ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Τμήμα Πληροφορικής

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Μελέτη γνωστών αλγορίθμων εκμάθησης γράφων, για την αντιμετώπιση του καρκίνου του τραχήλου της μήτρας»

«Μαρία Γαλάνη»
«p10013»

Επιβλέποντες:
Βέργαδος Δημήτριος
Κουτσούρης Δημήτριος

Πειραιάς, *«Ιούνιος» «2015»*

Επιτελική Σύνοψη

Η παρούσα διπλωματική εργασία ασχολείται με τη σύγκριση γνωστών αλγορίθμων εκμάθησης γράφων που μελετήθηκαν, οι οποίοι ταξινομούν αποτελέσματα εξετάσεων απαραίτητων για τη διάγνωση του καρκίνου του τραχήλου της μήτρας, ώστε να προκύψουν τα καλύτερα δυνατά αποτελέσματα για την αντιμετώπιση του.

Ο συγκεκριμένος καρκίνος είναι πολύ συχνός, κυρίως σε γυναίκες που δεν εξετάζονται προληπτικά, και οι αιτίες εμφάνισης του δεν έχουν διευκρινιστεί με ακρίβεια και εγκυρότητα. Ο συσχετισμός του βέβαια με διάφορους τύπους του ιού HPV, αποτέλεσε σημαντικό βήμα για τη διερεύνηση και άλλων αιτιών που τον προκαλούν. Εξέταση ορόσημο για τον καρκίνο αυτό, αποτελεί ακόμα και σήμερα, το τεστ Παπανικολάου, το οποίο εφευρέθηκε το 1943, από γιατρό ελληνικής καταγωγής, το Γεώργιο Παπανικολάου. Λόγω του συγκεκριμένου τεστ, το ποστοστό θνησιμότητας των γυναικών που πάσχουν από καρκίνο του τραχήλου της μήτρας, μειώθηκε στο 74%.

Παρόλα αυτά, η διάγνωση της νόσου μέσω του τεστ Παπανικολάου δεν είναι πλήρως έγκυρη και επιτυχής, καθώς υπάρχει ένα ποσοστό λανθασμένης πρόβλεψης, και το αποτέλεσμα του τεστ για τα πλακώδη κύτταρα απροσδιορίστου σημασίας (ASCUS - atypical squamous cells of undetermined significance) χρήζει περαιτέρω διερεύνησης γιατί μπορεί να εγκυμονεί κινδύνους. Τα κύτταρα αυτά δεν έχουν «ξεκάθαρο χαρακτήρα» και αποτελούν κάτι σα «γκρι ζώνη». Στα πλαίσια της σωστής διάγνωσης λοιπόν, είναι απαραίτητο να περιληφθούν νέες τεχνικές ανίχνευσης των τραχηλικών αλλοιώσεων, όπως είναι το HPV DNA τεστ, το mRna τεστ, το flow τεστ και το P16 τεστ. Σε περίπτωση δηλαδή που το τεστ Παπανικολάου του ασθενή βγει θετικό, ο ασθενής πρέπει να υποβληθεί και σε άλλα τεστ προκειμένου να εξακριβωθεί το τι ακριβώς συμβαίνει. Όταν το τεστ Παπανικολάου βγει θετικό, σε καμία περίπτωση δε σημαίνει ότι ο ασθενής πάσχει 100% από καρκίνο του τραχήλου της μήτρας.

Στην παρούσα εργασία λάβαμε δεδομένα από 700 γυναίκες που είχαν πραγματοποιήσει και τις 5 παραπάνω εξετάσεις. Τα Μπεϋζιανά δίκτυα λειτουργούν σε αυτή την περίπτωση ως τμήμα των συστημάτων υποστήριξης κλινικής απόφασης, και προτιμώνται στην περίπτωση της διάγνωσης καθώς διαχειρίζονται σε καλύτερο βαθμό το στοιχείο της αβεβαιότητας. Πιο συγκεκριμένα, στην εργασία επικεντρωθήκαμε στην προσπάθεια διευκρίνησης των περιπτώσεων που εμφάνιζαν ASCUS ως αποτέλεσμα του τεστ Παπανικολάου προκειμένου να αποκτήσει ο ιατρός καλύτερη εικόνα για την υγεία της εξεταζόμενης. Το διαγνωστικό συμπέρασμα από τη σύγκριση των αποτελεσμάτων των αλγορίθμων θα μπορούσε να λειτουργήσει συμβουλευτικά για τον κλινικό ιατρό.

Λέξεις – κλειδιά: Καρκίνος τραχήλου μήτρας, HPV, Pap test, HPV DNA test, mRNA test, flow test, P16 test, Δομημένο πιθανοτικό μοντέλο, Μπεϋζιανά δίκτυα, Συστήματα υποστήριξης κλινικής απόφασης, Εξόρυξη Γνώσης, Αλγόριθμος K2, Γενετικοί αλγόριθμοι, Αλγόριθμος Hill climbing, αλγόριθμος Simulated annealing

Abstract

The present diploma thesis deals with the comparison of well known algorithms which classify test results for the diagnosis of cervical cancer, to provide the best possible recommendations for the treatment of this type of cancer.

This particular cancer is very common, especially in women who are not examined proactively, however the causes for its appearance have not yet determined with accuracy and validity. An important step towards the investigation of what causes this particular type of cancer was its correlation with different types of HPV virus. Still the Pap test, which was invented in 1943 from Greek doctor Georgios Papanikolaou, is a big milestone in the proactive examination and treatment. Because of this test, the mortality rate from cervical cancer decreased to 74%, which is indicative of the importance of the invention.

However, the diagnosis of disease through the Pap test is not 100% valid and successful. The fault prediction rate especially for the squamous cells of undetermined significance (ASCUS - atypical squamous cells of undetermined significance) needs further investigation. To ensure correct diagnosis, it is necessary to add new techniques for detecting cervical lesions, such as HPV DNA test, mRna test, the flow test and the P16 test. Therefore, in case the Pap test of a patient is positive, the patient should be submitted to additional tests in order to investigate exactly what is happening.

In our analysis we used data from 380 women who had made all 5 tests above. The diagnostic conclusion from comparing the results of the algorithms could be used as an indicator for the clinician. In this case, the Bayesian networks function as part of clinical decision support systems, and are preferred in the case of diagnosis as well as managing a better degree the element of uncertainty. Specifically, the work focused on the effort of clarification of cases had resulted ASCUS Pap smear to obtain the doctor better picture of the health of the test.

Keywords: cervical cancer, HPV, Pap test, HPV DNA test, mRNA test, flow test, P16 test, Built probabilistic model, Bayesian networks, clinical decision support systems, Mining, Algorithm K2, Genetic algorithms, algorithm Hill climbing , Simulated annealing algorithm

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον κ.Δημήτρη Κουτσούρη και τον κ.Δημήτρη Βέργαδο, οι οποίοι μου έδωσαν τη δυνατότητα να πραγματοποιήσω αυτή τη διατμηματική διπλωματική εργασία.

Επίσης, θα ήθελα να ευχαριστήσω το Δρα. Χάρη Τσίρμπα για την υπομονή του και το χρόνο που αφιέρωσε σε μένα τόσο απλόχερα όλο αυτό το διάστημα.

«Ιούνιος 2015»
«Μαρία Γαλάνη»

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΙΝΑΚΑΣ ΕΙΚΟΝΩΝ

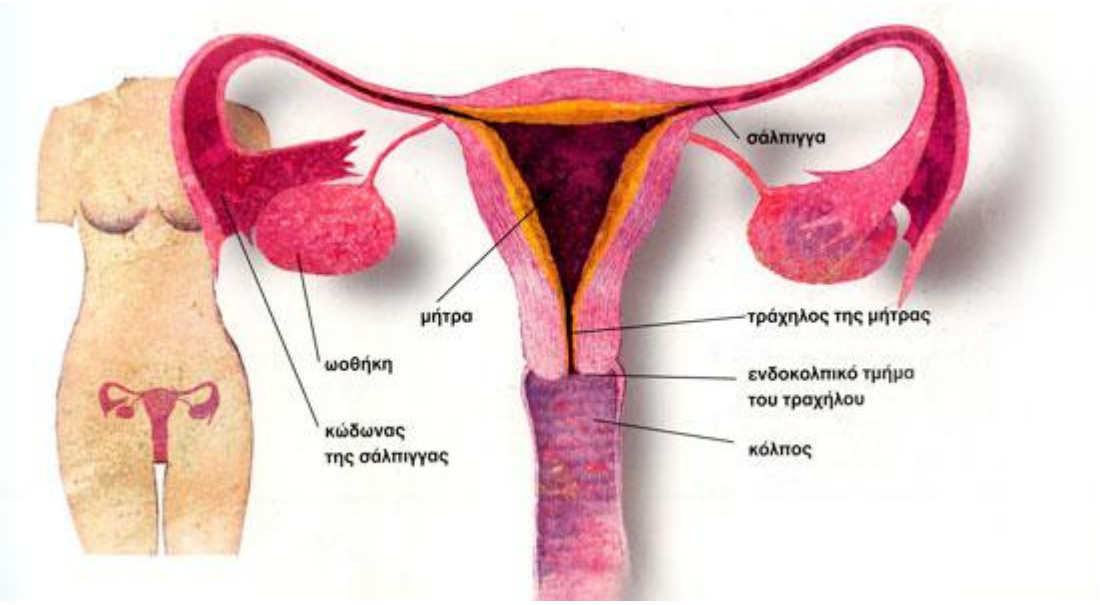
Κεφάλαιο 1^{ο}

1 Ο καρκίνος του τραχήλου της μήτρας

1.1 Η ανατομία του τραχήλου της μήτρας

Τα γεννητικά κύτταρα της γυναίκας, δηλαδή τα ωάρια και οι ορμόνες παράγονται από τις ωοθήκες. Οι δύο ωοθήκες της γυναίκας έχουν ανώμαλη επιφάνεια, με εξογκώματα, ενώ το χρώμα τους είναι ροζ. Τα κύτταρα αυτά, παίζουν σημαντικό ρόλο τόσο στον εμμηνορρυσιακό κύκλο όσο και στη δημιουργία των δευτερογενών χαρακτηριστικών του φύλου. Δηλαδή η λειτουργία τους είναι αναπαγωγική και ενδοκρινολογική ταυτόχρονα. Επίσης και οι σάλπιγγες είναι δύο και ο ρόλος τους είναι να μεταφέρουν το γονιμοποιημένο ωάριο στη μήτρα. Οι σάλπιγγες αποτελούνται από τον κώδωνα, τον ισθμό, τη μητριαία μοίρα και τη λήκυθο όπου συντελείται και η γονιμοποίηση του ωαρίου. Το μήκος τους είναι περίπου 10 με 12 cm.

