαθορισμένη τάξη.

Προσδιορίζοντας μια συνάρτηση σφάλματος που αναγνωρίζει και παρεμβαίνει σε λανθασμένες αποφάσεις, μια εκτιμήτρια Bayes είναι σε θέση να ελαχιστοποιήσει το μέσο κόστος σφάλματος και κατα συνέπεια να χρησιμοποιηθεί για να μοντελοποιήσει το άγνωστο σύστημα. Οι Duda και Hart (1973) δίνουν μια περιγραφή αυτής της διαδικασίας.

Γενικευμένη εκμάθηση κατά Bayes.

- Σκοπός της εκμάθησης κατά Bayes είναι να εκτιμηθεί η συνάρτηση πυκνότητας πιθανότητας p(X) για μια μεταβλητή X από ένα σύνολο Ω, η ανεξάρτητα επιλεγμένων, δειγμάτων $x_1, x_2, \ldots, x_n$ σύμφωνα με την p(X).
- Βασική υπόθεση:

  -Η μορφή της συνάρτησης p(X) να είναι γνωστή (για παράδειγμα Gaussian), αλλά η τιμή του διανύσματος παραμέτρων θ να μην είναι σαφώς καθορισμένη.

  -Η αρχική μας γνώση για την παράμετρο θ να περιέχεται σε μια a priori πιθανότητα p(θ).

  -Η εκτίμηση των τιμών της παραμέτρου θ να γίνει από το σύνολο Ω των n δειγμάτων.
- Η εξίσωση εκμάθησης είναι:

$$p(\theta \setminus \Omega) = \frac{p(\Omega \setminus \theta)\, p(\theta)}{\int p(\Omega \setminus \theta)\, p(\theta) d\theta}$$

και σύμφωνα με την υπόθεση ανεξαρτησίας,

$$p(\Omega \setminus \theta) = \prod_{k=1}^{n} p(x_k | \theta)$$

Η πιθανότητα p(θ) εκτιμάται από την παραπάνω εξίσωση σύμφωνα με το σύνολο Ω ή με ένα προς ένα τα δείγματα $x_1, x_2, \ldots, x_n$.

Θεωρία Εκμάθησης

Γραμμικές διαχωριστικές τεχνικές χρησιμοποιούνται στο στάδιο της εκμάθησης. Η ανάλυση και η ταξινόμηση δεδομένων σε δύο κατηγορίες μπορεί να επεκταθεί σε ανάλυση σε πολλαπλές κατηγορίες με ειδικές τεχνικές όπως οι κατασκευές Kesler (Duda και Hart (1973)).Στην παρούσα αναφορά περιοριζόμαστε στη ταξινόμηση δεδομένων σε δύο κατηγορίες.

Ας υποθέσουμε ότι υπάρχουν n σημεία ενός d-διάσταστου χώρου τα οποία κατηγοριοποιούνται ως ω_1 ή ω_2. Ένα υπερεπίπεδο που διαχωρίζει τα σημεία ω_1 από τα σημεία ω_2 ονομάζεται γραμικά διχοτομούν επίπεδο. Η συνάρτηση f(n,d) των γραμμικά διχοτομούντων επιπέδων ως προς τα 2^n διχοτομούντα επιπέδα αυτών των n σημείων δίνεται από τους Duda και Hart:

$$f(n,d) = \begin{cases} 1 & n \le d+1 \\ \frac{2}{2^n} \sum_{i=0}^{d} \binom{n-1}{i} & n > d+1 \end{cases}$$

Αυτό σημαίνει ότι υπάρχει πάντα ένα υπερεπίπεδο για να ταξινομεί d+1 σημεία σωστά. Η βασική προϋπόθεση ώστε να μην υποεκτιμηθεί η γραμμική διαχωριστική συνάρτηση είναι ο αριθμός των σημείων να μην είναι μικρότερος από d+1. Για τη τιμή n=2(d+1), η οποία ονομάζεται ικανότητα του υπερεπιπέδου, τα μισά από τα διχοτομούντα επίπεδα είναι γραμμικά.

Ανάστροφη μετάδοση σφάλματος (backpropagation)

Στη συνέχεια αναφέρεται πως επιτυγχάνεται η διαδικασία εκπαίδευσης των ΤΝΔ με τον κανόνα εκμάθησης «Ανάστροφη μετάδοση σφάλματος» ο οποίος περιγράφεται από τη σχέση:

$$\Delta W_{ji} = \eta \delta_j O_i$$

όπου W_{ji} είναι το βάρος από μια είσοδο O_i , η είναι ο ανεξαρτήτως επαναλήψεων ρυθμός εκμάθησης ($0<\eta<1$) και δ_j είναι η κλίση σφάλματος στον κόμβο j. Αν ένας κόμβος j είναι κόμβος εξόδου, τότε το δ_j υπολογίζεται από:

$$\delta_j = (T_j - O_j)F'(net_j)$$

όπου

$$net_j = \sum_i W_{ji} O_i$$

Αν F είναι μία σιγμοειδής συνάρτηση τότε:

$$O_j = F(net_j) = F\left(\sum_i W_{ji} O_i\right)$$

Αν ο κόμβος j είναι ένας κόμβος του κρυφού επιπέδου, τότε το δ_j υπολογίζεται από τη σχέση: $\delta_j = F_j'(net_j)\sum_k \delta_k W_{kj}$

Η διαδικασία ανάστροφης μετάδοσης σφάλματος ελαχιστοποιεί το κριτήριο σφάλματος

$$E = \frac{1}{2}\sum_j (T_j - O_j)^2$$

Η τεχνική κατιούσας κλίσης καταλήγει σε:

$$\Delta W_{ji} = -\eta\,(\partial E / \partial W_{ji})$$

και με τον κανόνα της αλυσίδας έχουμε ότι:

$$\partial E / \partial W_{ji} = (\partial E / \partial O_j)(\partial O_j / \partial W_{ji})$$

Σε περίπτωση που ο κόμβος j είναι κόμβος εξόδου, τότε :

$$\partial E / \partial O_j = -(T_j - O_j)$$
$$\partial O_j / \partial W_{ji} = F_j'(net_j) O_i$$

Άρα,

$$\begin{aligned} \partial E / \partial W_{ji} &= (\partial E / \partial O_j)(\partial O_j / \partial W_{ji}) \\ &= -(T_j - O_j) F_j'(net_j) O_i \\ &= -\delta_j O_i \end{aligned}$$

Έτσι καταλήγουμε στη σχέση:

$$\Delta W_{ji} = \eta \delta_j O_i$$

Όταν ο κόμβος j είναι κόμβος κρυφού επιπέδου, τότε T_j δεν δίνεται Τότε:

$$\partial E / \partial O_j = \sum_k (\partial E / \partial O_k)(\partial O_k / \partial O_j)$$

Η έξοδος του κόμβου k δίνεται από:

$$O_k = F\left(\sum_j W_{kj} O_j\right)$$

και

$$\partial O_k / \partial O_j = F'(net_k) W_{kj}$$

Ως αποτέλεσμα έχουμε,

$$\begin{aligned} \partial E / \partial O_j &= \sum_k (\partial E / \partial O_k)(\partial O_k / \partial O_j) \\ &= -\sum_k (T_k - O_k) F'(net_k) W_{kj} \\ &= -\sum_k \delta_k W_{kj} \end{aligned}$$

όπου,

$$\begin{aligned} \partial E / \partial W_{ji} &= (\partial E / \partial O_j)(\partial O_j / \partial W_{ji}) \\ &= -\left(\sum_k \delta_k W_{kj}\right) F_j'(net_j) O_j \\ &= -\delta_j O_i \end{aligned}$$

Άρα καταλήγουμε στο,

$$\Delta W_{ji} = \eta \delta_j O_i$$

Γενίκευση

Η ικανότητα γενίκευσης είναι η πιο σημαντική των νευρωνικών δικτύων και μετρά την απόδοση του δικτύου σε δεδομένα εκτός του συνόλου με το οποίο εκπαιδεύτηκε. Οι ευρέως χρησιμοποιούμενοι μέθοδοι είναι :

α) η εύρεση άνω και κάτω φράγματος για το σφάλμα γενίκευσης, β) η μέθοδος «cross validation» και γ) η PSE

α) Οι Vapnik και Chervonenkis (1971) προτείνουν τα ακόλουθα φράγματα για το σφάλμα γενίκευσης σε ένα δίκτυο με ένα κρυφό επίπεδο πλήρως συνδεδεμένο:

$$2\lfloor N_h / 2 \rfloor d \le VCdim \le 2 N_w \log(e N_n)$$

όπου d ο αριθμός των μονάδων εισόδου, N_h ο αριθμός των κρυφών μονάδων, N_w ο συνολικός αριθμός των βαρών του δικτύου και N_n ο συνολικός αριθμός των κόμβων του δικτύου.

β) Η γενικότερη μέθοδος είναι η «cross validation» όπου τα δεδομένα χωρίζονται στα δεδομένα εκπαίδευσης και στα δεδομένα ελέγχου και εφαρμόζονται οι γνωστές στατιστικές τεχνικές.

γ) Άλλη τεχνική που μετρά την ικανότητα γενίκευσης του δικτύου είναι η μέτρηση του προβλεπόμενου τετραγωνικού σφάλματος (predicted squared error-PSE) (Moody(1971)).

$$PSE = MSE + \frac{2N_w}{P}\sigma^2$$

όπου MSE είναι το μέσο τετραγωνικό σφάλμα του συνόλου εκπαίδευσης, N_w ο αριθμός των ελευθέρων βαρών , P ο αριθμός των σταδίων εκπαίδευσης και σ^2 η διακύμανση του θορύβου. Το μέτρο αυτό δίνει μια αμερόληπτη εκτίμηση του PSE για μη γραμμικά συστήματα όπως τα νευρωνικά δίκτυα.

Πιθανοθεωρητικά Νευρωνικά δίκτυα

Ο κανόνας απόφασης κατά Bayes είναι ο κανόνας εκείνος μέσω του οποίου επιλέγεται η κατηγορία με τον ελάχιστο δεσμευμένο κίνδυνο.
Δοθέντος ενός διανύσματος x , σε χώρο c κλάσεων $\omega_1, \omega_2, \ldots, \omega_c$, ο κανόνας *ρυθμού ελαχίστου σφάλματος θα ταξινομήσει το διάνυσμα x στη κλάση* ω_i αν ισχύει ότι:

$$P(\omega_i | x) > P(\omega_j | x) \text{ για } i \neq j$$

Εδώ η a posteriori πιθανότητα χρησιμοποιείται σαν διαχωριστική συνάρτηση. Μια άλλη επιλογή διαχωριστικής συνάρτησης για να επιτευχθεί ο ελάχιστος ρυθμός σφάλματος είναι η $P(x | \omega_i)P(\omega_i)$.

Για κάθε κατηγορία c κλάσεων , ας προσδιορίσουμε την συνάρτηση πυκνότητας πιθανότητας(PDF) ως εξής: $f_c(x) = P(x | c)$

Ο Parzen (1962) απέδειξε ότι μια ομαλοποιημένη συνεχής συνάρτηση μπορεί να προσεγγιστεί ασυμπτωτικά από μια τάξη PDF εκτιμητριών. Στηριζόμενος σε αυτή την ιδέα ο Specht (1990) ανέπτυξε μια ειδική εκτιμήτρια συνάρτηση για την τάξη PDF βασιζόμενος στα πρότυπα εκπαίδευσης:

$$f_c(x)=\frac{1}{(2\pi)^{d/2}\,\sigma^d}\,\frac{1}{m}\sum_{i=1}^{m}\exp\left[-(x-X_{ci})\cdot(x-X_{ci})/2\sigma^2\right]$$

όπου :

d=η διάσταση του επιπέδου εισόδου

m=ο συνολικός αριθμός των πρότυπων εκπαίδευσης

σ = παράμετρος εξομάλυνσης

X_{ci} = το i πρότυπο της κατηγορίας c.

Η $f_c(x)$ είναι το άθροισμα κατανομών Gauss επικεντρωμένο σε κάθε πρότυπο εκπαίδευσης. Σημειώνεται ότι το άθροισμα αυτό μπορεί να προσεγγίσει οποιαδήποτε ομαλοποιημένη συνάρτηση πυκνότητας μη περιοριζόμενο σε κατανομή Gauss.

Στα πιθανοθεωρητικά νευρωνικά δίκτυα, κάθε πρότυπο εκπαίδευσης X_{ci} κωδικοποιείται σαν ένα διάνυσμα βαρών εισόδου της μονάδας i , που ονομάζεται μονάδα προτύπου. Άρα $X_{ci} = W_i$. Η συνολική είσοδος στη μονάδα i ισούται με $z_i=x \cdot W_i$ με την υπόθεση ότι τόσο το x όσο και το W_i είναι κανονικοποιημένα στη μονάδα μήκους. Η παραπάνω εξίσωση μπορεί να μετασχηματιστεί στην

$$f_c(x)=\frac{1}{(2\pi)^{d/2}\,\sigma^d}\,\frac{1}{m}\sum_{i=1}^{m}\exp\left[(z_i-1)/\sigma^2\right]$$

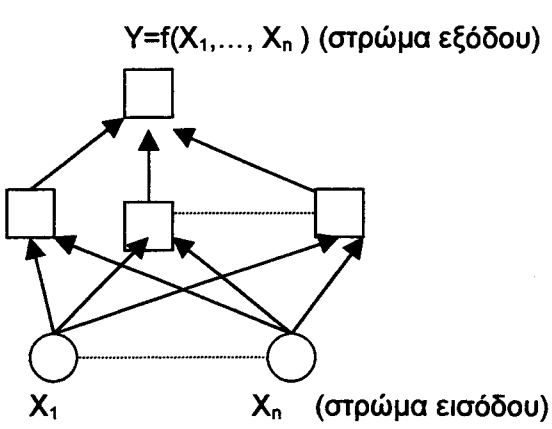
Η διαδικασία αυτή καθορίζει το σχεδιασμό του δικτύου, με τη συνάρτηση ενεργοποίησης της μονάδας i να ορίζεται από τη σχέση:

$$F(z)=\exp\left[(z-1)/\sigma^2\right].$$

$$O_j=\frac{1}{m}P(\omega_j)\sum_i W_{ji}\,O_i$$

Νευρωνικά δίκτυα ως Μαθηματικά Μοντέλα

Η βάση για τη χρήση των νευρωνικών δικτύων σαν μαθηματικά μοντέλα είναι η απεικόνιση (mapping).Το θεώρημα του Kolmogorov καθορίζει μια αυστηρή μαθηματική θεμελίωση για τα δίκτυα απεικόνισης. Τα δίκτυα αυτά προσεγγίζουν στην ουσία μια μαθηματική συνάρτηση , όπως φαίνεται στο διάγραμμα του Σχ. Π1 που ακολουθεί:



Σχ. Π1

Πιο αναλυτικά, τα δίκτυα απεικόνισης χρησιμοποιούν μια φραγμένη απεικόνιση ή συνάρτηση από ένα φραγμένο σύνολο n διαστάσεων σε ένα άλλο φραγμένο σύνολο m διαστάσεων (όπου n=m ή n≠m). Έστω ότι χρησιμοποιούμε ένα νευρωνικό δίκτυο για να προσεγγίσουμε μια συνάρτηση f. Το δίκτυο εκπαιδεύεται έτσι ώστε μια συνάρτηση F_{NN} που συνδέεται άμεσα με τη λειτουργία του δικτύου, να προσεγγίσει την f, δηλαδή $F_{NN} \approx f$. Ο αλγόριθμος της ανάστροφης μετάδοσης (backpropagation) είναι ο πλέον συνηθισμένος για την εκπαίδευση των δικτύων απεικόνισης.

Για την εκμάθηση του δικτύου, επιλέγεται ένα σύνολο στοιχείων εκμάθησης από το δείγμα. Κάθε στοιχείο εκπροσωπείται από ένα ζεύγος (x,y), όπου y=f(x). Ο στόχος μας είναι μετά την εκμάθηση να ισχύει $y=f(x)=F_{NN}$ για κάθε x του συνόλου εκμάθησης.

Στη διάρκεια της εκμάθησης προσπαθούμε να ελαχιστοποιήσουμε τη διαφορά μεταξύ F_{NN} και f. Υπάρχουν αρκετά κριτήρια καθορισμού της διαφοράς. Το πλέον διαδεδομένο είναι το κριτήριο του μέσου τετραγωνικού σφάλματος. Το τετράγωνο του σφάλματος για ένα στοιχείο εκμάθησης είναι

$$E^2(x) = \left| f(x) - F_{NN}(x) \right|^2$$

Το μέσο τετραγωνικό σφάλμα για N στοιχεία εκμάθησης δίνεται από τη σχέση:

$$\overline{E^2} = \frac{1}{N} \sum E^2(x)$$

Μπορούμε επίσης να χρησιμοποιήσουμε τη σχέση:

$$\overline{E^2} = \int E^2(x)\rho(x)dx$$

όπου ρ(x) είναι η συνάρτηση πυκνότητας πιθανότητας. Ο κανόνας ελαχίστου μέσου τετραγωνικού σφάλματος χρησιμοποιείται επιτυχώς και στη στατιστική σε διάφορα πεδία ελέγχου για αναγνώριση προτύπων.

Θεώρημα Kolmogorov

Ένα σημαντικό θεώρημα που αναδεικνύει την ικανότητα των πολυστρωματικών νευρωνικών δικτύων αποδείχθηκε από τον Kolmogorov και αναφέρεται από τον Lorentz (1976).Το θεώρημα αποδεικνύει ότι κάθε συνεχής συνάρτηση μπορεί να εκφρασθεί με τη βοήθεια μη γραμμικών συνεχών και αυξουσών συναρτήσεων μιας μόνο μεταβλητής.

Το θεώρημα ύπαρξης απεικόνισης νευρωνικών δικτύων του Kolmogorov απόδεικνύει ότι σε κάθε συνεχή συνάρτηση f, από τον n-διάστατο χώρο $[0,1]^n$ στο σύνολο των πραγματικών αριθμών, μπορεί να εφαρμοσθεί πλήρως ένα *εμπρόσθιο νευρωνικό δίκτυο τριών επιπέδων με n στοιχεία στο επίπεδο εισόδου*, 2n+1 στοιχεία στο κρυφό επίπεδο και m στοιχεία στο επίπεδο εξόδου. Τα στοιχεία του κρυφού επιπέδου χρησιμοποιούν ως συνάρτηση ενεργοποίησης την *ακόλουθη*:

$$h_k = \sum_{i=1}^{n} \lambda^k \psi(x_i + k\varepsilon) + k$$

όπου η συνεχής και αύξουσα πραγματική συνάρτηση ψ και η σταθερά $\lambda \in R$ είναι ανεξάρτητες από την συνάρτηση f, η σταθερά ε είναι ένας θετικός ρητός αριθμός όχι μεγαλύτερος από μία αυθαίρετα επιλεγμένη τιμή κατωφλίου, και x_i είναι η i είσοδος.

Τα στοιχεία του επιπέδου εξόδου χρησιμοποιούν την ακόλουθη συνάρτηση ενεργοποίησης: $y_j = \sum_{k=1}^{2n+1} g_j(h_k)$, όπου η πραγματική συνεχής συνάρτηση g_j εξαρτάται από την f και την c.

Η απόδειξη του θεωρήματος δεν υποδεικνύει τον τρόπο με τον οποίο θα βρεθούν οι μη γραμμικές συναρτήσεις ενεργοποίησης ψ και g, υποστηρίζει μόνο την ύπαρξη ενός δικτύου τριών επιπέδων που μπορεί να απεικονισθεί.

Ο Cybenko(1989) έδειξε επίσης, ότι ένα πεπερασμένο άθροισμα της μορφής

$$y = \sum_i W_i \sigma(a_i \cdot x + b_i)$$

είναι πυκνό στο χώρο των συνεχών συναρτήσεων με πεδίο ορισμού στο $[0,1]^n$ όπου η συνάρτηση σ είναι μια συνεχής διαχωριστική συνάρτηση και b_i, a_i, W_i αντιστοιχούν στη μεροληψία, το διάνυσμα βαρών εισόδου, και τα βάρη εξόδου αντίστοιχα, που συνδέονται με τον κρυφό κόμβο i.

Γενικευμένη παλινδρόμηση στο Νευρωνικά Δίκτυα

Η γενικευμένη παλινδρόμηση αναπτύχθηκε από τον Specht (1991) και αναφέρεται περισσότερο στα πιθανοθεωρητικά νευρωνικά δίκτυα. Χρησιμοποιείται κυρίως για την εκτίμηση συνεχών μεταβλητών και ακολουθεί γενικώς την πολυμεταβλητή στατιστική παλινδρόμηση.

Έστω f(x,y) η γνωστή συνεχής δεσμευμένη συνάρτηση πυκνότητας πιθανότητας μιας διανυσματικής τυχαίας μεταβλητής x και y μια μονόμετρη μεταβλητή. Αν X μια τιμή της μεταβλητής x, η υπο συνθήκη αναμενόμενη τιμή του y δοθέντος του X δίνεται από την ακόλουθη σχέση:

$$E[y|X] = \frac{\int_{-\infty}^{\infty} y\, f(X,y)\, dy}{\int_{-\infty}^{\infty} f(X,y)\, dy}$$

Όταν η συνάρτηση πυκνότητας πιθανότητας είναι άγνωστη, χρησιμοποιείται συνήθως η εκτιμήτρια συνάρτηση του Parzen. Αν X_i και Y_i ένα δείγμα παρατηρήσεων των μεταβλητών x και y, τότε η εκτιμήτρια είναι η ακόλουθη:

$$\hat{f}(X,Y) = \frac{1}{(2\pi)^{(d+1)/2}\sigma^{d+1}} \cdot \frac{1}{n}\sum_{i=1}^{n} \exp\left[-\frac{D_i^2}{2\sigma^2}\right] \cdot \exp\left[-\frac{(Y-Y_i)^2}{2\sigma^2}\right] \quad (1)$$

όπου n είναι ο αριθμός των παρατηρήσεων ,d είναι η διάσταση του διανύσματος της μεταβλητής x και $D_i^2 = (X-X_i)\cdot(X-X_i)$.

Με αντικατάσταση των ανωτέρω συναρτήσεων καταλήγουμε στην εκτιμηθείσα υπό συνθήκη αναμενόμενη τιμή:

$$\hat{Y}(X)=\frac{\sum_{i=1}^{n} Y_i \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}$$

Κύριο σημείο για την εύρεση βέλτιστης εκτιμήτριας είναι η επιλογή του πλάτους του σ (παράμετρος εξομάλυνσης). Ο Parzen απέδειξε ότι εκτιμήτριες της προαναφερθείσας μορφής συγκλίνουν ασυμπτωτικά σε όλα τα σημεία της συνάρτησης πυκνότητας πιθανότητας (συνεπείς εκτιμήτριες), όπου η συνάρτηση πυκνότητας είναι συνεχής, με την προϋπόθεση ότι σ(n) είναι μια φθίνουσα συνάρτηση έτσι ώστε:

$$\lim_{n\to\infty} \sigma(n)=0$$

και

$$\lim_{n\to\infty} n\sigma^d(n)=\infty$$

Ο πυρήνας του Gauss που χρησιμοποιήθηκε στην εξίσωση (1) μπορεί να αντικατασταθεί από την προσέγγιση του Parzen, όπως παρατηρούμε στην ακόλουθη εκτιμήτρια:

$$\hat{Y}(X)=\frac{\sum_{i=1}^{n} Y_i \exp\left(-\frac{C_i}{\sigma}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{C_i}{\sigma}\right)},$$

όπου $C_i=\sum_{j=1}^{d}\left|X^j - X_i^j\right|$ και j δηλώνει την j συνιστώσα του διανύσματος εισόδου.

Όταν ο αριθμός των παρατηρήσεων είναι μεγάλος, είναι ανέφικτο να αποδίδεται ένας διαφορετικός κόμβος σε κάθε περίπτωση. Τεχνικές clustering χρησιμοποιούνται για να ομαδοποιήσουν δεδομένα , έτσι ώστε μια ομάδα να αντιπροσωπεύεται από ένα κόμβο. Αυτή η διαδικασία καταλήγει στην κατασκευή δικτύου με μεγαλύτερη ικανότητα γενίκευσης.

Στην περίπτωση των m ομάδων η εκτιμήτρια έχει την ακόλουθη μορφή:

$$\hat{Y}(X)=\frac{\sum_{i=1}^{m} A_i \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}{\sum_{i=1}^{m} B_i \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}$$

όπου A_i το άθροισμα των τιμών της Υ μεταβλητής, B_i ο αριθμός των περιπτώσεων που ανήκουν στην ομάδα i και D_i η Ευκλείδεια απόσταση του διανύσματος εισόδου από το κεντροειδές της ομάδας. Έτσι X_i είναι τώρα το κεντροειδές της ομάδας i. Οι παράμετροι A_i και B_i προσαυξάνονται κάθε φορά που μια παρατήρηση εκπαίδευσης Y_j απόδίδεται στην ομάδα i.

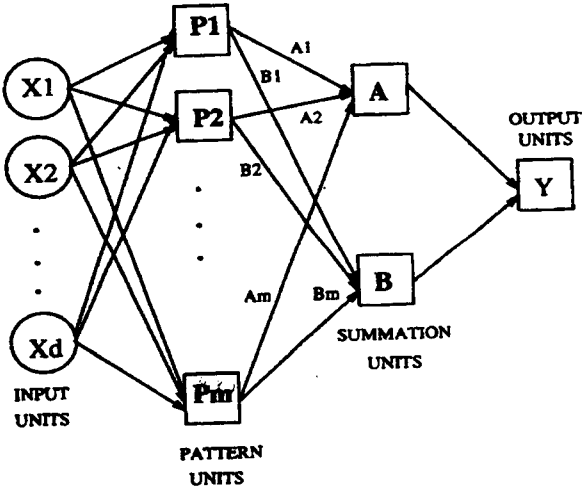
$$A_i (k) = A_i (k - 1) + Y_j$$

και

$$B_i (k) = B_i (k - 1) + Y_j$$

όπου k δηλώνει την τιμή της παραμέτρου μετά από k παρατηρήσεις.

Το διάγραμμα του Σχ.Π2 παρουσιάζει τη γενικευμένη παλινδρόμηση ενός νευρωνικού δικτύου. Κάθε μονάδα προτύπου συνδέεται με ένα υπόδειγμα ή με ένα κεντροειδές.



Σχ. Π2: Γενικευμένη παλινδρόμηση νευρωνικού δικτύου για την εκτίμηση συνεχών μεταβλητών.

ΠΑΡΑΡΤΗΜΑ Β

ΧΡΗΣΗ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ Profile Neural Applications-P.N.A

Σε αυτό το κεφάλαιο παρουσιάζονται συνοπτικά οι δυνατότητες του Profile Neural Applications-P.N.A (Brain Maker Professional ο πυρήνας ενός νευρωνικού δικτύου) Το P.N.A., απότελείται από τα εξής υποσυστήματα:

- Data Collector

Με τη χρήση αυτού του υποσυστήματος, ο χρήστης μπορεί να συλλέξει στοιχεία από τις Βάσεις της PROFILE σε τέτοια μορφή έτσι ώστε να είναι εύκολα επεξεργάσιμη από το επόμενο υποσύστημα.

- Net Maker

Είναι το υποσύστημα εκείνο το οποίο παρέχει τα εργαλεία, που μας επιτρέπουν να επεξεργαστούμε τα πρωτογενή στοιχεία που συλλέξαμε, έτσι ώστε να δημιουργήσουμε τα δεδομένα της εκμάθησης Ακόμα, μπορούμε να χωρίσουμε τα δεδομένα μας σε αυτό της εκπαίδευσης (training) και σε αυτά του ελέγχου (testing).

- Brain Maker

Είναι ο πυρήνας του Νευρωνικού Δικτύου, δηλαδή το υποσύστημα εκείνο που αναλαμβάνει την εκπαίδευση καθώς και την εκτέλεση του δικτύου.