Ο τράχηλος βρίσκεται κάτω μέρος της κοιλιακής χώρας της γυναίκας. Το ανώτερο μέρος της μήτρας λέγεται σώμα ενώ το κατώτερο μέρος της ονομάζεται τράχηλος της μήτρας και είναι ένας ινομυώδης σωλήνας, που φέρνει σε επικοινωνία το σώμα της μήτρας με τον κόλπο.



Εικόνα 1 Η ανατομία του γυναικείου αναπαραγωγικού συστήματος

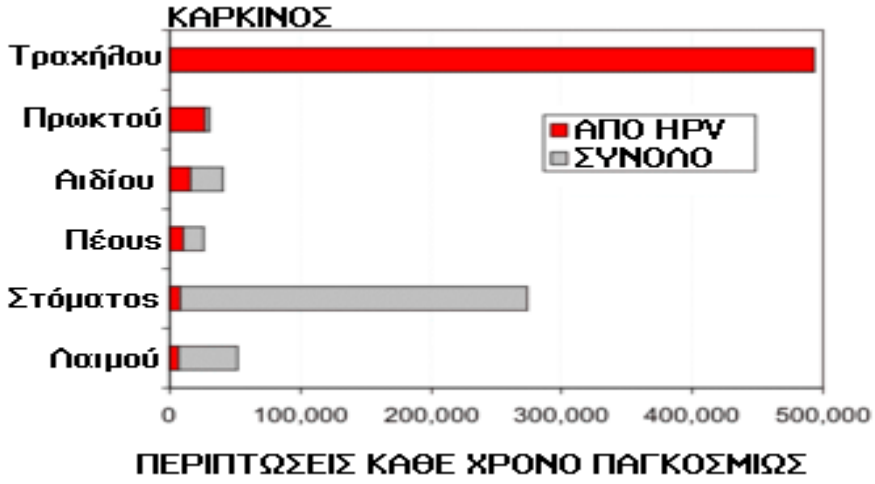
Στη συγκεκριμένη εργασία, επικεντρωνόμαστε στον τράχηλο της μήτρας, λόγω του ότι είναι το σημείο εκείνο που εμφανίζεται και αναπτύσσεται η νόσος του καρκίνου, μία

νόσος που μαστίζει τις γυναίκες σε όλο τον κόσμο. Στο κάτω μέρος της μήτρας βρίσκεται ο τράχηλος, ο οποίος διαθέτει άνοιγμα που βοηθάει την επικοινωνία του κόλπου με την κοιλότητα της μήτρας. Τα σπερματοζωάρια του άντρα μεταφέρονται από τον κόλπο στις σάλπιγγες μέσω του τραχήλου. Ο ατρακτοειδής σχήματος ενδοτραχηλικός σωλήνας διασχίζει την κυτταρική επιφάνεια, η οποία χωρίζεται στο πλακώδες και το αδενικό επιθήλιο. Το πλακώδες επιθήλιο αφορά την επιφάνεια του τραχήλου, ενώ το αδενικό την εσωτερική επιφάνεια του τραχηλικού σωλήνα [1].

1.2 Ο καρκίνος του τραχήλου σε παγκόσμια κλίμακα

Παλαιότερα, ο καρκίνος του τραχήλου της μήτρας αποτελούσε έναν από τους κυριότερους λόγους θνησιμότητας των γυναικών. Τελευταία έχουν αναπτυχθεί διάφορες μέθοδοι ανίχνευσης και θεραπείας του, με στόχο η θνησιμότητα να μειωθεί σημαντικά, με αποτέλεσμα να βρίσκεται στην τέταρτη θέση στην παγκόσμια κλίμακα εκτίμησης θανάτου λόγω καρκίνου.

Τα ποσοστά θνησιμότητας λόγω του καρκίνου του τραχήλου της μήτρας διαφέρουν αρκετά μεταξύ των αναπτυγμένων και των αναπτυσσόμενων χωρών. Οι χώρες της Αφρικής, της κεντρικής και νότιας Ασίας και της Λατινικής Αμερικής κρατούν τα πρωτεία των υψηλότερων ποσοστών εμφάνισης της συγκεκριμένης νόσου. Αντίθετα, τα χαμηλότερα ποσοστά κατέχουν η Βόρεια Αμερική, η Αυστραλία και η Δυτική Ασία. Σε χώρες όπου γίνεται συστηματική χρήση του τεστ Παπανικολάου, όπως είναι οι ΗΠΑ, τα ποσοστά εμφάνισης της νόσου, καθώς και η θνησιμότητα λόγω αυτής, είναι πολύ μικρότερη σε σχέση με τις υπόλοιπες χώρες, πόσο μάλλον σε σχέση με χώρες υποανάπτυκτες που στερούνται ιατρικής περίθαλψης και κατάλληλων υποδομών [1].



Εικόνα 2 Περιπτώσεις καρκίνου λόγω του ιού HPV κάθε χρόνο παγκοσμίως

1.3 Αιτίες ανάπτυξης του καρκίνου του τραχήλου της μήτρας

Παρόλο που οι αιτίες εμφάνισης του καρκίνου του τραχήλου της μήτρας έχουν διερευνηθεί από πολλούς επιστημονικούς φορείς, δεν έχουν διευκρινιστεί με εγκυρότητα και ακρίβεια. Όμως, το γεγονός αυτό δεν πρέπει να μας αποθαρρύνει, καθώς από αξιόπιστες μελέτες έχει βρεθεί ένας συνδυασμός παραγόντων από τον οποίο μπορεί να προκύψει η εμφάνιση της νόσου.

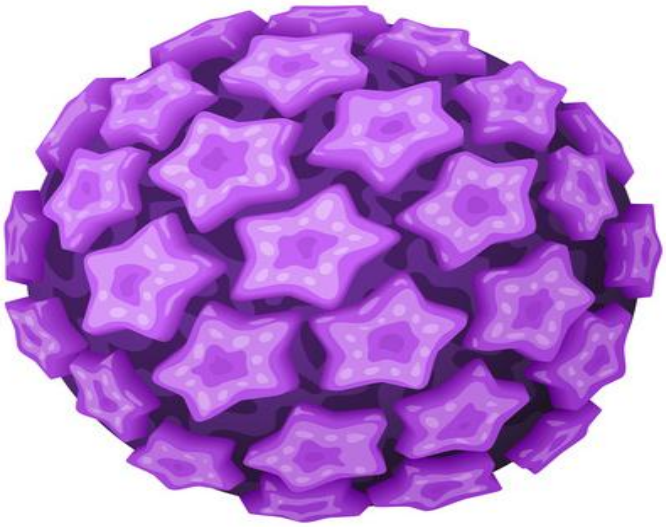
1.4 Ο ρόλος του ιού HPV στην εμφάνισης της καρκινικής νόσου

1.4.1 Ο ιός των ανθρωπίνων θηλωμάτων (Human papilloma virus- HPV)

Ο ιός HPV, ανήκει στην κατηγορία των ιών των ανθρωπίνων θηλωμάτων (papillomaviruses), η οποία έχει κάποια συγκεκριμένα χαρακτηριστικά, όπως είναι το μικρό και χωρίς εξωτερικό περίβλημα μέγεθος, η κοινή γενετική τους δομή, το κυκλικό διπλής έλικας DNA, και η ιδιότητα τους να προσβάλλουν τα επιθηλιακά κύτταρα του δέρματος, καθώς και των βλεννογόνων προκαλώντας πολύ γρήγορο πολλαπλασιασμό κυττάρων. Η τελευταία ιδιότητα είναι και η ιδιότητα η οποία συνδέει και τον συγκεκριμένο ιό με τον καρκίνο του τραχήλου της μήτρας (Monograph on Human Papillomavirus) (htt).

Όπως γνωρίζουμε, το καψίδιο του ιού αποτελείται από δύο δομικές πρωτείνες, την L1 και την L2, οι οποίες κωδικοποιούνται στο DNA. Τα γονίδια του ιού αντιπροσωπεύουν το γενετικό υλικό και η λειτουργική τους περιοχή χωρίζεται σε τρία (Wikipedia) μέρη:

- την πρώιμη περιοχή (E, Early) που κωδικοποιεί τις πρωτεΐνες E1 μέχρι E7 οι οποίες είναι απαραίτητες για την αναπαραγωγή του ιού και κατέχει το 50 % του DNA.
- την όψιμη περιοχή (L, Late) που κωδικοποιεί την κύρια πρωτεΐνη καψιδίου L1 και την δευτερεύουσα πρωτεΐνη καψιδίου L2 που είναι απαραίτητες για τη συγκέντρωση του ιού. Η συγκεκριμένη περιοχή κατέχει το 40 % του DNA.
- Ένα ευρύ μη κωδικοποιημένο τμήμα (LCR, Long Control Region ή αλλιώς NCR, Non Control Region ή αλλιώς URR, Upstream Regulatory Region) που σχετίζεται με την αναπαραγωγή του ιού (Wikipedia) [2].



Εικόνα 3 Το καψίδιο του ιού HPV

Δύο τμήματα πολυαδενυλίωσης (polyadenylation) χωρίζουν τις προαναφερθέντες περιοχές μεταξύ τους (early και late PA).

Ο συγκεκριμένος ιός έχει διάφορους «τύπους» (types). Με τον όρο «τύπο» (type), αναφερόμαστε σε μία μετάλλαξη του ιού, η οποία είναι δυνατόν να απομονωθεί. Μέχρι στιγμής έχουν αναγνωρισθεί 189 τύποι του ιού των θηλωμάτων και πάνω από 100 τύποι του ιού. Στην κατηγορία των σεξουαλικά μεταδιδόμενων συναντάμε περίπου τους 40 υπο-τύπους (subtypes), ενώ σύμφωνα με μελέτες οι 120 από τους υπο-τύπους αφορούν τον άνθρωπο.

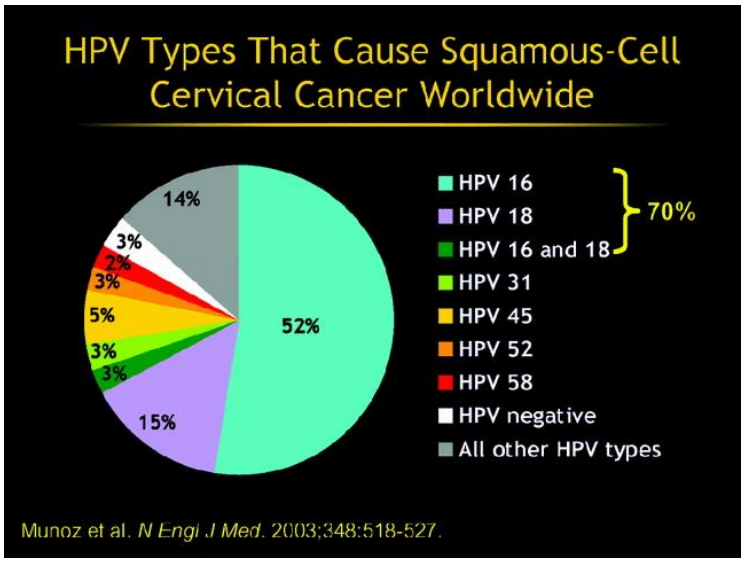
Γενικότερα, για να θεωρηθεί ένας τύπος διαφορετικός από τους ήδη υπάρχοντες τύπους, είναι απαραίτητο να καταγραφεί το σύνολο του γενετικού του κώδικα, καθώς και η ακολουθία του γονιδίου L1 να διαφέρει πάνω από 10% από τον πιο κοντινό γνωστό τύπο. Αν η διαφορά έγκειται μεταξύ 2% και 10%, τότε ορίζεται ένας υπό-τυπος όπως αναφέραμε παραπάνω, ενώ αν η διαφορές είναι μικρότερες του 2% τότε ορίζεται μία παραλλαγή (variant).

Η ομαδοποίηση των τύπων γίνεται με βάση την ομοιότητα στο γενικό τους κώδικα, άρα και στις ιδιότητες που παρουσιάζουν. «Γένη» (genus) ονομάζονται οι πιο υψηλόβαθμες ομάδες τύπων του ιού, ενώ τύποι από διαφοερικά γένη μπορεί να παρουσιάζουν ομοιότητα από 23% έως 43%. Ο διαχωρισμός κάθε γένους γίνεται σε «είδη» (species), με τα διαφορετικά είδη ενός γένους να παρουσιάζουν ομοιότητα από 60% με 70%. Μέχρι τώρα 16 είναι τα γένη που έχουν αναγνωριστεί με βάση τα ελληνικά γράμματα. Το πρώτο γένος είναι εκείνο που σχετίζεται με τον καρκίνο του τραχήλου της μήτρας και ονομάζεται Άλφα ιός των θηλωμάτων (Alpha Papilloma viruses).

Υπάρχουν τύποι του ιού HPV, οι οποίοι είναι υπεύθυνοι για την εμφάνιση κονδυλωμάτων και άλλοι για διάφορες μολύνσεις που ορισμένες φορές μπορεί να οδηγήσουν ακόμα και σε καρκινικές αλλοιώσεις.

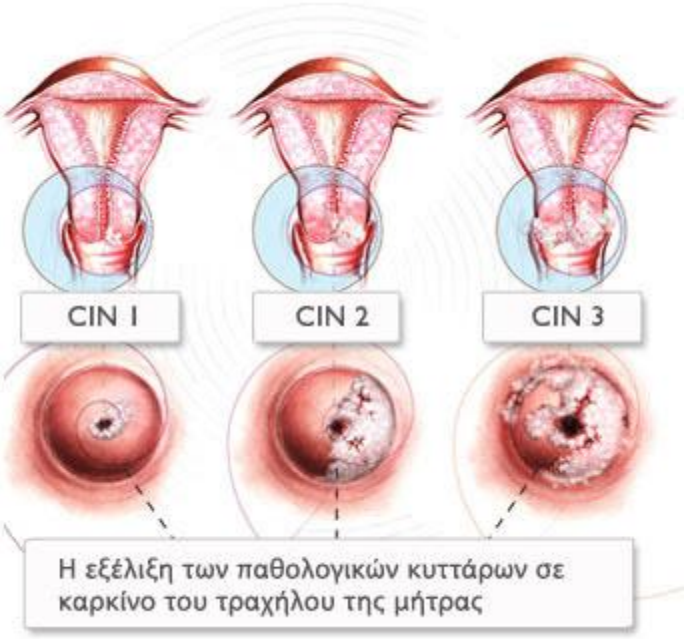
Κάποιοι τύποι είναι γνωστό ότι έχουν την τάση να ενσωματώνονται στο DNA του ανθρώπου, και χρήζουν ιδιαίτερης προσοχής λόγω του υψηλού βαθμού κινδύνου που φέρουν. Αυτοί οι τύποι είναι οι 16, 18, 31, 33, 35, 39, 45, 51, 52, 53, 56, 58, 59, 66, 68, 70, 73, 82, και 85.

Οι παραπάνω τύποι έχουν «ενοχοποιηθεί» για την πρόκληση προκαρκινικών αλλοιώσεων-δυσπλασιών που προκαλούν. Χρησιμοποιούμε τον όρο τραχηλική ενδοεπιθηλιακή νεοπλασία (Cervical Intraepithelial Neoplasia ή CIN) για να περιγράψουμε ένα ευρύ φάσμα ανωμαλιών- αλλοιώσεων που υφίστανται τα κύτταρα του τραχηλικού πλακώδους επιθηλίου. Ακόμη, με τον όρο «δυσπλασία» αναφερόμαστε σε όλες εκείνες τις διαταραχές διαφοροποιήσεως του πλακώδους επιθηλίου που δεν εκπληρώνουν τις προυποθέσεις του ενδοεπιθηλιακού καρκινώματος (in situ). Στο ενδοεπιθηλιακό καρκίνωμα δεν εμφανίζεται κάποια διαφοροποίηση σε ολόκληρο το πάχος του πλακώδους επιθηλίου.



Εικόνα 4 Οι διάφοροι τύποι του ιού HPV που προκαλούν καρκίνο παγκοσμίως

Οι δυσπλαστικές αλλοιώσεις καθορίζουν το πόσο σοβαρή είναι η δυσπλασία. Έτσι, ανάλογα τη βαρύτητα και την έκταση των αλλοιώσεων, η δυσπλασία διακρίνεται σε ελαφρά (CIN-I), μέτρια (CIN-II) και βαριά δυσπλασία και ενδοεπιθηλιακό καρκίνωμα (CIN-III). Η μέτρια δυσπλασία με τη σειρά της μπορεί να εξελιχθεί σε μία προκαρκινική κατάσταση, την αλλοίωση του πλακώδους επιθηλίου (SIL- Squamous Intraepithelial Lesion).



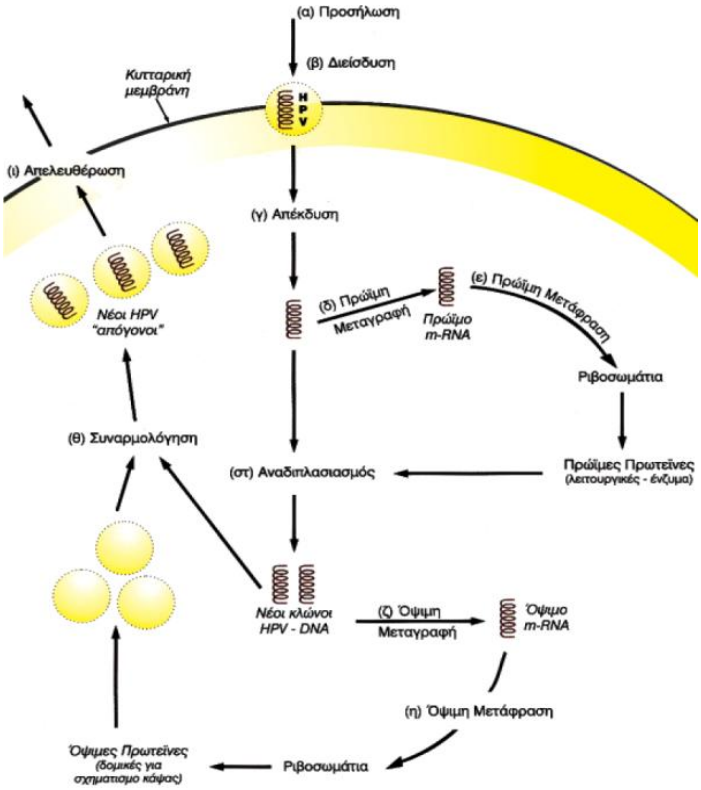
Εικόνα 5 Η εξέλιξη των παθολογικών κυττάρων σε καρκίνο του τραχήλου της μήτρας σε όλες τις φάσεις

Σύμφωνα με την Αμερικανική Κοινωνία Καρκίνου (American Cancer Society)[3], η μόλυνση από τον ιό HPV είναι προϋπόθεση για την ύπαρξη του καρκίνου του τραχήλου της μήτρας. Γνωρίζουμε βέβαια ότι πολλές έρευνες δείχνουν να υπάρχουν αρκετοί παράγοντες που δρουν συμπληρωματικά με το συγκεκριμένο ιό και ενισχύουν την διαδικασία εκδήλωσης της νόσου και ανάπτυξης του καρκίνου του τραχήλου. Αυτοί οι παράγοντες είναι το κάπνισμα, η σεξουαλική δραστηριότητα, η χρήση αντισυλληπτικών φαρμάκων, οι πολλαπλές κυήσεις, η μόλυνση από μικρόβια και μικροοργανισμούς όπως είναι τα χλαμύδια ο έρπις κτλ, η μη άρτια λειτουργία του ανοσοποιητικού συστήματος , το βεβαρυμμένο οικογενειακό ιστορικό καρκίνου του τραχήλου κ.α.

1.4.2 Τρόποι δράσης του ιού HPV στον άνθρωπο

Το πρώτο πράγμα που κάνει ο ιός είναι να προσβάλει τη ζώνη μετάπλασης και πιο συγκεκριμένα τα επιθηλιακά κύτταρα [4]. Στη συνέχεια, μεταφέρεται στον πυρήνα του ξενιστή, και ξεκινάει η διαδικασία αντιγραφής τους με το υγιές γονιδίωμα του κυττάρου-ξενιστή. Αυτό έχει ως αποτέλεσμα να παρατείνεται η διάρκεια ζωής του ιού μέσα στον οργανισμό.

Με την προσβολή αυξάνονται αρχικά οι συγκεντρώσεις των πρωτεϊνών E1 και E2, ώσπου να γίνει η ενσωμάτωση του υιού μέσα στον ξενιστή, άρα και αυξάνονται επίσης και οι συγκεντρώσεις των πρωτεϊνών E6 και E7, με αποτέλεσμα να αδρανοποιηθούν δύο ογκοκατασταλτικές πρωτεΐνες, η p53 και η pRb. Η αδρανοποίηση της πρώτης ογκοκατασταλτικής πρωτεΐνης οδηγεί στον ανεξέλεγκτο πολλαπλασιασμό των κυττάρων.