- Competitor

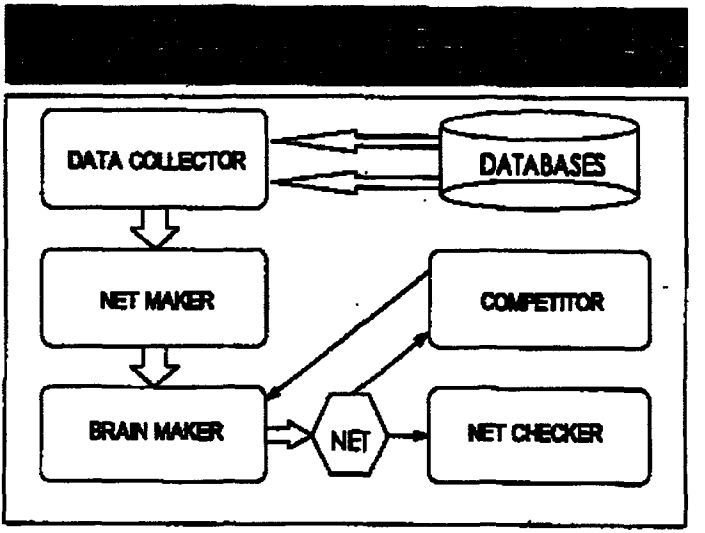
Είναι ένα σύνολο από εργαλεία τα οποία αναλαμβάνουν την εκπαίδευση και την εκτέλεση συγκεκριμένων δικτύων. Τέτοια δίκτυα είναι αυτά στα οποία δε ζητούμε να μας προβλέψουν μία τιμή, αλλά να πραγματοποιήσουν μία ταξινόμηση.

- Net Checker

Είναι ένα βοηθητικό πρόγραμμα, το οποίο ελέγχει το δίκτυο και ειδοποιεί για πιθανά προβλήματα.

- Genetic Training Options (GTOs)

Είναι το υποσύστημα εκείνο που αναλαμβάνει να βελτιώσει ένα νευρωνικό δίκτυο, χρησιμοποιώντας γενετικούς αλγορίθμους (μετάλλαξη και διασταύρωση).



1. Data Collector

Το υποσύστημα αυτό παρέχει εύκολη πρόσβαση στις Βάσεις τις PROFILE. Μέσω αυτού μπορούμε να επιλέξουμε τα στοιχεία που θέλουμε να επεξεργαστούμε.

Χρήσιμα εργαλεία είναι: α) Metastock, πληροφορίες για μακροοικονομικά δεδομένα από την αντίστοιχη Βάση, ημερήσια στοιχεία μετοχών κ.λ.π.

β) το Professional Fundamentals μπορούμε να επιλέξουμε χρηματοοικονομικά στοιχεία και αριθμοδείκτες για τις εταιρίες που επελέγησαν. Τέλος, κάνει μία επεξεργασία και ένα συνδυασμό αυτών των στοιχείων, έτσι ώστε να μπορούν να διαβαστούν για περαιτέρω επεξεργασία από το επόμενο υποσύστημα της εφαρμογής, το Net Maker.

2. Net Maker

Το υποσύστημα που επεξεργάζεται τα πρωτογενή στοιχεία, τα οποία τα έχουμε συλλέξει με το Data Collector, είναι το Net Maker. Οι λειτουργίες που παρέχει το Net Maker είναι παρόμοιες με τις λειτουργίες ενός spreadsheet, σε μια πιο απλή μορφή.

2.1 Λειτουργίες Στήλης

Αυτά που μπορούμε να εφαρμόσουμε σε μία στήλη, είναι

- να την αντιγράψουμε σε μία άλλη
- να τη μετακινήσουμε σε μία άλλη θέση
- να ανταλλάξουμε τις θέσεις δύο στηλών
- να διαγράψουμε μία στήλη
- να προσθέσουμε δύο στήλες
- να αφαιρέσουμε δύο στήλες
- να πολλαπλασιάσουμε δύο στήλες
- να διαιρέσουμε δύο στήλες
- να "σηκώσουμε" μία στήλη προς τα πάνω
- να "κατεβάσουμε" μία στήλη προς τα κάτω

Οι παραπάνω ενέργειες είναι όλες προφανείς, εκτός από τις δύο τελευταίες. Μία στήλη χρειάζεται να την μετακινήσουμε προς τα πάνω, όταν αυτή η στήλη αντιπροσωπεύει αυτό που θέλουμε να προβλέψουμε. Με αυτόν τον τρόπο και έχοντας πάντα στο μυαλό μας ότι τα στοιχεία επεξεργάζονται από τον πυρήνα του δικτύου κατά γραμμές, μεταφέρουμε τις τιμές των επόμενων περιόδων στη σειρά που μας ενδιαφέρει.

Αντίστοιχα, μετακινώντας μία στήλη προς τα κάτω, δημιουργούμε μία μορφή "μνήμης" κάποιων γεγονότων. Δηλαδή, μεταφέρουμε τιμές του παρελθόντος στο σήμερα, έτσι ώστε να είναι διαθέσιμες.

2.2 Λειτουργίες Γραμμής

Για μία γραμμή των δεδομένων μας, μπορούμε να εκτελέσουμε τα εξής:

- να αντιγράψουμε μία γραμμή
- να διαγράψουμε μία γραμμή
- να ανακατέψουμε τις γραμμές
- να δημιουργήσουμε μία γραμμή μεγίστων και ελαχίστων

Η διαγραφή μίας γραμμής είναι απαραίτητη όταν μετακινούμε τις στήλες προς τα πάνω ή προς τα κάτω. Μία τέτοια μετακίνηση συνεπάγεται την απώλεια πληροφορίας στην αρχή ή στο τέλος των δεδομένων μας.

Τέτοια κενά δεν επιτρέπονται για το λόγο ότι θα δημιουργήσουν εσφαλμένες παρατηρήσεις. Αυτό θα είχε δυσάρεστες συνέπειες στην εκπαίδευση του δικτύου και στην απόδοση του.

Το ανακάτεμα των γραμμών είναι μία ενέργεια που πρέπει να πραγματοποιηθεί αν δε χρησιμοποιούμε αναδρομικά γεγονότα (recurrent inputs, θα αναλυθεί στο επόμενο κεφάλαιο). Έχει παρατηρηθεί ότι. όταν η σειρά των γεγονότων εκπαίδευσης ενός δικτύου είναι χρονολογική, τότε το δίκτυο δεν εκπαιδεύεται καλά. Όταν έχουμε χρονολογική σειρά στα γεγονότα, γενικά οι απόκλίσεις μεταξύ δύο ημερών είναι σχετικά μικρές και αυτό δυσχεράνει τη διαδικασία εκμάθησης του. Αντίθετα, όταν τα διάφορα γεγονότα είναι ανακατεμένα, το δίκτυο σχηματίζει πιο γρήγορα άποψη για το σύνολο των δεδομένων για τα οποία εκπαιδεύεται.

Η δημιουργία γραμμών μεγίστων και ελαχίστων για κάθε στήλη, μας επιτρέπει να καθορίσουμε εμείς τις ελάχιστες και τις μέγιστες τιμές της. Τι σημαίνει όμως να καθορίσουμε τα μέγιστα και τα ελάχιστα, αφού αυτά μπορούν να υπολογιστούν εύκολα;

Αν επηρεάσουμε τις μέγιστες ή τις ελάχιστες τιμές αυτές, τότε, κατά την εκπαίδευση του δικτύου, όταν ο πυρήνας εκπαίδευσης συναντήσει μία τιμή μεγαλύτερη από το μέγιστο (ή μικρότερη από το ελάχιστο), τότε θα συνεχίσει τους υπολογισμούς του με την τιμή που του έχει δοθεί. Αυτή η λειτουργία έχει σαν συνέπεια την ταχύτερη και καλύτερη εκπαίδευση του δικτύου μας.

2.3 Ετικέτες

Με το σύνολο των λειτουργιών αυτών, μπορούμε να ονομάσουμε μία στήλη με το όνομα της επιλογής μας, αλλά και να μαρκάρουμε τις στήλες για την εκπαίδευση του δικτύου σαν:

- Input, δηλ. η στήλη αποτελεί δεδομένο εισόδου (ανεξάρτητη μεταβλητή)
- Pattern, δηλ. η στήλη περιέχει αυτό που θέλουμε να προβλέψουμε
- Annote, δηλ. η στήλη περιέχει απλώς σημειώσεις
- Not Used. δηλ. η στήλη να μη χρησιμοποιηθεί καθόλου στην εκπαίδευση

2.4 Αριθμοί-Σύμβολα

Με τις λειτουργίες των αριθμών, μπορούμε να εκτελέσουμε τις βασικές πράξεις ενός αριθμού με μία στήλη. Δηλαδή, για παράδειγμα, να πολλαπλασιάσουμε μία στήλη με το -1, έτσι ώστε να πάρουμε μία νέα στήλη που θα είναι το αντίθετο της πρώτης. Επίσης, μπορούμε να ζητήσουμε την αποκοπή των δεκαδικών ψηφίων από μία στήλη.

Οι λειτουργίες των συμβόλων μας επιτρέπουν να μετατρέψουμε αριθμούς σε σύμβολα. Αυτή η ιδιότητα χρησιμοποιείται όταν θέλουμε να επεξεργαστούμε τους αριθμούς σαν να ήταν σύμβολα. Ένα τέτοιο παράδειγμα είναι αν το 1 συμβολίζει σήμα αγοράς, το -1 σήμα πώλησης και το Ο σήμα παραμονής. Αν δε μετατρέψουμε τους αριθμούς αυτούς σε σύμβολα, τότε θα παίρνουμε αποτελέσματα της μορφής 0.8789. Αν όμως κάνουμε τη μετατροπή, θα λάβουμε για κάθε σύμβολο το αποτέλεσμα ισχύει-δεν ισχύει.

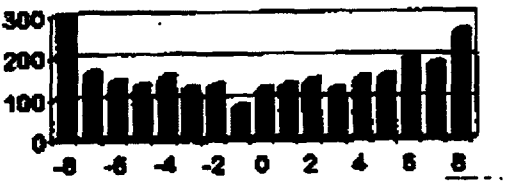
2.5 Στατιστικές Λειτουργίες

Οι λειτουργίες αυτές είναι από τις πιο σημαντικές για την κατασκευή ενός αποτελεσματικού δικτύου. Προσφέρουν τις πληροφορίες εκείνες που χρειαζόμαστε για να αναλύσουμε τα δεδομένα, καθώς και για να δημιουργήσουμε άλλες στήλες που να περιέχουν διάφορες στατιστικές πράξεις.

- Λειτουργίες ανάλυσης δεδομένων

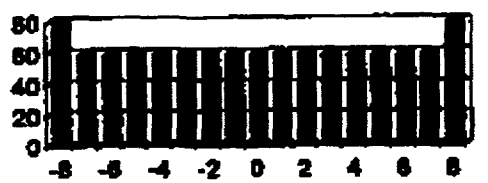
2.5.1 Ιστόγραμμα / Γράφημα

Με το ιστόγραμμα έχουμε τη δυνατότητα να δούμε την κατανομή των τιμών μίας στήλης. Αυτό έρχεται σε συνδυασμό με τον καθορισμό των μεγίστων και ελαχίστων των στηλών, που αναφέραμε στα προηγούμενα. Το ιστόγραμμα δείχνει πώς κατανέμονται οι τιμές των δεδομένων σε σχέση με τα μέγιστα και ελάχιστα που δόθηκαν. Αν η κατανομή έχει σχήμα U, σημαίνει ότι το διάστημα ελάχιστο-μέγιστο είναι μικρό. Αν έχει σχήμα Λ τότε το εύρος είναι μεγάλο. Αν η κατανομή είναι ομοιόμορφη, (εξαιρετική περίπτωση) τότε οι τιμές που δόθηκαν είναι οι κατάλληλες.



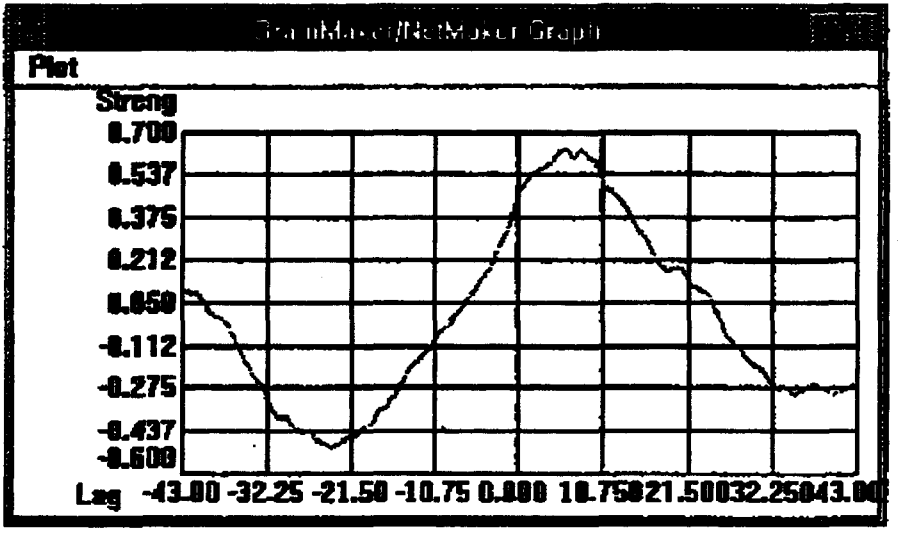
Στενό Εύρος Μέγιστων / Ελάχιστων

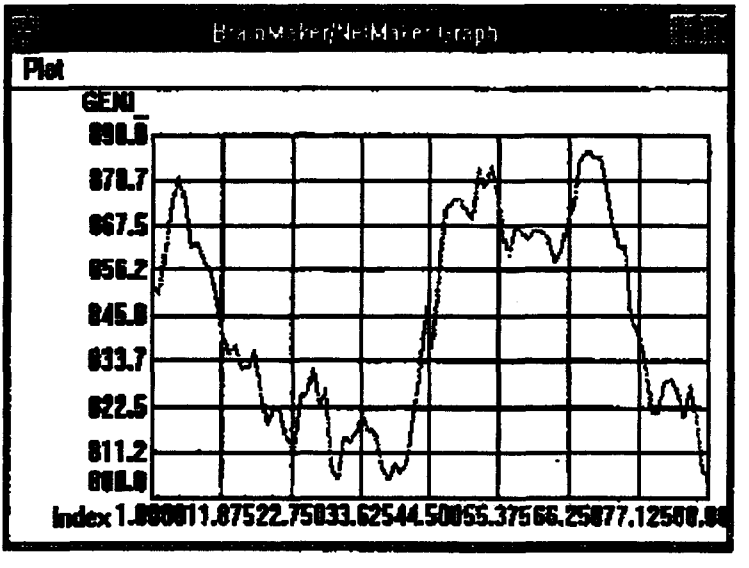
Πλατύ Εύρος Μέγιστων / Ελάχιστων



Θεωρητικά Τέλεια Κατανομή

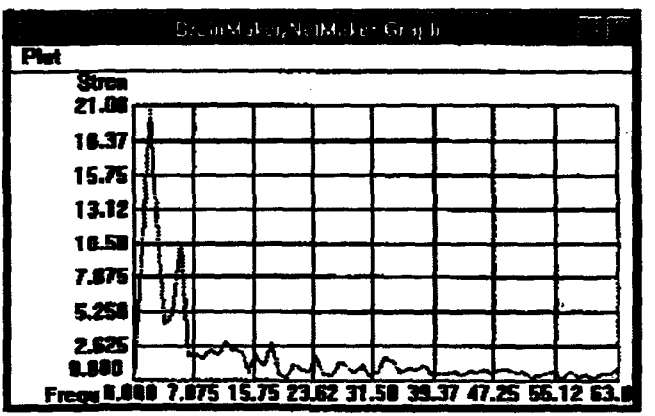
Με το γράφημα μπορούμε να ζητήσουμε τη γραφική παράσταση μίας η περισσοτέρων στηλών των δεδομένων. Με αυτόν τον τρόπο μπορούμε να έχουμε μία πρώτη εικόνα για τα δεδομένα που επεξεργαζόμαστε. Οι δυνατότητες που μας δίνονται είναι να επιλέξουμε γραμμικό ή λογαριθμικό διάγραμμα, πώς θα εμφανίζονται οι διάφορες στήλες κά.


Plot
Streng
0.700
0.537
0.375
0.212
0.050
-0.112
-0.275
-0.437
-0.600
Lag -43.00 -32.25 -21.50 -10.75 0.000 10.750 21.500 32.250 43.00


Plot
GEN
878.7
867.5
856.2
845.0
833.7
822.5
811.2
Index 1.000 11.875 22.750 33.625 44.500 55.375 66.250 77.125

2.5.2 Κυκλική Ανάλυση

Με το εργαλείο αυτό μπορούμε να βρούμε πιθανούς κύκλους στη στήλη που επεξεργαζόμαστε. Η μεγάλη χρησιμότητα αυτής της πληροφορίας είναι ότι ένα Νευρωνικό Δίκτυο εκπαιδεύεται πολύ πιο εύκολα όταν του δώσουμε να προβλέψει ένα γεγονός πάνω στον κύκλο του, δηλαδή πάνω στην περιοδικότητά του. Στο παρακάτω γράφημα κυκλικότητας, βλέπουμε ότι ο ισχυρότερος κύκλος εμφανίζεται 1.7644 φορές στο δείγμα και ο δεύτερος πιο ισχυρός 5.9534 φορές. Επειδή τα δεδομένα απεικονίζουν 128 χρονικές περιόδους, σημαίνει ότι ο ισχυρότερος κύκλος εμφανίζεται κάθε (128/1.7644=) 72.54 περιόδους, και ο άλλος κάθε (128/5.9534=) 21.5 περιόδους.



2.5.3 Συσχέτιση δεδομένων

Από τη στιγμή που έχουμε απόφασίσει για το τι θα προβλέπει το δίκτυο που κατασκευάζουμε, μεγάλη χρησιμότητα έχει να βρούμε ποια από τα δεδομένα (input) επηρεάζουν περισσότερο την πρόβλεψη. Την εργασία αυτή αναλαμβάνει η επιλογή του Net Maker, Data Correlator, που μας δείχνει πόσο ισχυρή είναι η συσχέτιση μεταξύ δύο στηλών.

Στη γραφική παράσταση βλέπουμε τη δύναμη της συσχέτισης στον κάθετο άξονα (Strength) και στον οριζόντιο τη χρονική διαφορά (Lag). Όταν η δύναμη είναι θετική, ερμηνεύεται ότι οι δύο στήλες κινούνται παράλληλα. Αρνητική δύναμη σημαίνει ότι οι στήλες κινούνται αντίθετα. Επίσης, όσο η δύναμη είναι πιο

κοντά στο 1 ή στο -1, τόσο πιο ισχυρή είναι η συσχέτιση. Ο οριζόντιος άξονας, που παριστάνει το lag. είναι χωρισμένος στη μέση, όπου βρίσκεται το 0 Το αρνητικό lag σημαίνει ότι η πρώτη στήλη προηγείται της δεύτερης, θετικό lag σημαίνει ότι η πρώτη στήλη έπεται της δεύτερης.

Με βάση τα παραπάνω, αυτό που ψάχνουμε στο γράφημα, είναι να βρούμε που βρίσκεται η πιο ψηλή κορυφή και να δούμε το lag της.

2.5.4 Στατιστικές Στήλες

Εκτός από τα χρήσιμα στοιχεία που μπορούμε να πάρουμε από τα προηγούμενα εργαλεία, έχουμε και τη δυνατότητα να δημιουργήσουμε νέες στήλες, που να περιέχουν διάφορα στατιστικά δεδομένα, τα οποία αναλύονται στα επόμενα:

- Διαφορά Στήλης

Μπορούμε να κατασκευάσουμε μία στήλη η οποία να είναι η διαφορά της τιμής της στήλης σε μία γραμμή από μία άλλη γραμμή. Για παράδειγμα, για ημερήσια δεδομένα, μπορούμε να δημιουργήσουμε μία στήλη που να περιέχει τη διαφορά της τιμής κλεισίματος από το κλείσιμο της προηγούμενης μέρας.

Την παραπάνω διαφορά μπορούμε να την εκφράσουμε με τρεις τρόπους:

- Σαν απλή διαφορά (απόλυτες τιμές)
- Επί τοις εκατό διαφορά (ποσοστιαία)
- Λογαριθμική διαφορά (διαφορά λογαρίθμων)

- Απόλυτη Τιμή

Με αυτή τη λειτουργία μπορούμε να κατασκευάσουμε μία νέα στήλη, η οποία να είναι η απόλυτη τιμή της αρχικής

2.5.5 Κινητοί Μέσοι Όροι / Κινητή Ενδιάμεση Τιμή

Με αυτήν τη λειτουργία, όπως επίσης και με τις λειτουργίες που θα αναφερθούν στην επόμενη παράγραφο, μπορούμε να προσθέσουμε και Τεχνική Ανάλυση στα δεδομένα του δικτύου . Συγκεκριμένα, μπορούμε να κατασκευάσουμε μία στήλη η οποία θα περιέχει τον κινητό μέσο όρο κάποιων περιόδων που ορίζουμε εμείς. Επίσης, έχουμε τη δυνατότητα να επιλέξουμε τον τρόπο υπολογισμού του κινητού αυτού μέσου, από τις παρακάτω επιλογές:

- Απλός Κινητός Μέσος
- Εκθετικός Κινητός Μέσος
- Γεωμετρικός Κινητός Μέσος
- Σταθμισμένος Κινητός Μέσος
- Κινητή Μέση Απόκλιση Τετραγώνου
-Κινητή Ασυμμετρία.

Τέλος, υπάρχει η δυνατότητα να ζητήσουμε μία νέα στήλη, που να περιέχει την κινητή ενδιάμεση τιμή (moving median price), για κάποιον αριθμό περιόδων που ζητάμε.

2.5.6 Δείκτες Τεχνικής Ανάλυσης

Εκτός από τους κινητούς μέσους όρους που αναφέρθηκαν στα προηγούμενα, με το Net Maker μπορούμε επίσης να κατασκευάσουμε και στήλες που να περιέχουν τεχνικούς δείκτες. Οι επιλογές που μπορούμε να κάνουμε όσον αφορά αυτούς τους δείκτες είναι οι εξής.

-MACD
-RSI
- Stochastic
- On-balance Volume

Δε θα αναφερθούμε καθόλου στην ερμηνεία των παραπάνω δεικτών. Κάτι τέτοιο θα είχε σαν απότέλεσμα να ξεφύγουμε από το σκοπό αυτού του κείμενο.

2.5.7 Λειτουργίες Αρχείων

Τέλος, είναι σκόπιμο να αναφέρουμε ότι το Net Maker προετοιμάζει τα αρχεία για να τα αξιοποιήσει το Brain Maker. Εδώ μπορούμε να καθορίσουμε τα εξής:

-Το μέγεθος του αρχείου ελέγχου (test file), ως ποσοστό των αρχικών γεγονότων.
- Αν θα χρησιμοποιήσουμε την αναδρομική ιδιότητα και για πόσες χρονικές περιόδους.
-Τα μέγιστα και τα ελάχιστα, σαν αριθμό Μέσων Απόκλίσεων Τετραγώνων.

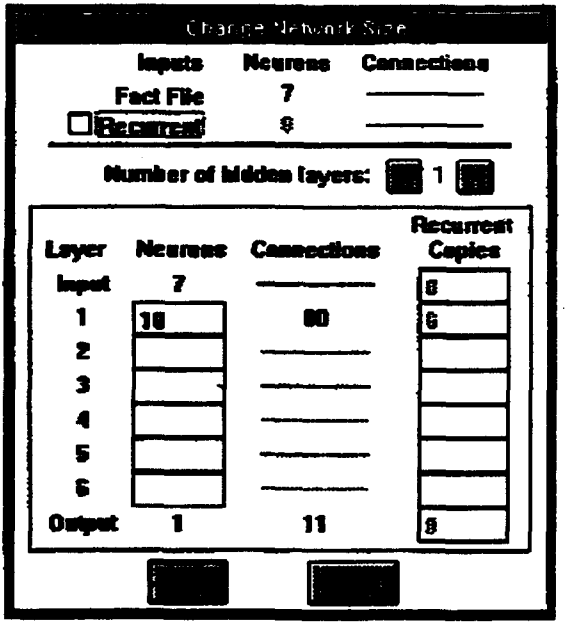
3. Brain Maker

Ο πυρήνας του νευρωνικού δικτύου αναλαμβάνει την εκπαίδευση, αλλά και την εκτέλεση του δικτύου. Κατά την εκπαίδευση, ο χρήστης μπορεί να καθορίσει διάφορες παραμέτρους. Επίσης, έχει τη δυνατότητα να παρακολουθεί την εκπαίδευση του δικτύου. Δεν είναι απαραίτητο ο καθορισμός των παραμέτρων να γίνει στην αρχή της εκπαίδευσης του δικτύου, αλλά και κατά τη διάρκεια αυτής.

Παράμετροι Αρχιτεκτονικής

3.1 Μέγεθος Δικτύου

Με τις παραμέτρους αυτές μπορούμε να καθορίσουμε τον αριθμό των ενδιάμεσων επιπέδων που θα χρησιμοποιήσουμε, τον αριθμό των νευρώνων σε κάθε επίπεδο, αν θα υπάρχει αναδρομή στο δίκτυο και πόσα "γεγονότα" πριν θα "θυμάται". Ένα δείγμα των παραπάνω, παρουσιάζεται στο διάγραμμα που ακολουθεί:



3.2 Συναρτήσεις μεταφοράς.

Εδώ μπορούμε να καθορίσουμε τη συνάρτηση εκείνη που εφαρμόζεται σε κάθε τιμή εισόδου ενός νευρώνα και το αποτέλεσμα της είναι η τιμή εξόδου του νευρώνα. Επίσης, μπορούμε να καθορίσουμε τη μεγαλύτερη και τη μικρότερη τιμή που μπορεί να εξαχθεί από το νευρώνα, καθώς και το κέντρο της συνάρτησης.

Οι συναρτήσεις μεταφοράς που είναι διαθέσιμες είναι:

- Linear Transfer Function
- Linear Threshold Transfer Function
- Step Transfer Function
- Sigmoid Transfer Function
- Gaussian Transfer Function

Συνήθως χρησιμοποιείται η σιγμοειδής συνάρτηση (Sigmoid). Όσον αφορά τα όρια της συνάρτησης, καλό είναι να αποδεχθούμε τις προεπιλεγμένες τιμές (default values) που μας παρέχει το Brain Maker.

3. 3 Θόρυβος στις Συνδέσεις

Με αυτήν την επιλογή μπορούμε να προσθέσουμε θόρυβο μεταξύ των συνδέσεων των νευρώνων. Ο θόρυβος που προστίθεται ακολουθεί την κατανομή Gauss με τυπική απόκλιση τετραγώνου τη μέση τιμή θορύβου, που εισάγει ο χρήστης. Όσο μεγαλύτερη είναι αυτή η τιμή, τόσο περισσότερο "βλάπτει" το δίκτυο. Όσο μικρότερη, τόσο λιγότερο. Αν η τιμή που εισάγουμε είναι μεγαλύτερη της μονάδας, τότε οποιαδήποτε εκπαίδευση είχαμε κάνει μέχρι τώρα χάνεται.

3.4 Αποκοπή Μικρών Συνδέσμων

Με τη δυνατότητα αυτή μπορούμε να αποκόψουμε τις συνδέσεις που είναι μικρότερες από κάποια τιμή, που ορίζουμε για κάθε επίπεδο. Αυτή τη λειτουργία τη χρησιμοποιούμε σε ένα εκπαιδευμένο δίκτυο και μετά από την

εκτέλεση αυτής, συνεχίζουμε την εκπαίδευση του δικτύου. Αν έχουμε καθορίσει να κλειδωθούν αυτές οι μικρές συνδέσεις στο μηδέν, τότε στη μετέπειτα εκπαίδευση του δικτύου θα παραμείνουν αυτές στο μηδέν (πρακτικά θα εξαφανιστούν). Η σημασία της λειτουργίας αυτής είναι ότι μπορεί να οδηγήσει σε ένα δίκτυο που να γενικεύει καλύτερα.

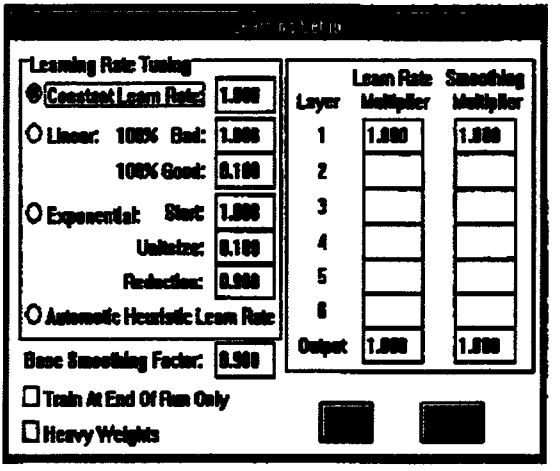
Παράμετροι Εκπαίδευσης

3.5. Καθορισμός Εκμάθησης

Με αυτές τις παραμέτρους επηρεάζουμε τον αλγόριθμο εκπαίδευσης του δικτύου. Αναφερόμαστε κυρίως στη τιμή μεταβολής των τιμών των συνδέσεων στην περίπτωση που βρεθεί ένα δεδομένο ταξινομημένο λάθος (ή αντίστοιχα σωστό). Μπορούμε να επιλέξουμε σταθερό ρυθμό εκπαίδευσης, γραμμικό, εκθετικό ή τον αυτόματου καθορισμού, ο οποίος βασίζεται σε ευρετικούς αλγόριθμους. Αν η τιμή του ρυθμού εκπαίδευσης είναι μεγάλη τότε το δίκτυο θα "τρελαθεί". Αν είναι πολύ μικρή, τότε ο χρόνος εκπαίδευσης του δικτύου μπορεί να αυξηθεί πολύ. Συνήθως χρησιμοποιείται ένας ρυθμός εκπαίδευσης γύρω στη μονάδα.