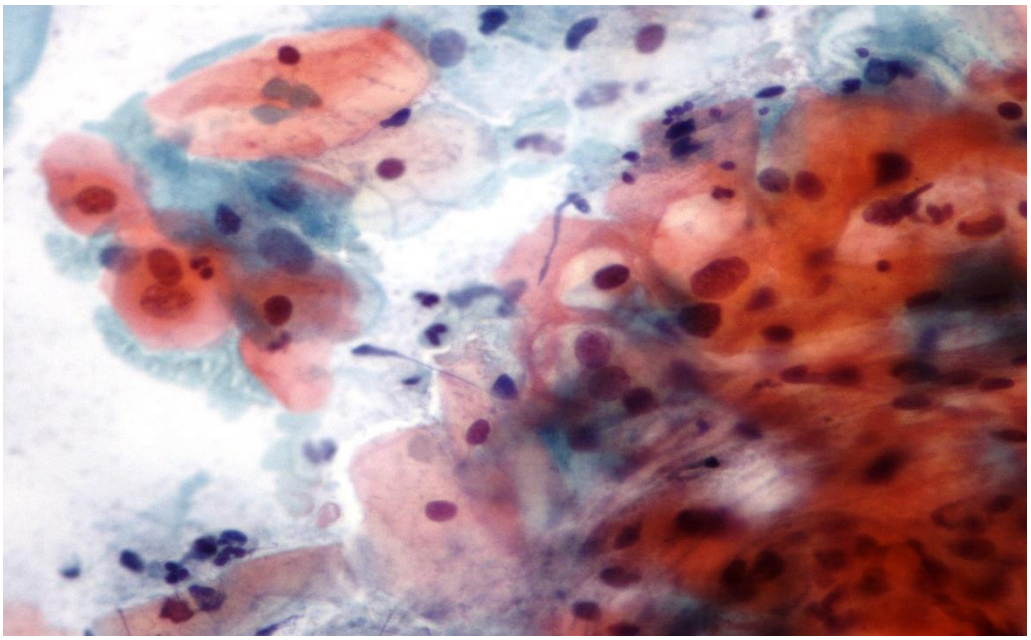


Εικόνα 6 Τρόπος δράσης του ιού HPV

1.4.3 Τρόποι αντίδρασης του ανοσοποιητικού συστήματος στην προσβολή από τον ιό HPV

Η λειτουργία του ανθρώπινου ανοσοποιητικού συστήματος είναι αρκετά αποδοτική, και αυτό φαίνεται από το γεγονός ότι μεγάλος αριθμός τύπων του ιού αντιμετωπίζεται αποτελεσματικά., χωρίς να προκληθεί καρκίνος. Βέβαια, η βασική προϋπόθεση για να συμβεί αυτό είναι η άρτια και άμεση λειτουργία του ανοσοποιητικού συστήματος [4]. Παρόλα αυτά πολλές φορές παρουσιάζονται δυσκολίες στην αποτελεσματική αντιμετώπιση του ιού, οι οποίες αναλύονται παρακάτω:

- Ο ιός δημιουργεί κάποιο είδος φλεγμονής στον οργανισμό, αφού είναι υπεύθυνος για τον ανεξέλεγκτο πολλαπλασιασμό κυττάρων, και όχι την καταστροφή τους.
- Επιπλέον, τα κύτταρα που προσβάλλονται από τον ιό είναι τα επιθηλιακά κύτταρα, τα οποία βρίσκονται μακριά από τα λεμφοκύτταρα, τα οποία είναι υπεύθυνα για την ενίσχυση του ανθρώπινου ανοσοποιητικού συστήματος.
- Τέλος, το αμυντικό σύστημα «παραπλανάται» και δεν αναγνωρίζει τον ιό, λόγω του ότι εμποδίζεται η παραγωγή της ιντερφερόνης από τις πρωτείνες E6 και E7.



Εικόνα 7 Η δράση του ιού HPV στον οργανισμό του ασθενή

1.5 Άλλες αιτίες εμφάνισης του καρκίνου του τραχήλου της μήτρας

Εκτός από την παρουσία του ιού HPV, υπάρχουν και άλλοι παράγοντες που σχετίζονται με την εκδήλωση του καρκίνου του τραχήλου της μήτρας. Αυτοί οι παράγοντες είναι:

- Σεξουαλική δραστηριότητα: Παράγοντας ο οποίος παίζει βασικό ρόλο στη μετάδοση του ιού HPV και στην έναρξη της διαδικασίας μόλυνσης από αυτόν.
- Αντισυλληπτικά σκευάσματα: Αμφιλεγόμενος παράγοντας που χρήζει περαιτέρω διερεύνησης. Πολλές έρευνες (Διεθνή Υπηρεσία Έρευνας Καρκίνου- 2002), έδειξαν ότι η συνεχής και παρατεταμένη χρήση αντισυλληπτικών αυξάνει τον κίνδυνο εμφάνισης καρκίνου, στην περίπτωση που το άτομο έχει προσληφθεί παλιότερα από κάποιο τύπο του ιού HPV, που ανήκει στην ομάδα υψηλού κινδύνου (συνήθως HPV 16 ή 18).
- Κάπνισμα: Η γενικότερη σχέση οποιουδήποτε καρκίνου με το κάπνισμα έχει στοιχειοθετηθεί πολλές φορές. Ο συνδυασμός καπνίσματος και HPV μπορεί να επιτείνει τη διαδικασία δημιουργίας κακοηθών επιθηλιακών κυττάρων. Για παράδειγμα, γυναίκες θετικές σε HPV 16/18 που καπνίζουν, έχουν μεγαλύτερη πιθανότητα να προσληφθούν από καρκίνο του τραχήλου, σε σχέση με μη καπνίστριες.
- Μικρόβια- μικροοργανισμοί: Η παρουσία τους μπορεί να επιταχύνει τη διαδικασία της καρκινογένεσης.
- Μη επαρκής λειτουργία του ανοσοποιητικού συστήματος: Αυτό μπορεί να είναι αποτέλεσμα πρόσληψης του οργανισμού από κάποια άλλη νόσο, η οποία να τον

εξασθένησε, με αποτέλεσμα να μην μπορεί να αντιμετωπίσει ακόμα και τύπου του ιού HPV που ανήκουν σε ομάδες χαμηλού κινδύνου.

1.6 Οι ιστολογικοί τύποι του καρκίνου του τραχήλου της μήτρας

Οι κύριοι ιστολογικοί τύποι του καρκίνου του τραχήλου της μήτρας είναι οι εξής:

- Τα καρκινώματα από πλακώδες επιθήλιο
- Τα αδενοκαρκινώματα

Τα συχνότερα εμφανιζόμενα σε ποσοστό 85-90% είναι τα καρκινώματα από πλακώδες επιθήλιο και διακρίνονται ιστολογικά σε αυτά που αφορούν μεγάλα κερατινοποιημένα κύτταρα, μεγάλα μη κερατινοποιημένα κύτταρα και μικρά κύτταρα.

Τα αδενοκαρκινώματα είναι σπανιότερα με ποσοστό εμφάνισης 10-15%. Ο πιο συχνός τύπος τους είναι το βλεννώδες ενδοτραχηλικό αδενοκαρκίνωμα. Συγκρίνοντας το φυσιολογικό αδενικό κύτταρο με το καρκινικό, διακρίνονται σε υψηλά, μέτρια και χαμηλά. Άλλοι τύποι καρκινώματος είναι τα αδενοπλακώδη, τα αδενοκυστικά, τα μεταστατικά, και έχουν ποσοστό εμφάνισης μόνο 3-5% [4].

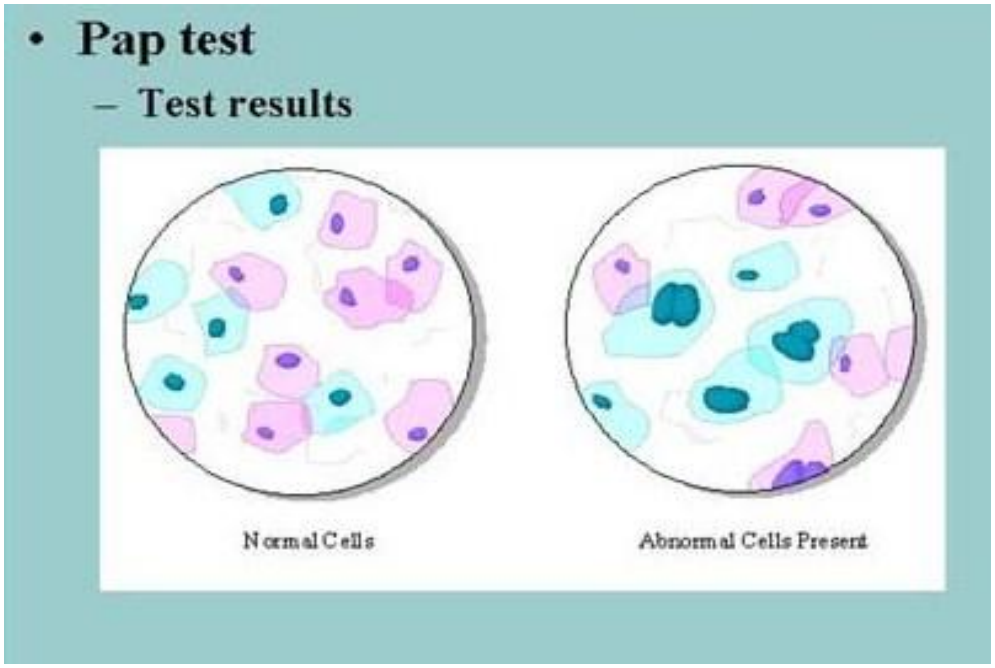
1.7 Προληπτικός πληθυσμιακός έλεγχος και ανίχνευση του καρκίνου του τραχήλου

Η διαδικασία της πρόληψης του καρκίνου του τραχήλου διακρίνεται σε πρωτογενή και δευτερογενή. Η πρωτογενής πρόληψη αφορά τη διαδικασία την οποία πρέπει να ακολουθήσει η γυναίκα προκειμένου να ελαχιστοποιήσει την πιθανότητα εμφάνισης του καρκίνου. Η συγκεκριμένη διαδικασία περιλαμβάνει τον εμβολιασμό της γυναίκας, τη λήψη μέτρων σεξουαλικής προφύλαξης, ακόμα και την αποφυγή του καπνίσματος. Η δευτερογενής πρόληψη αφορά το συστηματικό προληπτικό πληθυσμιακό έλεγχο (screening) στον οποίο οφείλουν να υπόκεινται οι γυναίκες ανά τακτά χρονικά διαστήματα [4].