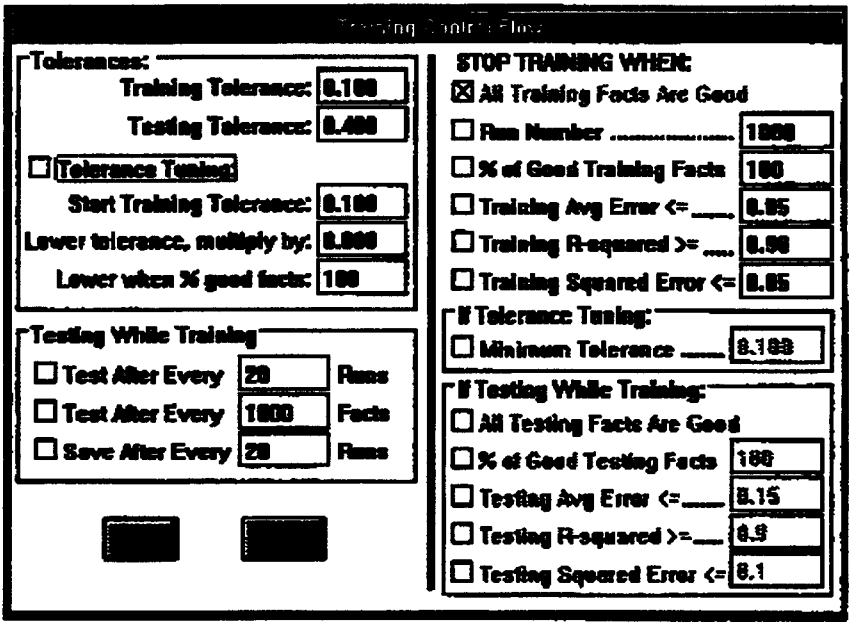
Από ότι φαίνεται από την παρακάτω οθόνη, μπορούμε να έχουμε διαφορετικούς ρυθμούς εκπαίδευσης για κάθε επίπεδο του δικτύου. Επίσης, μπορούμε να καθορίσουμε έτσι ώστε οι διάφορες μεταβολές στις τιμές των συνδέσεων να επιτελούνται στο τέλος της επεξεργασίας όλων των δεδομένων, αντί μετά από καθένα ξεχωριστά.

Τέλος με την επιλογή των "Βαρέων Βαρών" (Heavy Weights) μειώνουμε το ρυθμό μεταβολής των τιμών των συνδέσεων οι οποίες έχουν μεγάλη τιμή. Πολλές φορές αυτή η επιλογή έχει σαν απότέλεσμα να μπορέσει να εκπαιδευτεί ένα δίκτυο, που διαφορετικά δε θα μπορούσε. Η οθόνη με τις παραμέτρους είναι η ακόλουθη:



3.5.1 Καθορισμός Εκπαίδευσης

Με το σύνολο των παραμέτρων αυτών μπορούμε να καθορίσουμε το επίπεδο ανοχής που θέλουμε για την εκπαίδευση και για τον έλεγχο του δικτύου. Το επίπεδο ανοχής καθορίζει το εύρος εκείνο των τιμών των εξαγομένων αποτελεσμάτων που θεωρείται αποδεκτό. Μπορούμε επίσης να καθορίσουμε το χρόνο που θα σταματήσει η εκπαίδευση του δικτύου. Παρατηρούμε ότι είναι δυνατό να καθορισθεί διαφορετικό επίπεδο ανοχής για την εκπαίδευση και διαφορετικό για τον έλεγχο του δικτύου. Συνήθως στην αρχή της εκπαίδευσης το επίπεδο ανοχής είναι σχετικά μεγάλο, το οποίο όμως καθώς προοδεύει το δίκτυο ελαττώνεται. Για την ακρίβεια, ελαττώνεται κατά ένα ποσοστό που ορίζουμε όταν τα ορθά γεγονότα ενός "τρεξίματος" (run) ξεπεράσουν κάποιο ποσοστό των συνολικών.



Ανάλογα με τις παραμέτρους των επιπέδων ανοχής καθορίζεται ο χρόνος που θα σταματήσει η εκπαίδευση του δικτύου. Οι συνήθεις επιλογές είναι οι ακόλουθες:

- Όταν όλα τα προς εκπαίδευση γεγονότα ή ένα ποσοστό τους είναι ορθό.
- Όταν το επίπεδο ανοχής φθάσει στο επιθυμητό.
- Όταν όλα τα δεδομένα ελέγχου ή κάποιο ποσοστό τους είναι ορθό.

Επίσης, από αυτό το σημείο, καθορίζουμε αν το δίκτυο θα ελέγχεται κατά τη διάρκεια της εκπαίδευσής του και αν θα σώζεται αυτόματα μετά από κάποιον αριθμό κύκλων εκπαίδευσης (runs).

3.5.2 Θόρυβος στα Δεδομένα

Με το σύνολο αυτών των παραμέτρων μπορούμε να καθορίσουμε αν θέλουμε να υπάρχει θόρυβος στα δεδομένα. Θόρυβος είναι, μικρές αποκλίσεις της πραγματικής τιμής. Μπορούμε να επιλέξουμε να υπάρχει θόρυβος στα δεδομένα εκπαίδευσης, ελέγχου, πρόβλεψης ή σε κάποιο συνδυασμό από τα παραπάνω.

Έχει παρατηρηθεί από μερικούς χρήστες των νευρωνικών δικτύων ότι όταν έχουμε αριθμητικά δεδομένα, ο θόρυβος οδηγεί σε δίκτυα που μπορούν να γενικεύουν καλύτερα. Επίσης, όταν το μέγεθος του δείγματος που έχουμε στη

διάθεση μας είναι μικρό, καλό είναι να προσθέσουμε θόρυβο. Όμως, η ύπαρξη θορύβου καθυστερεί την εκπαίδευση του δικτύου. Μία μέση λύση είναι να αφήσουμε το δίκτυο να εκπαιδευτεί χωρίς θόρυβο μέχρι να φθάσει σε κάποιο ικανοποιητικό σημείο και τότε να του προσθέσουμε θόρυβο και να το αφήσουμε να συνεχίσει την εκπαίδευση του.

3.5.3 Παράμετροι Αναδρομής

Με αυτές τις παραμέτρους καθορίζεται ποια δεδομένα θα χρησιμοποιηθούν για την αναδρομή των πρώτων γεγονότων, όταν προηγούμενα στοιχεία δεν είναι διαθέσιμα, όπως επίσης αν τα δεδομένα της αναδρομής θα ληφθούν από την προβλεπόμενη τιμή ή από την πραγματική τιμή. Βέβαια, τις παραμέτρους αυτές τις καθορίζουμε μόνο στην περίπτωση που έχουμε επιλέξει αναδρομή στην αρχιτεκτονική του δικτύου.

3.5.4 Μεταβολή Μεγέθους Δικτύου

Μία άλλη δυνατότητα που προσφέρει το σύστημα είναι η αυτόματη προσθήκη *νευρώνων στα εσωτερικά επίπεδα του δικτύου κατά τη διάρκεια της* εκπαίδευσης.

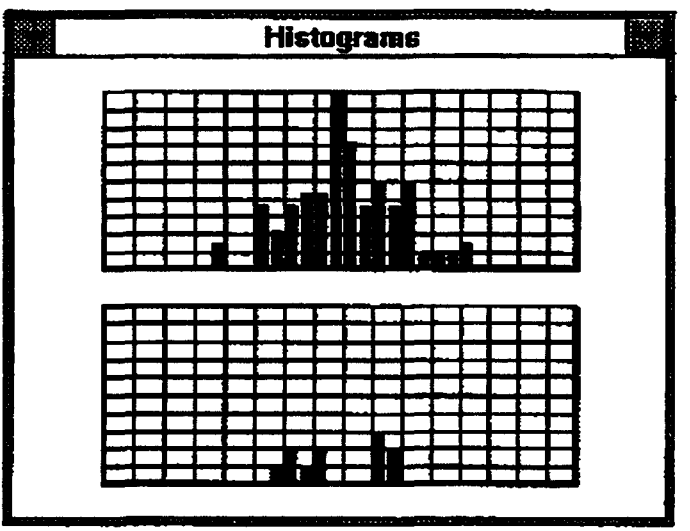
Αντίστοιχα, καθορίζεται η αυτόματη αφαίρεση νευρώνων κατά την εκπαίδευση. Για την υλοποίηση των παραπάνω, θέτουμε τις συνθήκες που πρέπει να ισχύουν για να εκτελεστούν οι ενέργειες αυτές και το εσωτερικό εκείνο επίπεδο όπου θα προστεθούν οι νευρώνες.

Εξέλιξη Εκπαίδευσης

Εκτός όμως από τα εργαλεία προσδιορισμού της αρχιτεκτονικής και του καθορισμού των παραμέτρων εκπαίδευσης, το Brain Maker παρέχει και εργαλεία για να παρακολουθείται η εξέλιξη της εκπαίδευσης του δικτύου.

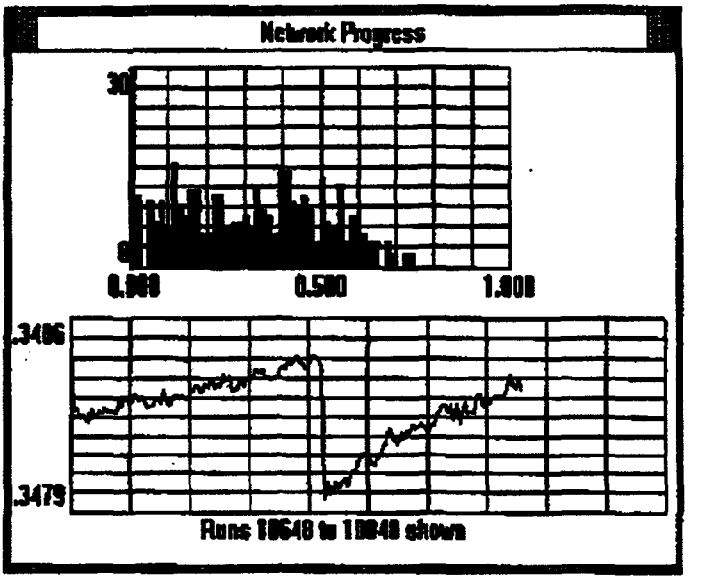
3.6 Ιστογράμματα

Μία ένδειξη της προόδου του δικτύου είναι τα ιστογράμματα που παρουσιάζουν τις συνδέσεις μεταξύ των επιπέδων. Ανάλυση για αυτά τα ιστογράμματα έχει γίνει στα προηγούμενα. Εδώ απλώς θα δούμε πώς ακριβώς εμφανίζονται αυτά στο Brain Maker,



3.6.1.Πρόοδος του Δικτύου

Παρατηρούμε την κατανομή των λαθών στον τελευταίο κύκλο, καθώς και την εξέλιξη της τιμής της τετραγωνικής ρίζας του μέσου τετραγωνικού σφάλματος (RMS Error). Το ζητούμενο είναι τα λάθη να κατανέμονται όσο το δυνατό πλησιέστερα στο μηδέν και το RMS Error να μειώνεται.



3.6.2 Ανάλυση Βασικής Οθόνης

Εκτός όμως από τα διαγράμματα, παρατηρούμε και τα στοιχεία που φαίνονται στην οθόνη, όσον αφορά τα ορθά και τα λανθασμένα δεδομένα. Τέτοια στοιχεία είναι :

- Ο αριθμός ορθών δεδομένων στον τρέχοντα κύκλο (run).
- Αριθμός λανθασμένων δεδομένων στον τρέχοντα κύκλο.
- Αριθμός ορθών δεδομένων στον προηγούμενο κύκλο.
- Αριθμός λανθασμένων δεδομένων στον προηγούμενο κύκλο.

Από αυτά μπορούμε να δούμε την εξέλιξη των ορθών δεδομένων.

3.6.3 Στατιστικά

Άλλες ενέργειες που μπορούμε να ζητήσουμε από τον πυρήνα του νευρωνικού δικτύου, είναι η φύλαξη στατιστικών στοιχείων για μετέπειτα επεξεργασία. Με αυτόν τον τρόπο, μπορούμε να φυλάξουμε σε διαφορετικό αρχείο τα αποτελέσματα κάθε κύκλου εκπαίδευσης και σε διαφορετικό αυτά του ελέγχου.

Μερικές από τις πληροφορίες που φυλάσσονται σε αυτά τα αρχεία είναι οι:

- Αριθμός κύκλων
- Συνολικός αριθμός δεδομένων που έχουν επεξεργαστεί
- Αριθμός ορθών δεδομένων
- Αριθμός λανθασμένων δεδομένων
- Πόσοι από τους νευρώνες εξόδου έδωσαν λάθος αποτελέσματα
- Συνολικός αριθμός λανθασμένων δεδομένων
- Ρυθμός εκπαίδευσης
- Επίπεδο ανοχής
- Μέσο Λάθος
- RMS Error

Αυτές τις πληροφορίες μπορούμε να τις επεξεργαστούμε μέσω του Net Maker που αναφέρθηκε στα προηγούμενα. Η χρησιμότητα αυτών των πληροφοριών έγκειται στην επιλογή του δικτύου μέγιστης απόδοσης, το οποίο δεν είναι κατ' ανάγκη το δίκτυο που προέκυψε στη φάση της εκπαίδευσης.

3.6.4 Ενέργειες στο Δίκτυο

Τέλος, όπως είναι αναμενόμενο, ο πυρήνας του νευρωνικού δικτύου παρέχει και τις απαραίτητες εντολές για την εκπαίδευση, τον έλεγχο και την εκτέλεση του δικτύου. Αξιοσημείωτο είναι ότι παρέχεται η δυνατότητα να εκτελούμε ή να ελέγχουμε το δίκτυο με ένα-ένα δεδομένο. Με αυτόν τρόπο και τη δυνατότητα να επηρεάσουμε την τιμή ενός στοιχείου εισόδου, μπορούμε να ελέγξουμε και να κατανοήσουμε καλύτερα το δίκτυο.

3.7 Competitor

Τον Competitor τον χρησιμοποιούμε για να ταξινομήσουμε μία σειρά από δεδομένα, βάσει της χρονολογικής εξέλιξης της τιμής των χαρακτηριστικών τους (δεικτών), για μερικές περιόδους. Αυτά τα στοιχεία τα εισάγουμε στον Competitor, ο οποίος αναλαμβάνει να εκπαιδεύσει κατάλληλα το δίκτυο.

Το τελευταίο βήμα είναι να πάρουμε τις τιμές των δεικτών σήμερα και να ζητήσουμε από τον Competitor να ταξινομήσει τα δεδομένα. Η σειρά των δεδομένων, έτσι όπως προβλέπεται από το νευρωνικό δίκτυο παρουσιάζεται στην οθόνη μας.

3.8 Net Checker

Ο Net Checker είναι ένα εργαλείο που παρέχει το PROFILE Neural Applications για τον έλεγχο του δικτύου. Προσπαθεί να εντοπίσει λάθη στη δομή και ελέγχει για αντιφατικά δεδομένα στα δεδομένα εκπαίδευσης.

Μερικά από τα λάθη που αναγνωρίζει είναι :

- Μη αναγνωρίσιμες γραμμές
- Κενές γραμμές στο τέλος ή στη μέση των αρχείων
- Περιττά στοιχεία σε γραμμές που παριστάνουν γεγονότα
- Ελλιπή στοιχεία σε γραμμές που παριστάνουν γεγονότα
- Αντιφατικά ή σχεδόν αντιφατικά δεδομένα

Αντιφατικά δεδομένα είναι αυτά για τα οποία εξάγεται διαφορετικό αποτέλεσμα με τα ίδια ή παρόμοια στοιχεία εισόδου.

3.9 Genetic Training Options

Τα Genetic Training Options (GTOs) μπορούν να απλοποιήσουν την εργασία μας, αλλά και να καταλήξουν σε αποδοτικότερα δίκτυα, μπορούν να χρησιμοποιηθούν στην αρχή της εκπαίδευσης του δικτύου με δύο τρόπους

- Βοηθώντας μας να προσδιορίσουμε τις καλύτερες παραμέτρους εκπαίδευσης για το Δίκτυο και
- Βοηθώντας μας να ξεκινήσει η εκπαίδευσή μας όχι με τυχαία βάρη αλλά με τέτοια βάρη ώστε να πάρουμε πιο γρήγορα τα αποτελέσματά μας.

Εκτός όμως από τη βοήθεια που μας προσφέρουν στην αρχή της εκπαίδευσης, μπορούν να χρησιμοποιηθούν και μετά το πέρας αυτής, έτσι ώστε να καταλήξουμε σε αποδοτικότερο Δίκτυο.

Για να το πετύχουμε αυτό, παίρνουμε το εκπαιδευμένο Δίκτυο (ή γενικότερα το Δίκτυο που μας δίνει μέχρι στιγμής τα καλύτερα αποτελέσματα) και δίνουμε εντολή στα GTOs να δοκιμάσουν "κοντινά" Δίκτυα και να μας τα παρουσιάσουν.

Τότε τα GTOs ανάλογα με τις παραμέτρους που έχουμε καθορίσει, εφαρμόζουν γενετικούς αλγόριθμους Μετάλλαξης και Διασταύρωσης. Μέσω αυτών κατασκευάζουν και ελέγχουν εναλλακτικά Δίκτυα και μας προτείνουν τα καλύτερα, βάσει κριτηρίων.

ΠΑΡΑΡΤΗΜΑ Γ
ΠΙΝΑΚΕΣ ΣΥΧΝΟΤΗΤΩΝ

ΠΙΝΑΚΑΣ Π1: ΜΕΣΟΣ, ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ ΚΑΙ ΣΥΝΤΕΛΕΣΤΗΣ ΑΣΥΜΜΕΤΡΙΑΣ ΤΗΣ ΜΕΤΑΒΛΗΤΗΣ «ΗΛΙΚΙΑ»

	ΠΛΗΘΟΣ	ΜΕΣΟΣ	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ	ΔΙΑΚΥΜΑΝΣΗ	ΣΥΝΤΕΛΕΣΤΗΣ ΑΣΥΜΜΕΤΡΙΑΣ	ΚΥΡΤΩΣΗ
ΗΛΙΚΙΑ	660	55.061	8.097	65.562	- 0.163	0.190

ΠΙΝΑΚΑΣ Π2: ΑΡΤΗΡΙΑ "RCA"

ΠΟΣΟΣΤΟ ΕΜΦΡΑΞΗΣ	0	(0 – 70)	[70 – 100]
ΘΕΣΗ	ΑΡΙΘΜΟΣ ΑΤΟΜΩΝ		
PROX	376 (57%)	103 (15,6%)	181 (27,9%)
MID	445 (67,4%)	99 (15%)	116 (17,6%)
DIST	574 (87%)	43 (6,5%)	43 (6,5%)
PD	614 (93%)	15 (2,3%)	31 (4,7%)

ΠΙΝΑΚΑΣ Π3: ΜΕΣΟΣ, ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ ΚΑΙ ΣΥΝΤΕΛΕΣΤΗΣ ΑΣΥΜΜΕΤΡΙΑΣ

ΑΡΤΗΡΙΑ ΘΕΣΗ	ΜΕΣΟΣ	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ	ΣΥΝΤΕΛΕΣΤΗΣ ΑΣΥΜΜΕΤΡΙΑΣ
PROX	0,3181	0,4088	0,741
MID	0,2238	0,3529	1,205
DIST	0,085	0,2386	2,762
PD	0,0477	0,1833	3,853

ΠΙΝΑΚΑΣ Π4: ΑΡΤΗΡΙΑ "LAD"

ΠΟΣΟΣΤΟ ΕΜΦΡΑΞΗΣ	0	(0 – 70)	[70 – 100]
ΘΕΣΗ	ΑΡΙΘΜΟΣ ΑΤΟΜΩΝ		
PROX	309 (46,8%)	90 (13,6%)	261 (39,5%)
MID	353 (53,5%)	76 (11,5%)	231 (35%)
1D	531 (80,5%)	38 (5,8%)	91 (13,8%)
APIC	603 (91,4%)	28 (4,2%)	29 (4,4%)
2D	628 (95,2%)	7 (1,1%)	25 (3,8%)

ΠΙΝΑΚΑΣ Π5 : ΜΕΣΟΣ, ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ ΚΑΙ ΣΥΝΤΕΛΕΣΤΗΣ ΑΣΥΜΜΕΤΡΙΑΣ

ΘΕΣΗ	ΜΕΣΟΣ	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ	ΣΥΝΤΕΛΕΣΤΗΣ ΑΣΥΜΜΕΤΡΙΑΣ
PROX	0,4189	0,4294	0,230
MID	0,3661	0,4215	0,437
1D	0,1437	0,3061	1,815
APIC	0,056	0,1959	3,568
2D	0,0384	0,1752	4,526

ΠΙΝΑΚΑΣ Π6: ΑΡΤΗΡΙΑ "LCX"

ΠΟΣΟΣΤΟ ΕΜΦΡΑΞΗΣ	0	(0 – 70)	[70 – 100]
ΘΕΣΗ	ΑΡΙΘΜΟΣ ΑΤΟΜΩΝ		
PROX	400 (60,6%)	93 (14,1%)	167 (25,3%)
DIST	487 (73,8%)	71 (10,8%)	102 (15,5%)
OM	554(83,9%)	18 (2.7%)	88 (13,3%)
PL	623 (94,4%)	16 (2,4%)	21 (3,2%)
PD	653 (98,9%)	0 (0%)	7 (1.0%)

ΠΙΝΑΚΑΣ Π7: ΜΕΣΟΣ, ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ ΚΑΙ ΣΥΝΤΕΛΕΣΤΗΣ ΑΣΥΜΜΕΤΡΙΑΣ

ΘΕΣΗ	ΜΕΣΟΣ	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ	ΣΥΝΤΕΛΕΣΤΗΣ ΑΣΥΜΜΕΤΡΙΑΣ
PROX	0,2845	0,3884	0,851
DIST	0,1817	0,3286	1,479
OM	0,1305	0,3070	2,030
PL	0,03947	0,1720	4,493
PD	0,0089	0,08739	9,903

ΠΙΝΑΚΑΣ Π8: ΣΤΕΛΕΧΟΣ "LMCA"

ΠΟΣΟΣΤΟ ΕΜΦΡΑΞΗΣ	0	(0 – 70)	[70 – 100]
ΑΡΙΘΜΟΣ ΑΤΟΜΩΝ	613 (92,9%)	31(4,7%)	16(2,4%)

ΜΕΣΟΣ	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ	ΣΥΝΤΕΛΕΣΤΗΣ ΑΣΥΜΜΕΤΡΙΑΣ
0,0394	0,1530	4,045

ΠΙΝΑΚΑΣ Π9: ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ

Variable	B	S.E.	Wald	Df	Sig	R	Exp (B)
Sex (1)	145,5522	4095,9870	,0013	1	,9717	,0000	1,631E+63
Age	,1002	,1036	,9356	1	,3334	,0000	1,1054
EF	-62,8281	17,1395	13,4373	1	,0002	-,2271	,0000
Lipids (1)	5,6508	546,9188	,0001	1	,9918	,0000	284,5050
HBP (1)	14,7281	951,1802	,0002	1	,9876	,0000	2490812,9
Smoke (1)	43,3029	863,5402	,0025	1	,9600	,0000	6400E+18
Diabetes (1)	34,2412	2906,6387	,0001	1	,9906	,0000	7,426E+14
Obecity (1)	-94,7177	1428,0343	,0044	1	,9471	,0000	,0000
Family (1)	-17,4725	429,0831	,0017	1	,9675	,0000	,0000
Sedentar (1)	-68,4690	1298,2148	,0028	1	,9579	,0000	,0000
Type A	-56,8470	2084,2418	,0007	1	,9782	,0000	,0000
Int_1	-12,6491	745,1286	,0003	1	,9865	,0000	,0000
Int_2	-41,1458	759,5388	,0029	1	,9568	,0000	,0000
Int_3	11,0579	410,5773	,0007	1	,9785	,0000	63444,976
Int_4	-49,3539	816,6405	,0007	1	,9518	,0000	,0000
Int_5	-18,9634	2252,8670	,0001	1	,9933	,0000	,0000
Int_6	-33,3524	721,3408	,00021	1	,9631	,0000	,0000
Int_7	-39,2205	1486,3601	,0007	1	,9789	,0000	,0000
Int_8	-21,7288	1174,0222	,0003	1	,9852	,0000	,0000
Int_9	-7,9114	3,2504	5,9242	1	,0149	-,1330	,0004
Int_10	3,6423	4,1770	,7603	1	,3832	,0000	38,1785
Int_11	-2,1690	3,7232	,3394	1	,5602	,0000	,1143
Int_12	-1,5169	2,3118	,4305	1	,5117	,0000	,2194
Int_13	-2,1605	3,8337	,3176	1	,5731	,0000	,1153
Int_14	-21,3621	721,2831	,0009	1	,9764	,0000	,0000
Int_15	29,8577	571,9146	,0027	1	,9584	,0000	9,269E+12
Int_16	-4,8012	4,0006	1,4403	1	,2301	,0000	,0082
Int_17	22,0558	202,7334	,0118	1	,9134	,0000	3,791E+09
Int_18	7,0758	2,9518	5,7460	1	,0165	,1300	1183,0043
Int_19	18,227	101,6892	,0343	1	,8532	,0000	149480090
Int_20	13,0235	389,7229	,0011	1	,9733	,0000	452937,98
Int_21	-22,2153	1590,4471	,0002	1	,9889	,0000	,0000
Int_22	3,8184	202,6945	,0004	1	,9850	,0000	45,5332
Int_23	-,9136	3,5863	,0649	1	,7989	,0000	,4011
Int_24	-3,6042	16,0289	,0506	1	,8221	,0000	,0272
Int_25	-9,9734	282,3817	,0012	1	,9718	,0000	,0000
Int_26	-48,1817	839,1180	,0033	1	,9542	,0000	,0000
Int_27	14,1330	202,6786	,0049	1	,9444	,0000	1373659,9
Int_28	8,9621	202,7251	,0020	1	,9647	,0000	7801,5916
Int_29	-13,7919	913,5921	,0002	1	,9880	,0000	,0000
Int_30	-4,1137	1774,1548	,0000	1	,9981	,0000	,0163
Int_31	,8078	5,0690	,0254	1	,8734	,0000	2,2430
Int_32	37,4050	750,8119	,0025	1	,9603	,0000	1,757E+16
Int_33	58,8224	1657,8405	,0013	1	,9717	,0000	3,518E+25
Int_34	23,1563	721,6918	,0010	1	,9744	,0000	1,139E+10
Int_35	3,3517	660,9042	,0000	1	,9960	,0000	28,5514
Int_36	89,2633	847,3175	,0111	1	,9161	,0000	5,842E+38
Constant	32,8225	3380,5034	,0001	1	,9923		

1.. SEX
AGE
EF
LIPIDS
HBP
SMOKE
DIABETES
OBECITY
FAMILY
SEDENTAR
TYPEA
LIPIDS * SEX
HBP * SEX
SEX * SMOKE
DIABETES + SEX
OBECITY * SEX
FAMILY * SEX
SEDENTAR * SEX
SEX * TYPEA
HBP * LIPIDS
LIPIDS * SMOKE
DIABETES * LIPIDS
LIPIDS * OBECITY
FAMILY * LIPIDS
LIPIDS * SEDENTAR
LIPIDS * TYPEA
HBP * SMOKE
DIABETES * HBP
HBP * OBECITY
FAMILY * HBP
HBP * SEDENTAR
HBP * TYPEA
DIABETES * SMOKE
OBECITY * SMOKE
FAMILY * SMOKE
SEDENTAR * SMOKE
SMOKE * TYPEA
DIABETES * OBECITY
DIABETES * FAMILY
DIABETES * SEDENTAR
DIABETES * TYPEA
FAMILY * OBECITY
OBECITY * SEDENTAR
OBECITY * TYPEA
FAMILY * SEDENTAR
FAMILY * TYPEA
SEDENTAR * TYPEA