1.7.1 Τεστ Παπανικολάου

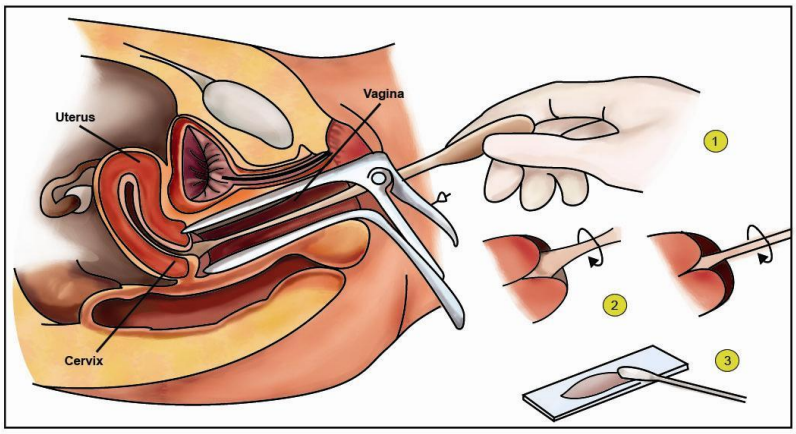
Το τεστ Παπανικολάου είναι το προληπτικό πληθυσμιακό τεστ με τη μεγαλύτερη περίοδο εφαρμογής, αφού μόνο στην Αμερική χρησιμοποιείται για τουλάχιστον 40 έτη. Η μείωση των θανάτων εξαιτίας του καρκίνου του τραχήλου της μήτρας, μετά τη χρήση του συγκεκριμένου τεστ ανέρχεται στο 74%. Το όνομα του τεστ, είναι το όνομα του ιατρού Γεωργίου Παπανικολάου, που ήταν αυτός που το ανακάλυψε. Ο ελληνικής καταγωγής Αμερικανός ανατόμος, διέκρινε ότι καρκινικά κύτταρα μπορούν να παρατηρηθούν στις κολπικές εκκρίσεις γυναικών που πάσχουν από καρκίνο του τραχήλου της μήτρας και το 1943 δημοσίευσε το άρθρο με τίτλο «Διάγνωση του καρκίνου του τραχήλου της μήτρας από το κολπικό επίχρισμα»[5].

Με το συγκεκριμένο τεστ, συλλέγονται όλα τα είδη των κυττάρων από την περιοχή της μήτρας, δηλαδή τα πλακώδη, αδενικά και μεταπλαστικά κύτταρα. Εφόσον πρόκειται για τα επιθηλιακά κύτταρα, το δείγμα συλλέγεται με χρήση σπάτουλας. Τα κύτταρα του ενδοτραχηλικού σωλήνα συλλέγονται με τη χρήση ειδικού βουρτσακίου. Στη συνέχεια τα κύτταρα τοποθετούνται σε γυάλινα πλακίδια και με τη χρήση σπρέι ή διαλύματος αλκοόλης στερεώνονται πάνω σε αυτά, για να εξεταστούν στο μικροσκόπιο.



Εικόνα 8 Αποτελέσματα του τεστ Παπανικολάου φυσιολογικών και μη φυσιολογικών κυττάρων

Το τεστ Παπανικολάου έχει εξελιχθεί πολύ τα τελευταία χρόνια και έτσι γίνεται ταυτόχρονη συλλογή αδενικών και πλακωδών κυττάρων, με μία ειδική συσκευή. Μετά τη συλλογή των κυττάρων, αυτά μαζί με το τελικό τμήμα της συσκευής, τοποθετούνται σε μία ειδική φιάλη και με τη βοήθεια ενός άλλου ειδικού μηχανήματος τα κύτταρα στρώνονται σε μονή στοιβάδα. Έτσι, αποφεύγεται η αλληλοκάλυψη των κυττάρων. Η συγκεκριμένη τεχνική ονομάζεται Κυτταρολογία υγρής φάσης (LBC- Liquid Based Cytology).



Εικόνα 9 Διαδικασία με την οποία πραγματοποιείται το Τεστ Παπανικολάου

Τα αποτελέσματα του τεστ Παπανικολάου, βασίζονται στο σύστημα Bethesda και είναι τα εξής:

1. Φυσιολογική κατηγορία (WNL): Παθολογικά ευρήματα εντός φυσιολογικών και αποδεκτών ορίων
2. Καλοήθεις κυτταρικές αλλοιώσεις, φλεγμονή και αντιδραστικές αλλοιώσεις.
3. Κατηγορία «Άτυπα πλακώδη κύτταρα απροσδιορίστου σημασίας» (ASCUS): Τα άτυπα πλακώδη κύτταρα απροσδιόριστου σημασίας, αποτελούν ευρήματα με κάποια ατυπία, χωρίς όμως να μπορεί να βρεθεί με ακρίβεια η σοβαρότητα και η βαρύτητά της. Η ατυπία αυτή μπορεί να προκαλέσει κάποια αλλοίωση η οποία μπορεί να διαπιστωθεί μέσω κολποσκόπησης.
4. Κατηγορία «Χαμηλού επιπέδου ενδοεπιθηλιακές αλλοιώσεις» (LSIL): Η συγκεκριμένη κατηγορία περιλαμβάνει την εύρεση κυττάρων στα οποία παρατηρούνται χαμηλού βαθμού αλλοιώσεις ή αλλοιώσεις που προκαλούνται από θηλωματοϊούς ή από την ελαφριά δυσπλασία (CIN-I). Τα ευρήματα της συγκεκριμένης κατηγορίας είναι ανησυχητικά από την άποψη του πλήθους των περιστατικών που εμφανίζονται στη συνέχεια ως υψηλού επιπέδου ενδοεπιθηλιακές αλλοιώσεις (HSIL).
5. Κατηγορία «Υψηλού επιπέδου ενδοεπιθηλιακές αλλοιώσεις» (HSIL): Στην κατηγορία αυτή τα κύτταρα που εντοπίζονται έχουν υποστεί υψηλού βαθμού αλλοιώσεις. Στη συγκεκριμένη κατηγορία, ανήκει η μέτρια δυσπλασία (CIN-II), η βαριά δυσπλασία (CIN-III), και το ενδοεπιθηλιακό καρκίνωμα.
6. Καρκίνωμα εκ πλακωδών κυττάρων.
7. Καρκίνωμα εξ αδενικών κυττάρων.
8. Άλλα κακοήθη νεοπλάσματα.

Εικόνα 10 Τα αποτελέσματα του τεστ Παπανικολάου

Η παραπάνω ταξινόμηση χρησιμοποιήθηκε μέχρι το 2001. Το 2001 το σύστημα Bethesda υπέστει κάποια αναθεώρηση, η οποία φαίνεται παρακάτω:

- Οι δύο πρώτες κατηγορίες αποτελούν πλέον μία κατηγορία η οποία είναι η «αρνητική για ενδοεπιθηλιακή αλλοίωση ή κακοήθεια». Έτσι, οι αντιδραστικές μεταβολές χαρακτηρίζονται ως «καθαρές».
- Οι αλλοιώσεις ASCUS που προκαλούν δίλλημα ως προς το πως θ' αντιμετωπιστούν, χωρίζονται σε δύο κατηγορίες: καλοήθεις αλλοιώσεις (ASCUS) και άτυπα πλακώδη κύτταρα που δεν αποκλείουν την ύπαρξη υψηλού βαθμού πλακώδους ενδοεπιθηλιακής βλάβης(ASC-H). Το ASC-H έχει μεγαλύτερη πιθανότητα να είναι προκαρκινική κατάσταση.
- Άλλη κατηγορία.
- Όπως και σε κάθε άλλη ιατρική εξέταση, έτσι και στο τεστ Παπανικολάου η αποτελεσματικότητα του δεν είναι καθολική. Η αναποτελεσματικότητα μπορεί να οφείλεται στο μέγεθος και στη θέση των CIN αλλοιώσεων. Η διάμετρος τους μπορεί να είναι μικρότερη από 0.5cm και όσων αφορά τη θέση τους, μπορεί να βρίσκονται ψηλά στον ενδοτραχηλικό σωλήνα. Αναποτελεσματικότητα μπορεί να προκείψει επίσης λόγω της ανεπιτυχούς λήψης του δείγματος και λόγω της υποκειμενικότητας του εξεταστή.

Για τους παραπάνω λόγους, η γυναίκα πρέπει να εξετάζεται συχνά, και γι' αυτό βέβαια υπάρχουν και συμπληρωματικές τεχνικές, ώστε να ελαχιστοποιείται η πιθανότητα λάθους.

1.8 Νέες τεχνικές ανίχνευσης καρκίνου του τραχήλου της μήτρας

1.8.1 HPV DNA τεστ

Στην περίπτωση που άλλα διαγνωστικά τεστ παρουσιάζουν αδυναμίες, η ανάλυση του DNA «έρχεται» να τις καλύψει. Βασίζεται σε όλες εκείνες τις μοριακές τεχνολογίες που είναι δυνατόν να ανιχνεύσουν το DNA του ιού, σε δείγματα κυττάρων από την περιοχή του τραχήλου [6]. Οι χρησιμοποιούμενες μοριακές τεχνολογίες χωρίζονται σε δύο (2) κατηγορίες:

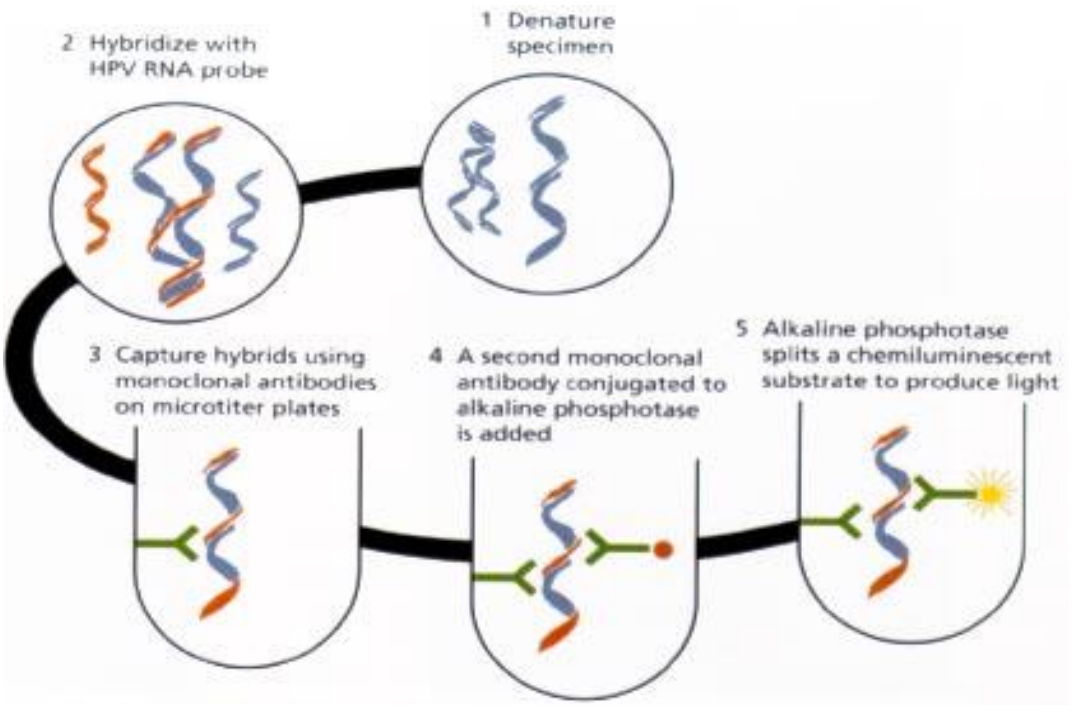
1. Στις τεχνολογίες που δεν υφίστανται καμία ενίσχυση, όπως είναι τα τεστ ανίχνευσης νουκλεικών οξέων.