ΠΑΡΑΡΤΗΜΑ Δ

ΙΑΤΡΙΚΟ ΔΕΛΤΙΟ

NAME: ΑΔΑΜΟΥΤΗΣ ΔΗΜΗΤΡΗΣ GROUP: CAD+ ISI CODE:

ADDRESS: ΚΑΛΙΣΟΥΡΑΣ 21 ΠΕΡΙΣΤΕΡΙ PHONE: 57 1 301

JOB: ΥΔΡ/ΧΟΣ ΑΝΟΥΚΑΔΟΠΟΣ MARITAL STATUS: M KIDS: 3

H: 168 W: 91 BSA: 2.006 BMI: 32.2 AGE: 63

P	P	HISTORY		STI		CATH		ANGIO	
	32	MI (?)		RR	<45	date: 28.5.86		RCA PROX	100
				QT	395	film: 4004		MID	-
03	34			Q2	405	cath: 5070		DIST	-
04	35	ANGINA: 5		EI	370	ESP	125	PD	
05	36	EFFORT	5	1-2	340	EDP	18	LMCA	
	37	SPONTANEOUS	-	ER	71	ER	70	LAD PROX	75
		NOCTURNAL	5	Q2-I	523	SEP	16	MID	50
		PERSONALITY		PEP-I	144	EDA	80	1-D	-
09		(E)	56.9	ET-I	379	EDI.	115	APIC	-
		J	403	QT-I	512	ESA	44	2-D	
		(S)	50.2	QTc	430	ESL	10.3	LCX PROX	-
12		Σ	426	Q-1	65	EDVI	106.0	DIST	-
13	44	COS	67,6	Q-1c	91	ESVI	55.5	OM	-
14	45	EF	52,7	IVC	50	EF	47.6	PL	-
	46	CCC	3,6	IVCc	52	CI	3.528	PD	-
16		RF: 4		PEP/ET	.397	SWI	73	RCA 4/4	28.2
17	48	(1) LIPIDS		QT/Q2	.775	CCNI	.692	LM -	-
		✗ HIGE BP		ET/IVC	5.8	Vcf	1.35	LAD -	13
		(2) SMOKE		DYSPNOEA: 1		MNSER	1.762	LCX -	-
20	51	✗ DIABETES		ANGINA: 2		CRWI	707	COS: 41.8	L
21		(3) OBESITY		CTR: 43		AKINESIA		VD: 2	
	53	(6) SEDENTARY		ECG : 4		ANEVRYSM		CCC 455	
23	54	✗ FAMILY		Q SCORE	3	EDCT	63 -32	LAD → R:4	
24	55	✗ URIKEMIA		R SCORE	56	ESCT	75 -52		
		✗ TYPE A		>R	51.5	C R	.32 131		

		COMMENTS/FOLLOW-UP
27	58	
	60	
	61	

% SHORTENING

1/2 AXES	UPPER	LOWER	L
BASAL	51.8	18.0	
MIDDLE	40.5	28.9	
APICAL	25.9	72.0	

DISPLACEMENT : - 4.5

VENTRICULOGRAPHY SCORE: 7	NORMAL	HYPO MOD	HYPO SEV	AKI	DYS	ANE
ANTEROBASAL	✓					
ANTEROLATERAL	✓					
APICAL		✓				
DIAPHRAGMATIC	✓					
POSTEROBASAL		✓				

ANEURYSM mm

ml/m²

% ESVI

ESA-A:

ΔEF:

ET loss:

ΒΙΒΛΙΟΓΡΑΦΙΑ

Akay, M., Welkowitz W., (1993), "Acoustical detection of coronary occlusions using neural networks", *J. Biomed. Eng., **15**, no.6, 469-473.*

Akay, M. (1995) Detection of coronary atery disease using fuzzy neural networks. Proceedings of the international joint conference of CFSA/IFIS/SOFT'95 on fuzzy theory and application. *pp 543-8.*

Aleksander Igor, Burnett P.,, "Thinking Machines, the Search for A1", *(NY: Alfred A. Knopf, 1987).*

Αλεξανδρόπουλος Α., Παλιατσος Α., Σοφιανός Γ. (1995):Βασικά θέματα Αριθμητικής Ανάλυσης, Σύγχρονη Εκδοτική, Αθήνα 1995.

AMA (1991) . The Heart, *The family Medical*, Library C.B. Clayman (Ed.) Doling Kindersley Ltd, 1991. Ελληνική μετάφραση από τις εκδόσεις Μανιατέα 1993.

Amendolia, S. R., Bertolucci, E., Biadi, O., Bottigli U., Caravelli, P., Fantacci, M. E., Fidecaro, E., Mariani, M., Messineo, A., Rosso,V., Stefanini, A. (1993). Neural network expert system for screening coronary heart disease. *Physica Medica. **Vol. 9** ,no.1 pp. 13-17.*

Anderson, G., T. Zeng, J., Clark, M., R., Wyeth, R., P., the POSCH Group. (1993). Proceedings of the annual conference on engineering in medicine end Biology. *Vol **15** pp 257-258.*

Bernardo AM., (1993) "Toward psychosocial risk profile in CHD: Analysis of the toxic components of the Type A pattern of conduct" *DAI-C **54/01** p.213, Spring 1993.*

Bologna, G., Rida, A., Pellegrini, C. (1997). Intelligent assistance for coronary heart disease diagnosis: A comparison study. *Artificial intelligence in medicine. 6th conference on artificial intelligence in medicine Europe, AIME p.199-210.*

Bowerman B.L., O'Connell R.T., "Linear Statistics Models: An Applied Approach", *PWS-Kent, Boston, 1992*

Breun T. and Arnsen E. "Selecting test factors: A comparison of discriminant analyses, logistic regression and Cox' s regression model using data fian the TROMS Heartstudy statistics in medicine *Vol. 4, 413 – 423 (1985)*

Byth, K., McLachlang, G. Logistic regression compared to normal discrimination for non-normal populations. *Austral. J. Stat. 1980; 22 (2) : 188-196.*

Cacoullos T., "Discriminant analysis and applications" *ed Academic Press Inc. 1973.*

Caldwell R.B (1994). Design of Neural Network-based Financial Forecasting Systems: Data selection and data processing, *Neurovest Journal, Sept.,pp12-17.*

Campbell M. J. – Machin D., "A commonsense approach" *Medical statistics ed. John Wiley and Sons.*

Carpenter , W., Hoffman, M (1995). Training backdrop neural networks. *AI Expert,* Mar, 30-33

Cattell R.B (1966), "The meaning and strategic use of factor analysis" in R.B. Cattell (ed.) *Handbook of multivariate experimental psychology, Rand McNally, Chicago.*

Chia-Lun Chao, Por-Jau-Huang, Chau-Chung Wu, Su-Jane Shen, Poon-Ung Chieng, Cheng-Tau Su, Yuang-Teh Lee. "Correlation between quantitative severity of stress Thallium-201 myocardial perfusion defect and severity of coronary stenosis". *J Formos Med Assoc. 1996; **95**: 105-109.*

Cybenko G. (1989), "Approximation by superposition of a sigmoid function", *Mathematics of Control Signals and Systems* **2,** *pp.303-314.*

Denker J. S., Wittner B.S, "Network generality, training required and precision required", IEEE conf. on Neural Information Processing Systems, November 1987.

Denolet J., Sys SU., Stroobant N., Rombouts H., Gillebert TC., Brutsaert DL.(1996) "Personality as independent predictor of long-term mortality in patients with CHD", *Lancet. 1996 Feb.17; 347(8999): 417-21.*

Derby C.A. (1990) "The use of population controls in a study of arteriographically-defined CAD (arteriosclerosis)" *DAI-B **50/07** p.2871, Jan 1990.*

Duda R.O and Hart P.E (1973) "Pattern classification and Scene Analysis", *J.Wiley & Sons, New York.*

Durrenberg RE. (1984) "CHD and Type A behavior pattern:Discriminamt analyses of selected variables in a rural sample" *MAI **22/01** p.80, Spring 1984.*

Edwards KL. (1996) "Genetic Epidemiology of the insulin resistance syndrome: A multivariate approach (CHD, diabetes mellitus)", *DAI-B* ***57/07****, p.4328 Jan 1997.*

Edwards KL., Austin MA., Newman B., Mayer E., Krauss RM., Selby JV. (1994) " Multivariate analysis of the insulin resistance syndrome in women" *Arterioscler-Thromb. 1994 Dec;* ***14(12)*** *1940-5*

Efron B. (1975) "The efficiency of Logistic regression compared to normal discriminant analysis". *JASA,* ***70****. 892-898.*

Fang CL. (1986) "Comparison of selected CHD indices in runners, weightlifters and controls (blood lipids, training mortality, pulse wave velocity, coronary)". *DAI-B* ***47/04*** *p.1437, Oct 1986.*

Fisher, R. A., "The use of multiple measurements in taxonomic problems". *Annals of Augenics, 1936,* ***7****, 179-188.*

Gary J. Martin., *The American Medical Association (A.M.A) 1991*

Garfagnini, A., Devoto, G., Posselini, P., Boggiano, P., Venturini, M.(1995) Relationship between HDL-cholesterol and apolipoprotein A1 and the severity of coronary artery disease. *European Heart Journal 1995,* ***16****: 465-470.*

Gensini GG. A more meaningful scoring system for determing the severity of coronary artery decease. *Am J. Cardiol 1983;* ***51****: 606.*

Georgiakodis F.(1986) "Discriminant techniques for medical prognosis following severe head injury" *A thesis submitted to the University of Bradford for the degree of Doctor of Philosophy, Bradford 1986.*

Γεωργιακώδης Μ., και Γεωργιακώδης Φ. (1993) Ειδικά Θέματα Γραμμικής Άλγεβρας με Εφαρμογές στη Στατιστική, Εκδόσεις βιβλίων Μ. Βαρβαρήγου, Πειραιάς 1993.

Godsland IF., Leyva F., Walton C., Worthington M., Stevenson JC. (1998), "Associations of smoking , alcohol and physical activity with risk factors for CHD and diabetes in the first follow-up cohort of the Heart Disease and Diabetes Risk Indicators in a Screened Cohort study", *J-Intern-Med. (1998) Jul;* ***244(1)*** *33-41*

Goreska B., Dzielinska Z., Jodkowski J., Ruzyllo W., Szmaus P., Dabrowski M. (1996), "Intracoronary Doppler versus new method of the assessment of coronary flow using computer analysis of conventional coronary angiograms", *in Conputers in Cardiology, 1996, p.405-408,*

Murray A., Arzbaecher R. (Eds), IEEE, NY.

Hayano J., Kimura K., Hosaka T., Shibata N., Fukunishi I., Yamasaki K., Mono H., Maeda S. (1997) "Coronary disease-prone behavior among Japanese men: job-centered lifestyle and social dominance. Type A Behavior Pattern Conference" *Am-Heart-J.1997 Dec;* ***134(6)*** *1029-36.*

Hazard B. Munro "Fa health care Research" *ed J.B. Lippincott company 1993.*

Hinton G. E., "Connectionist Learning Procedures", *Technical Report, Computer Science Department, Carnegie-Mellon University, June*

Hornik K., Stinchcombe M., White H. (1989) Multilayer Feedforward Wetworks are Universal Approximates, *Neural Network*, No 2 p.p. 359 - 66

Hosmer D., S. Lemeshow (1989) Applied Logistic Regression, *John Wiley & Sons 1989*

Jain, R., Mazumbar, J., Moran. W. (1995). A comparative study of artificial intelligent techniques in the detection of coronary artery disease. *Proceedings electronic technology directions to the year 2000. pp. 113-120.*

Jerwood D., Price D. J. and Georgiakodis F..*(1989),* , Predicting the potential for recovery within 24 hours of a head injury. *Med. Inform,,****14(4)*** *287-296*

Kan Takayanagi, MD, Hirokazou Yamagushi, MD, Shigenori, Morooka, MD, Yataka, Takabatake, MD. *(1992),* Higher Gensini Score of coronary arteries in acute inferior myocardial infarction with precordial ST-segment depression. *Japanese Heart Journal ,* ***33****; 25-39.*

Kandel Eric, Schwartz James,(1985) "Principles of Neural Science", *(NY: Elsevier Publishing,)*

Kasaoka S., Okuda F., Satoh A., Miura T., Kohno M., Fujii T., Katayama K., Ogawa H., Matsuzaki M. *(1997),* "Effect of Coronary Risk Factors on Coronary Angiographic Morphology in Patients with Ischemic Heart Disease, *Jpn Circ. J.* ***61*** *390-395*

Kimmel SE., Berlin JA., Strom BL., Laskey WK. (1995) "Development and validation of simplified predictive index for major compilation in contemporary percutaneous transuminal coronary angioplasty practice" *J-Am-Cardiol. 1995 Oct;* ***26(4)*** *931-8.*

Kingdon, J., (1997) Intelligent Systems and Financial Forecasting, *Springer-Verlag London Ltd 1997*

Klecka W.R (1984) Discriminant Analysis. *Sage Publications /Beverly Hills/London 1984.*

Knight, K. (1990). Connectionist ideas and algorithms. *Communications of the ACM, **33(11)**, 59-74, 1985.*

Knoke, J.,D (1982). Discriminant analysis with discrete and continuous variables. *Biometrics 1982; **38**: 191-200.*

Kohonen T.(1984), "Self Organization and Associative Memory", *(Berlin: Springer-Verlag, 1984).*

Κόκλα Α.Μ., Γεωργιακώδης Φ., Συριόπουλος Κ., Μάρκελλος Ρ. (2000), «Σύγκριση Διαχωριστικών τεχνικών και ΤΝΔ σε δείγμα στεφανιαίων ασθενών, *Πρακτικά 13ου Πανελληνίου Συνεδρίου ΕΣΥ, Φλώρινα (υπό έκδοση).*

Kokla AM., Georgiakodis F., Tzavelas G. (2000) "A Mathematical model for the Gensini Index", *accepted for oral presentation in the IBS-EMR Conference, Haifa 8-11/1/2001.*

Kostuk WJ., Ehsani AA, Kalriner JS et al.(1973), Self ventricular performance after myocardial infraction assessed by radioisotope angiocardiography Circulation, *1973; **47**; 242*

Kramer B., Brill M. ,Bruhn A., Kubler W. (1986), "Relationship between the degree of coronary artery disease and of left ventricular function and the duration of the QT- interval in ECG", *The European Heart Journal 1986; **7**: 14-24.*

Krishnaswami, S., Joseph, G., Punnoose, E., Chandy, ST. Coronary angiographic findings in-patients with diabetes: An exercise in cardiovascular Epidemiology. *JAPI 1996; **44(3)**169-171.*

Krzanowski W. J. "Mixtures of Continuous and Categorical variables in diocriminant analysis: A Hypothesis-testing approach Biometrics, *December 1982.*

Kuffler S., 1. Nichols, A. Martin, "From Neuron to Brain", *(Sunderland, MA: Sinauer Assoc.).*

Kyriakides, M., Petropoulakis, P., Androulakis, A., Antolopoulos, A., Apostolopoulos, T., Barbetseas, J., Vyssoulis, G., Toutouzas P. J. (1995), J-Clin. Epidemiol. 1995, Sex Differences in the anatomy of coronary artery disease. ***48(6)**:723-730.*

Κυριακίδης Μ. (1987) "Το οξύ έμφραγμα του μυοκαρδίου" *εκδόσεις Γ. Παρισιάνος, 1987.*

Lachenbruch P. A.(1975) Discriminant analysis. New York- Hafner Press.

Lapuerta, P., Azen PS., LaBree, L. (1995). "Use of Neural Networks in predicting the risk of coronary artery disease". *Computer and biomedical research.* ***28(1)****, 38-52.*

Lawrence J. (1993), "Introduction to Neural Networks", *Califomia Scientific Software Press, Appendix C, 303-308.*

Le Cun Y., "Generalization and Network Design Strategies", *Technical Report, 1989 -4, University of Toronto.*

Lehmann E.L Theory of point estimation p.26 ,J. Wiley and Sons1983.

LiMin Fu (1994) Neural Networks in Computer Intelligence, *McGraw-Hill*

Lippmann R.P. , "An introduction to computing with neural nets", *IEEE ASSP Magazine. April 1987, 4-22.*

Low KG. (1991) "Psychosocial variables, Type A behavior pattern and CHD in women", *DAI-B* ***52/01*** *p.85, Jul 1991.*

Mardia K.N., Kent J.T. Bibby J.T., Multivate analysis, *Academic Press.*

McClelland, J., Rumelhart, D. (1988). *Explorations in Parallel Distributed Processing: A handbook of Models, Programs and Exercises.* MIT Press, Lancaster, M.A.

Mc Cullock και Pitts (1943). Bulletin of Mathematical Biophysics, *115-133.*

Menard S. (1980), Applied logistic regression analysis. *Sage Publication.*

Mendes de Leon CF. (1988) " Behavioral and emotional precursors of acute CHD" *DAI-B* ***49/06*** *p.2151, Dec 1988.*

Menotti A., Kromhout D., Blackburn H., Fidanza F., Buzina R., Nissinen A. (1999), Food intake patterns and 25-year mortality from CHD:cross-cultural correlations in the Seven Countries Study. The Seven Countries Study Research Group. *Eur-J-Epidemiol., 1999 Jul.* ***15(6)*** *507-15.*

Mobley BA., Leasure R., Davidson L. (1995) "Artificial network predictions of lengths of stay on a post-coronary care unit" *Heart-lung. 1995 May-Jun;****24(3)*** *251-6.*

Morise AP., Diamond GA., Detrano R., Bobio M., Gunel E.(1996) "The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model" *Med-Decis-Making 1996 Apr-Jun;****16(2)*** *133-42.*

Morrison (1976). "Multivariate statistical methods" *ed Mc Graw–Hill Kogakusha LTD.*

Myers PJ McGee (1993), "Associations of body fat, alcohol consumption and diet with the risk of CHD in women from the Nhanes I Epidemiologic follow-up survey" *DAI-B **53/11** p.p.5650, May 1993.*

Ohno-Machado, L., Musen, A.,M. (1997). Sequential versus networks for pattern recognition: *An exampe using the domain of coronary heart disease. Computers in Biology and Medicine. **27** 267-81.*

Παπαϊωάννου Τ. και Κ. Φερεντίνος (1999). Ιατρική Στατιστική και Στοιχεία Βιομαθηματικών, Ιωάννινα 1999

Park HA. (1988) " Simulation of a population-besed model of CHD morbidity and mortality" *DAI-B **48/12** p.3482, Jun 1988.*

Parzen E. (1962) "On estimation of a probability density function and mode", *Ann. Math. Statis. **33**, pp.1065-1076.*

Pattillo R.W. et al.(1996) Predictors of prognosis by quantitative assessment of coronary angiography single photon emission computed tomography thallium imaging, and treadmill exercise testing *AM. Heart J. 1996; **131**; 582 - 590*

Pedhazur, J.,E. (1982) "Multiple regression in behavioral studies." *2nd ed. The Dryden Press.*

Poppius E., Tenkanen L., Kalimo R., Heinsalmi P. (1999) The sense of coherence, occupation and the risk of CHD in Helsinki Heart Study *Soc-Sci Med 1999 Jul; **49(1)** 109-20*

Press, J., Wilson, S. (1978) Choosing Between Logistic Regression and discriminant Analysis. *Jasa (1978), **73**, No 364:699-705.*

Profile Neural Applications™, User Guide and Reference Manual, Profile Systems & Software SA., Athens, Greece (1995)

Pyorala M., Miettinen H., Halonen P., Laakso M., Pyorala K. (2000) "Insulin resistance syndrome predicts the risk of CHD and stroke in healthy middle-aged men:the 22-year follow-up results of the Helsinki Policemen Study" (2000) *Feb; **20(2)** 538-44.*

Rees BB. (1996) "Influences of CAD knowledge, anxiety, social support and self-efficacy on adaptive health behaviors of patients treated with percutaneous transluminal coronary angioplasty" *DAI-B **56/07** p.3696, Jan 1996.*

Refenes A. N., Alippi C.*(1991),* "Histological image understanding by error backpropagation", *Microprocessing and Microprogramming, **32(3)** 437-46,*

Refenes A. N., Bentz Y., Burgess N.*(1993),* "Neural networks in investment management", *Journal of Communications and Finance, **8**, 95-101.*

Rosenberg EL. (1995), " Emotion, facial expression and CAD", *DAI-B **55/09** p.4172, Mar 1995.*

Rumelhart D.E, Hinton G.E, Williams R.J, "Learning internal representation by error back-propagation", in PDP. Elaboration in the Microstructure of cognition" Vol. I, Bradford Books, Cambridge, MA.

Rumelhart, D., Hinton, G., Williams, R (1986), "Learning representations by back-propagating errors". *Nature, **323,** 533-536.*

Sampath G., "Stochastic Models for Spike Trains of Single Neurons", Lecture Notes in Biomathematics *(Berlin, NY: Springer-Verlag, 1977).*

Selker HP., Griffith JL., Patil S., Long WJ., D'Agostino RB. (1995) "A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients" *J-Investig-Med. 1995 Oct;**43(5)** 468-76.*

Selzer, A (1982). "On the limitation of therapeutic intervention trials in ischemic heart disease: a clinician' s viewpoint". *Am I cardiol 1982; **49**; 252-255.*

Sharma Subhash. Applied Multivariate Techniques. (1996) *John Wiley and Sons, Inc.*

Shen, Z., Clarke, M., Jones. (1994). "Assessing the risk rank of coronary heart disease by use of neural networks : A comparison with a astatistical method". *Proceedings of the international conference of neural networks and expert systems in medicine and health case.204-210.*

Shen, Z., Clarke, M., Jones, R. W., Alberti, T. (1993). Detecting the risk factors of coronary heart disease by use of neural networks. *Proceedings of the 15th annual international conference of the IEEE engineering in medicine and biology society. **15 (1)** , 277-278.*

Shen, Z., Clarke, M., Jones, R. W., Alberti, T. (1993). A neural network approach to the detection of coronary artery disease. *Computers in cardiology **93CH3384-5**, 221-224.*

Shunji Kasaoka, MD, Fumio Okuda, MD, Akira Satoh, MD, Toshiro Miura, MD, Michihito Kohno, MD, Takashi Fujii MD, Kazuhito Katayama, MD, Hiroshi Ogawa, MD, Masunoni Matsuzaki, MD.(1997), "Effect of *Japanese Circulation Journal 1997, 61: 390-395.*

Sietsma J., Dow R.F.J.(1991), "Creating artificial neural networks that Generalize", *Neural Networks, **4**, 67-69.*

Silberberg j., Fryer J., Wlodarczyk J., Robertson R., Dear K. (1999) "Comparison of family history measures used to identify high risk of CHD", *Genet-Epidemiol. 1999, **16(4)**, 344-55.*

Silvey S.D. Statistical interference , *Library of university mathematics, published by Penguin Education 1970.*

Siriopoulos C., A. Sofikitis, (1990), "Technical Analysis: Comparison of Strategies and the Possibility of Ruin as a Measure of the Associated Risk, Some Future Directions using Artificial Intelligence Techniques", *Revue Mondes en Developpement **189** (1990) 19.*

Siriopoulos C., S. Perantonis and G. Karakoulas, (1994), "Artificial Intelligence Models for Financial Decision Making", *Journal of Information Strategy, **11(1)** 49-54.*

Siriopoulos C., Markelos R.N., Sirlantzis K. (1995), "Applications of Artificial Neural Networks in Emerging Financial Markets", *Proceedings of the Third International Conference on Neural Networks in the Capital Markets, 286-288, 1995.*

Siriopoulos C., G. Doukidis, G. Karakoulos, S. Perantonis, S. Varoufakis (1992), "Applications of Neural Networks and Knowledge Based Systems in Stock Investment Management : A comparison of performances", *Journal of Neural Network World **6** (1992), 785.*

Slater JC. (1985) " Physical activity and CHD: an epidemiological study of factors influencing their association and prospects for intervention" *DAI-B **45/08** p.2522, Feb 1985.*

Συριόπουλος Κ. (1997), Εναλλακτικά υποδείγματα των χρηματιστηριακών τιμών: τεχνητά νευρωνικά δίκτυα, *«Ο μηχανισμός των τιμών στα χρηματιστήρια αξιών», σελ.167-186,* Γ. Καραθανάσης και Μ. Γκλεζάκος (Eds), εκδόσεις Παπαζήση 1997.

Συριόπουλος Κ. (1996), Ανάλυση και Έλεγχοι Μονομεταβλητών Χρηματοοικονομικών Σειρών. Τυπωθήτω- Γ. Δαρδανός

Specht D.F (1991), " A general regression neural network", *IEEE Transactions on Neural Networks, **2(6)**, pp.568-576.*

SPSS (1990), Advanced Statistics User's Guide, Marija J. Norusis.

Suhnoon A. (1991). "Correlation of hypercholesterolemia, borderline cholesterol level and traditional risk factors for CHD in the elderly", *MAI **29/04** p.664, Winter 1991.*

Szaboki F., Khoor S., Nieberl J., Fugedi K., Kail E., Kekes E. " Statistical and Neural Network Model for Non-Invasive Estimation of Coronary Anatomy", *www.kard.akh-wien.ac.at*

Sztandera, L. M., Goodenday, L. S., Cros, K. J. (1996). "A neuro-fuzzy algorithm for coronary artery stenosis". *Computers in Biology and medicine **26.** 97-111.*

Takayanagi K. et al.(1991), "Higher Gensini Score of Coronary Arterias in Acute Interior Myocardial Infarction with Precordial St – segment Depression". *Department of Cardiology, Koshigaya Hospital Dokkyo University School of Medianc Japan. (1991)*

Townsend ST. (1995) "The multidimensional aspect of anger/hostility : relationships to CHD". *DAI-B **56/05** p.2889, Nov.1995.*

Τσίμπος Κ., Γεωργιακώδης Φ. (1999) Περιγραφική και Διερευνητική Στατιστική, εκδόσεις Σταμούλη.

Tuomi K. (1994) "Characteristics of work and life predicting CHD. Finnish research project on aging workers" *Soc-Sci- Med. 1994 Jun; **38(11)** 1509-1519.*

Twisk JW., Kemper HC., van-Mechelen W., Post GB. (1997) "Which lifestyle parameters discriminate high- from low-risk participants for CHD risk factors. Longitudinal analysis covering adolescence and young adulthood" *J-Cardiovask-Risk 1997, Oct-Dec; **4(5-6)** 393-400.*

Tzung-Dau Wang, MD, Chau-Chung Wu, MD, Wen-Jone Chen, MD, Chii-Ming Lee, MD, Ming-Fong Chen, MD, Chiau-Suong Liau, MD, Fung-Chang Sung, PhD, and Yuan-Teh Lee, MD.(1998), "Dyslipidemias have a detrimental Effect on left ventricular systolic Function in patients with a first acute myocardial infarction". *Am J. Cardiol, **81** (1998) 531-537.*

Zodpey SP., Kulkarni HR., Vasudeo ND., Chaubey BS. (1994) "A risk scoring system for prediction of CHD based on multivariate analysis: development and validation" *Indian-Heart-J. 1994 Mar-Apr;* ***46(2)*** *77-83.*

Wang XL., Tam C., McCredie RM., Wilcken DE. (1994) "*Deteminants of severity of CAD in Australian men and women" Circulation 1994 May;* ***89(5)*** *1974-81.*

Webb WR., Parker F.B, Nevill J.F, (1973) "Retrograde Pressures and Flows in Coronary Arterial Disease", *The Annals of Thoracic Surgery, 1973,* ***15,*** *256-262.*

Weigend A.S., B.A. Huberman., D.E. Rumelhart (1991), "Generalisation by Weight-Elimination with Application to Forecasting" in *Advances in Neural Information Processing Systems (1991)* **3** *,875.*

Wielgosz AT. (1984) "Self-labelling as a determinant of outcome in symptomatic patients with minimal or no CAD", *DAI-B* ***44/12*** *p.3733, Jun 1984.*

Zhuo Z., Ackerman E., Gatewood L., Kottke T., Wu Shu-Chen, Park Hyeoun-Ae (1991), " Polychotomous Multivariate Models for Coronary Heart Disease Simulation. 1. Tests of a Logistic Model., *Int. J. Biomed Comput,* ***27*** *(1991) 133-148.*

Zhuo Z., Ackerman E., Gatewood L., Kottke T. (1995), " Polychotomous Multivariate Models for Coronary Heart Disease Simulation. 3.Model sensitivities and risk factor interventions, *Int. J. Biomed Comput,* ***28*** *no.3 205-220.*

Zhuo Z., Tsai Yj., Ackerman E., Gatewood L., (1994), "Polychotomous Multivariate Models for Coronary Heart Disease Simulation. 4. The impact of physiological aging. *Int. J. Biomed Comput.* ***37,*** *no.3, 287-296.*