2. Στις τεχνολογίες οι οποίες εκμεταλλεύονται τη διαδιακασία της ενίσχυσης, όπως η αλυσιδωτή αντίδραση πολυμεράσης (PCR).

Οι τεχνικές πολλαπλασιασμού/ ενίσχυσης διαιρούνται με τη σειρά τους σε τρεις (3) επιμέρους κατηγορίες:

1. Την ενίσχυση στόχου.
2. Την ενίσχυση σήματος.
3. Την ιχνηθέτηση.

Γενικότερα, είναι πολύ σημαντικός και ο προσδιορισμός του τύπου του ιού, αφού κάθε τύπος έχει διαφορετικό ογκογενετικό δυναμικό και πρέπει ν' αντιμετωπιστεί με διαφορετικό τρόπο.



Εικόνα 11 HPV RNA τεστ

Παρόλο που η HPV DNA εξέταση παρουσιάζει υψηλή ευαισθησία, έχει χαμηλή ειδικότητα, με χαρακτηριστική την περίπτωση θετικής HPV DNA και αρνητικής κυτταρολογικής εξέτασης.

Πρέπει να τονιστεί ότι το εν λόγω τεστ δεν χρησιμοποιείται αποκλειστικά και μόνο για την ανίχνευση του καρκίνου του τραχήλου της μήτρας. Συχνά το συναντάμε ως βοήθημα στην κυτταρολογία αλλά και ως μέσο παρακολούθησης μετά το πέρας της θεραπείας της ασθενούς. Στόχος αυτής της περαιτέρω παρακολούθησης είναι η προστασία της ασθενούς από μια πιθανή υποτροπή στη νόσο, καθώς σύμφωνα με τις στατιστικές μελέτες υπάρχει

ένα ποσοστό από 5% έως 19% των χειρουργημένων γυναικών στις οποίες είχε εντοπιστεί αλλοίωση τύπου CIN-II ή CIN-III που υποτροπίασαν μετά το χειρουργείο.

1.8.2 mRNA τεστ

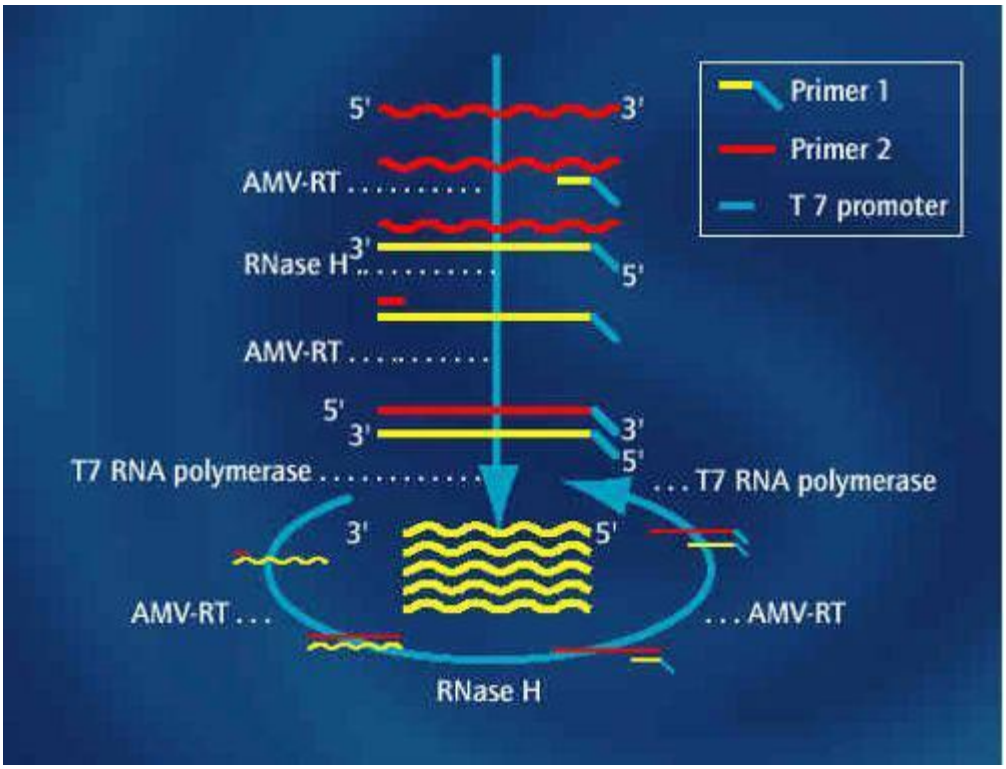
Ο καρκίνος του τραχήλου της μήτρας έχει συνδεθεί με την έκφραση των ογκογονιδίων E6/E7 και επομένως με την ανίχνευση του m-RNA μεταγράφου και το mRNA τεστ βασίζεται σ' αυτό το γεγονός [7]. Η ανίχνευση αυτής της έκφρασης μπορεί να γίνει με δύο (2) τρόπους, είτε μέσω της τεχνολογίας NASBA είτε μέσω τεχνικών αλυσιδωτών αντιδράσεων πολυμεράσης.

1.8.2.1 Τεχνολογία NASBA

Η τεχνολογία Nasba, που είναι η ενίσχυση μιας ειδικής αλληλουχίας νουκλεϊνικών οξέων χρησιμοποιεί εκκινητές ανάλογα με το τι τύπος του ιού είναι προς ανίχνευση, σε μία ενζυματική αντίδραση που πραγματοποιείται στους 41 °C. Στη συνέχεια χρησιμοποιείται μία τεχνική που χρησιμοποιείται στην έρευνα της μοριακής βιολογίας για να μελετήσει την γονιδιακή έκφραση με την ανίχνευση του RNA (ή απομονωμένο mRNA) σε ένα δείγμα, με τη βοήθεια της οποίας οπτικοποιείται το προϊόν της αντίδρασης. Επίσης, χρησιμοποιείται ένας ειδικός ιχνηθέτης (ολιγονουκλεοτίδιο) ή ένα μη ραδιενεργό ένζυμο (ELGA).

Αναλυτικότερα η διαδικασία έχει ως εξής: Ξεκινάει από τον εκκινητή και ακολουθεί η αντίστροφη μεταγραφή του τμήματος της RNA πολυμεράσης, που εμπεριέχεται στον εκκινητή μέσω της AMV και ακολουθεί η υδρόλυση της ακολουθίας στόχου RNA από την RNase. Τότε, ενεργοποιείται ένας δεύτερος εκκινητής, και προστίθεται στην ακολουθία του RNA. Η T7 RNA χρησιμοποιείται ως υποκινητής και με τη βοήθεια της μεταγράφεται αντίστροφα ο αρχικός στόχος. Μετά την ενίσχυση, μέσω των πολλαπλών αντιγραφών, προκύπτει ένα σήμα παραγόμενου φθορισμού, η ανίχνευση του οποίου συνεπάγεται την ύπαρξη του ιού.

Οι τύποι του ιού που εντοπίζονται μέσω αυτής της διαδικασίας είναι οι 16, 18, 31, 33, 45, δηλαδή οι τύποι με την υψηλότερη επικινδυνότητα.



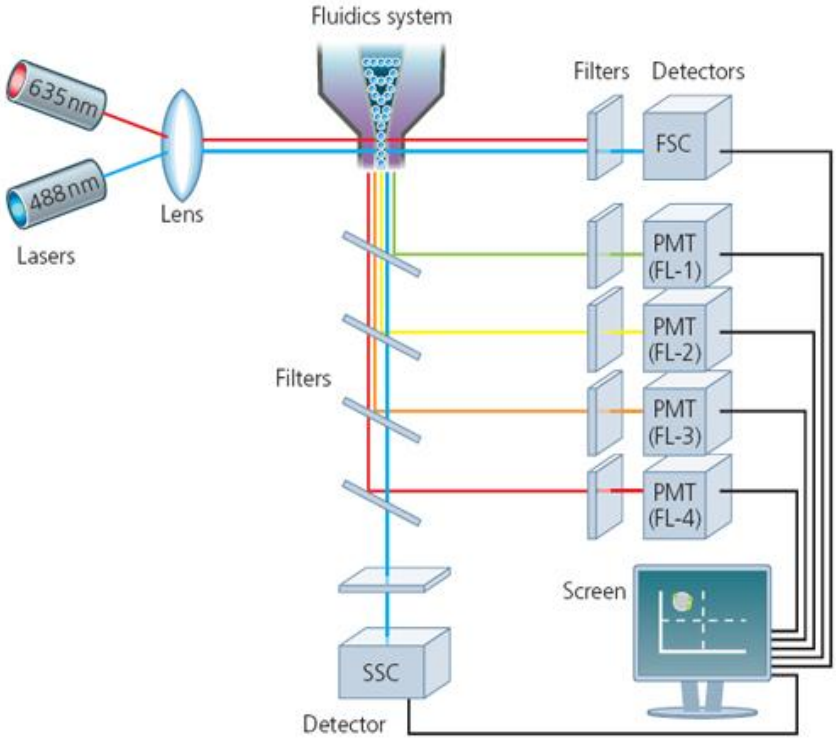
Εικόνα 12 Τεχνολογία NASBA

1.8.2.2 Τεχνική Αλυσιδωτής αντίδρασης πολυμεράσης (PCR)

Η τεχνική αυτή είναι μία τεχνική η οποία χρησιμοποιεί φθορίζοντα μόρια έτσι ώστε να είναι δυνατή η οπτική ανίχνευση του προιόντος της αντίδρασης. Η αρχική ποσότητα του νουκλεικού οξέος που περιέχει το προς εξέταση δείγμα συσχετίζεται άμεσα με την αρχική ποσοτική αύξηση της πολυμεράσης του προιόντος. Ακόμη είναι γνωστό ότι το πλήθος των αλληλουχιών που παράγονται σε κάθε θερμοδυναμικό κύκλο είναι ανάλογο με το φθορισμό που παράγεται σε κάθε κύκλο.
Οι μέθοδοι με τις οποίες μπορεί να πραγματοποιηθεί η τεχνική αυτή είναι: ιχνηθέτες υβριδισμού/ υδρόλυσης, παράγοντες που προσδένονται σε δίκλωνο DNA, και ιχνηθέτες υβριδισμού.

1.8.2.3 Κυτταρολογία ροής

Η κυτταρολογία ροής (flow cytometry) «εκμεταλλεύεται» τα φυσικά φαινόμενα της σκέδασης, του φθορισμού και της υδροδυναμικής, για να δώσει πληροφορίες για τη φυσική και τη χημική δομή ενός σωματιδίου.



Εικόνα 13 Κυτταρολογία ροής

Όπως φαίνεται παραπάνω, δέσμη φωτός συγκεκριμένου μήκους κύματος «οδηγείται» μέσω ειδικού φακού και μιας υδροδυναμικά συγκλίνουσας ροής υγρού. Σωματίδια με μέγεθος 0.2-150 νανομέτρων αιωρούνται στο εσωτερικό του υγρού, με ικανότητα σκέδασης του φωτός προς μία κατεύθυνση. Η φθορίζουσα ακτινοβολία συγκεντρώνεται σε διαφορετική κατεύθυνση. Το μήκος της ακτινοβολίας που εκπέμπει η διεγερμένη ουσία διαφέρει από το μήκος από το μήκος της πηγής. Η σκεδαζόμενη ακτινοβολία χωρίζεται σε εμπρόσθια και πλάγια. Η εμπρόσθια (FSC) δίνει πληροφορίες για τον όγκο του σωματιδίου, ενώ η πλάγια (SCC) δίνει πληροφορίες για την εσωτερική δομή του σωματιδίου. Η ακτινοβολία και στις δύο κατευθύνσεις συλλέγεται με τη βοήθεια ειδικών φίλτρων, ανιχνευτών και ενισχυτών του φωτεινού σήματος και αφού συλλεχθεί οδηγείται στον ηλεκτρονικό υπολογιστή προκειμένου να πραγματοποιηθεί η απαραίτητη επεξεργασία των δεδομένων. Η παραπάνω τεχνική χρησιμοποιείται για την ανίχνευση του μετεγγραφόμενου mRNA, αφού γίνεται μέτρηση της έκφρασης των ογκογονιδίων E6 και E7 [8].
Η κυτταρολογία ροής όταν συνδυάζεται με το τεστ Παπανικολάου και το HPV DNA τεστ γίνεται ακόμα πιο αποτελεσματική.

1.8.3 P16 τεστ

Ο ιός HPV είναι υπεύθυνος για την αδρανοποίηση της πρωτεΐνης ρετινοβλαστώματος pRb και των σχετικών γονιδίων ογκοκαταστολής, καθώς στοχεύει σε ένα αριθμό νουκλεικών

οξέων και πρωτεινών. Η πρωτεΐνη E7 είναι υπεύθυνη για αυτό, με αποτέλεσμα την παρουσία νεοπλασιών και παράλληλης εξέλιξης της υπερέκρασης της πρωτεΐνης $p^{16INK4a}$.
Το συγκεκριμένο τεστ λαμβάνει θετική τιμή όταν έχει ανιχνευτεί η υπερέκραση της συγκεκριμένης πρωτεΐνης, ενώ σε αντίθετη περίπτωση λαμβάνει αρνητική τιμή [9].

Πιο συγκεκριμένα, σχετικά με το $p^{16INK4a}$, είναι γνωστό ότι πρόκειται για ένα ογκοκατασταλτικό γονίδιο. Στην περίπτωση που παρατηρηθεί αυξημένη συγκέντρωση του γονιδίου αυτού προκύπτει διαταραχή του κυτταρικού κύκλου (φάση G1 – S) μέσω της E7 (ογκοπρωτεΐνη). Συγκεκριμένα, το p16 (εν συντομία) δεν επιτρέπει τη σύνδεση των κυκλίνων (cyclins) με τις πρωτεΐνες cyclin – dependent kinases (CDKs) με αποτέλεσμα να μην εξελίσσεται ο κυτταρικός κύκλος και να εξαπλώνεται ο όγκος. Λόγω της ιδιότητας του γονιδίου αυτού , η ανίχνευση του έχει συνδεθεί με τους τύπου του ιού HPV που ανήκουν στην κατηγορία του υψηλού κινδύνου.

Είναι επίσης γνωστό ότι η χρησιμοποίηση του προαναφερθέντος βιοδείκτη οδηγεί στη μείωση των ψευδώς αρνητικών και θετικών αποτελεσμάτων και κατά συνέπεια η ασθενής ακολουθεί την κατάλληλη θεραπεία ανάλογα με την διάγνωσή της.

1.9 Στατιστικά μέτρα για την απόδοση των διαφόρων τεχνικών ανίχνευσης

Υπάρχουν κάποιες βασικές έννοιες σχετικά με την αξιολόγηση των αποδόσεων των διαφόρων τεχνικών ανίχνευσης του καρκίνου του τραχήλου της μήτρας [1]. Αυτές οι έννοιες είναι οι εξής:

1. Ευαισθησία και ειδικότητα

Ευαισθησία

Η ευαισθησία αναφέρεται στην ικανότητα του τεστ να αναγνωρίζει σωστά τους ασθενείς. Μετρά δηλαδή το ποσοστό των αληθώς θετικών αποτελεσμάτων του τεστ στο σύνολο των πραγματικών θετικών αποτελεσμάτων.

Ειδικότητα

Αναφέρεται στην ικανότητα του τεστ να αναγνωρίζει σωστά εκείνους που δεν είναι ασθενείς. Μετρά δηλαδή το ποσοστό των αληθώς αρνητικών αποτελεσμάτων του τεστ στο σύνολο των πραγματικών αρνητικών αποτελεσμάτων.

Στη βέλτιστη περίπτωση έχουμε 100% ευαισθησία και 100% ειδικότητα. Αυτό σημαίνει ότι έχουν προβλεφθεί όλοι οι ασθενείς ως ασθενείς και όλοι οι υγιείς ως υγιείς [42].
Τα παραπάνω μέτρα είναι υπεύθυνα για την αξιολόγηση ενός νέου διαγνωστικού τεστ. Έστω λοιπόν ότι θέλουμε να γίνει μία αξιολόγηση ενός τεστ σχετικό με τον προληπτικό ελέγχο σε μία ομάδα ατόμων που περιλαμβάνει και ασθενείς και υγιείς ανθρώπους. Κάθε

άνθρωπος που συμμετέχει στο συγκεκριμένο τεστ, μπορεί να είναι ασθενής, μπορεί και όχι. Το αποτέλεσμα του τεστ μπορεί να είναι είτε αρνητικό είτε θετικό. Αρνητικό είναι στην περίπτωση που το άτομο δεν πάσχει από την ασθένεια και θετικό είναι όταν το άτομο πάσχει. Τα αποτελέσματα του τεστ μπορεί να συμπίπτουν με την αληθινή κατάσταση του ατόμου αλλά μπορεί και όχι [43]. Τότε μπορούν να υπάρξουν τα παρακάτω αποτελέσματα:

- Αληθώς θετικό (true positive- TP): Ο ασθενής αναγνωρίστηκε σωστά ως ασθενής.
- Ψευδώς θετικό (false positive- FP): Το υγιές άτομο αναγνωρίστηκε λανθασμένα ως ασθενής.
- Αληθώς αρνητικό (true negative- TN):Το υγιές άτομο αναγνωρίστηκε σωστά ως υγιές
- Ψευδώς αρνητικό (false negative- FN): Ο ασθενής αναγνωρίστηκε λανθασμένα ως υγιής.

Έστω a τα αληθώς θετικά αποτελέσματα, b τα ψευδώς θετικά, c τα ψευδώς αρνητικά, και d τα αληθώς αρνητικά αποτελέσματα.

Συνεπώς, το άθροισμα a+c είναι το άθροισμα όλων των ατόμων που πάσχουν από την ασθένεια. Αντίθετα, το άθροισμα b+d είναι το άθροισμα όλων όσων δεν πάσχουν από την ασθένεια.
Οι μαθηματικοί τύποι που συσχετίζουν τα δύο αυτά μέτρα απόδοσης μεταξύ τους είναι οι παρακάτω:

- $\text{Ευαισθησία}=\frac{a}{a+c}=\frac{\text{αληθώς θετικά αποτελέσματα}}{\text{άθροισμα όλων όσων έχουν την ασθένεια}}$
- $\text{Ειδικότητα}=\frac{d}{b+d}=\frac{\text{αληθώς αρνητικά αποτελέσματα}}{\text{άθροισμα όλων όσων δεν έχουν την ασθένεια}}$

Από τα παραπάνω προκύπτει ότι ένα τεστ με υψηλή ειδικότητα και θετικό αποτέλεσμα χρησιμοποιείται για να επιβεβαιώσει την ασθένεια. Αντίστοιχα, ένα τεστ με υψηλή ευαισθησία και αρνητικό αποτέλεσμα, χρησιμοποιείται για να αποκλείσει την ασθένεια.
Στην συγκεκριμένη περίπτωση του καρκίνου του τραχήλου της μήτρας, τα αποτελέσματα ερμηνεύονται με τους εξής τρόπους:

- το ψευδώς αρνητικό αποτέλεσμα σημαίνει ότι κατά την κυτταρολογική εξέταση δεν ανιχνεύθηκαν τα άτυπα κύτταρα τα οποία ήταν παρόντα στο επίχρισμα, και
- το ψευδώς θετικό αποτέλεσμα σημαίνει ότι κατά την κυτταρολογική εξέταση ανιχνεύθηκαν κατά λάθος άτυπα κύτταρα στο εξεταζόμενο επίχρισμα χωρίς αυτά να υφίστανται.

Επίσης μπορούν να υπολογιστούν οι πιθανότητες οι οποίες έχουν ως δεδομένη την πραγματική κατάσταση του ατόμου. Αυτές είναι:

- Πιθανότητα ψευδώς θετικών αποτελεσμάτων (False Positive Rate or Fraction)

  FPF = $\alpha = \frac{FP}{(FP+TN)}$ = 1– specificity

- Πιθανότητα ψευδώς αρνητικών αποτελεσμάτων (False Negative Rate or Fraction)

  FNF = β == $\frac{FN}{(TP+FN)}$ = 1 – sensitivity = 1 – power

- Πιθανότητα αληθώς θετικών αποτελεσμάτων (True Positive Rate or Fraction)
  TPF = sensitivity = power

- Πιθανότητα αληθώς αρνητικών αποτελεσμάτων (True Negative Rate or Fraction)
  TNF = specificity

1. Ορθότητα και ακρίβεια

Ευρέως χρησιμοποιούνται και δύο στατιστικά μέτρα απόδοσης, η ορθότητα και η ακρίβεια.

Ορθότητα

Η ορθότητα αναφέρεται στη διαφορά μεταξύ του μέσου όρου μιας σειράς μετρήσεων από μία τιμή που είναι αποδεκτή ως η αληθής (true) ή ορθή (correct) τιμή της μετρούμενης ποσότητας. Η ορθότητα δηλαδή, αποτελεί το βαθμό προσέγγισης της πραγματικότητας, δηλαδή το πόσο κοντά είναι οι μετρούμενες τιμές στις πραγματικές τιμές. Ουσιαστικά, δείχνει την ικανότητα του διαγνωστικού τεστ να προβλέπει σωστά την ύπαρξη της ασθένειας ή όχι στα άτομα.
Με βάση τα παραπάνω, ο μαθηματικός τύπος του συγκεκριμένου μέτρου είναι:

Ορθότητα= $\frac{a+d}{a+b+c+d}$= $\frac{\text{αληθώς θετικά+αληθώς αρνητικά}}{\text{άθροισμα όλων όσων εξετάστηκαν}}$

Ακρίβεια

Η ακρίβεια είναι ένα μέτρο που αναφέρεται στην προσέγγιση της συμφωνίας μεταξύ των επαναλαμβανομένων αποτελεσμάτων της μεθόδου και το βαθμό στον οποίο επαναλαμβανόμενες μετρήσεις (κάτω από τις ίδιες συνθήκες) δίνουν ίδια αποτελέσματα. Η ακρίβεια είναι ουσιαστικά η ποσότητα που μετρά τη διασπορά (dispersion) των αποτελεσμάτων, όταν η αναλυτική μεθοδολογία επαναλαμβάνεται στο ίδιο δείγμα. Η διασπορά των αποτελεσμάτων προκαλείται από διάφορες τυχαίες πηγές και βρίσκεται γύρω από την αναμενόμενη τιμή του αποτελέσματος εάν δεν υπάρχει συστηματικό σφάλμα [12].

Με βάση τα παραπάνω, ο μαθηματικός τύπος της ακρίβειας είναι:

$$\text{Ακρίβεια} = \frac{a}{a+b} = \frac{\text{αληθώς θετικά}}{\text{αληθώς θετικά} + \text{ψευδώς θετικά}}$$

Κεφάλαιο 2ο

2 Μπεϋζιανά δίκτυα (Bayesian Networks)

2.1 Εισαγωγή

Η θεωρία των πιθανοτήτων παρέχει μια βάση για την επίλυση πολλών επιστημονικών προβλημάτων. Η Τεχνητή νοημοσύνη, και πιο συγκεκριμένα η μηχανική μάθηση (machine learning), είναι ένας από τους τομείς που έχει αξιοποιήσει πλήρως τη θεωρία των πιθανοτήτων για να αναπτύξει νέα θεωρήματα και αλγόριθμους. Πολύ δημοφιλή πιθανοτικά γραφικά μοντέλα (PGMs) είναι τα Μπεϋζιανά δίκτυα (Bayesian Networks), τα οποία συνδυάζουν Θεωρίες γράφων και πιθανοτήτων για να αποκτήσουν μια πιο κατανοητή αναπαράσταση της από κοινού κατανομής πιθανότητας (probability distribution). Αυτό το εργαλείο μπορεί να βοηθήσει στην επίλυση προβλημάτων και στη λήψη αποφάσεων, ιδίως σε τομείς όπου υπάρχει έντονο το στοιχείο της αβεβαιότητας.
Το Μπεϋζιανό δίκτυο αποτελείται από κατευθυνόμενους άκυκλους γράφους, οι οποίοι αντιπροσωπεύουν το σύνολο κάποιων τυχαίων μεταβλητών και τις εξαρτήσεις τους. Οι ακμές αντιπροσωπεύουν τις εξαρτήσεις και οι κόμβοι τις μεταβλητές [10].
Στη συνέχεια ακολουθεί ένα παράδειγμα εισαγωγής πίνακα μαζί με την αντίστοιχη λεζάντα.

Το Μπεϋζιανό δίκτυο αποτελείται από κατευθυνόμενους άκυκλους γράφους, οι οποίοι αντιπροσωπεύουν το σύνολο κάποιων τυχαίων μεταβλητών και τις εξαρτήσεις τους. Οι ακμές αντιπροσωπεύουν τις εξαρτήσεις και οι κόμβοι τις μεταβλητές [10].

2.2 Μπεϋζιανή Θεωρία Απόφασης

Έχοντας ως δεδομένο ένα πρόβλημα ταξινόμησης Μ κλάσεων, με ω_1, ω_2, ..., ω_M, και άγνωστο δείγμα με χαρακτηριστικό x, σχηματίζονται Μ υπό συνθήκη πιθανότητες, $P(\omega i|x), i = 1,2, \dots, M$, οι οποίες ονομάζονται εκ των υστέρων πιθανότητες (a posteriori probabilities). Κάθε $P(\omega i|x)$ αναπαριστά αναπαριστά την πιθανότητα ένα άγνωστο δείγμα να ανήκει στην κλάση ωi, αφού δίνεται ότι το διάνυσμα χαρακτηριστικών του δείγματος παίρνει την τιμή x.

Έστω ότι έχουμε ένα πρόβλημα ταξινόμησης με δύο κλάσεις, και τα δείγματα μπορούν να ταξινομηθούν στις κλάσεις $\omega 1$ και $\omega 2$. Οι εκ των προτέρων πιθανότητες να ανήκει ένα δείγμα σε μία κλάση, $P(\omega 1)$ και $P(\omega 2)$ θεωρούμε ότι είναι γνωστές. Ακόμα και να μην είναι γνωστές όμως, είναι εύκολο να υπολογιστούν από τα διαθέσιμα δείγματα εκπαίδευσης, αφού αν N είναι ο αριθμός των δειγμάτων εκπαίδευσης, και $N1$ αυτά που

ταξινομούνται στην κλάση $\omega 1$, και $N2$ αυτά που ταξινομούνται στην κλάση $\omega 2$, τότε $P(\omega 1) \approx \frac{N1}{N}$ και $P(\omega 2) \approx \frac{N2}{N}$.

Επίσης, θεωρούνται γνωστές οι συναρτήσεις πυκνότητας υπό συνθήκης πιθανότητας $p(x|\omega i), \mathrm{i} = 1,2, \ldots, \mathrm{M}$, που περιγράφουν την κατανομή των διανυσμάτων χαρακτηριστικών σε κλάσεις. Η συνάρτηση $p(x|\omega i)$ αναφέρεται και ως συνάρτηση πιθανοφάνειας (likelihood function) των ωi ως προς x.

Ο κανόνας του Bayes χρησιμοποιεί τις παραπάνω πιθανότητες ως εξής:

$$P(\omega i|x) = \frac{p(x|\omega i)P(\omega i)}{p(x)}$$

Όπου $\mathrm{p(x)}$ είναι η συνάρτηση πυκνότητας πιθανότητας η οποία υπολογίζεται ως εξής:

$$p(x) = \sum_{i=1}^{2} p(x|\omega i)P(\omega i)$$

Άρα, ο κανόνας ταξινόμησης του Bayes εκφράζεται ως εξής:

Αν $P(\omega 1|x) > P(\omega 2|x)$, το x ταξινομείται στην κλάση ω1.

Αν $P(\omega 1|x) < P(\omega 2|x)$, το x ταξινομείται στην κλάση ω2.

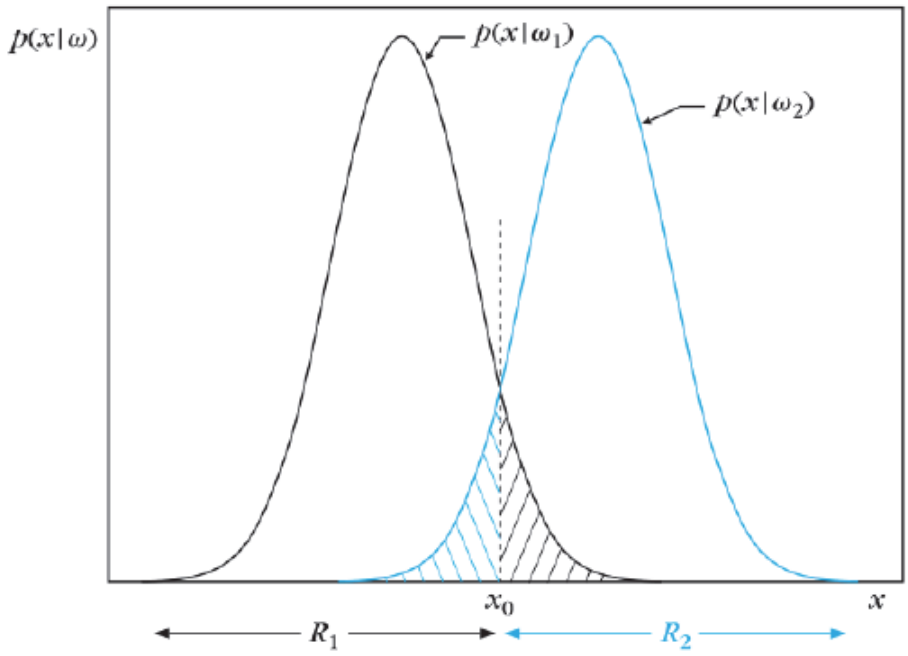
Προφανώς στην περίπτωση που οι πιθανότητες είναι ίσες, το δείγμα μπορεί να ταξινομηθεί σε οποιαδήποτε κλάση. Ουσιαστικά η απόφαση ταξινόμησης βασίζεται στις ανισότητες:

$$p(x|\omega 1)P(\omega 1) \gtrless p(x|\omega 2)P(\omega 2)$$

Το $p(x)$ είναι ίδιο για όλες τις κλάσεις, με αποτέλεσμα να μην επηρεάζει την απόφαση και ουσιαστικά δε λαμβάνεται υπόψη. Αν $P(\omega 1) = P(\omega 2) = 1/2$, οι ανισότητες γίνονται:

$$p(x|\omega 1) \gtrless p(x|\omega 2)$$

Άρα, θα πρέπει να αναζητήσουμε τη μέγιστη τιμή των συναρτήσεων πυκνότητας της υπό συνθήκης πιθανότητας υπολογισμένη στο σημείο x [11].



Εικόνα 14 Συναρτήσεις πυκνότητας υπό συνθήκης πιθανότητας για δύο κλασεις ω1 και ω2

Στην παραπάνω εικόνα παρουσιάζεται ένα δείγμα δύο ισοπίθανων κλάσεων και οι δύο συναρτήσεις $p(x|\omega i) i = 1,2$ για τις διάφορες τιμές του x . Στην εικόνα βλέπουμε και μία διαχωριστική γραμμή x_0 , η οποία ορίζει ένα κατώφλι διαχωρισμού, για το διαχωρισμό του διαχωριστικού χώρου χαρακτηριστικών σε δύο περιοχές R1 και R2. Για τις τιμές του x στην περιοχή R1, ο Bayes ταξινομητής επιλέγει την κλάση ω1, ενώ για τις τιμές του x στην περιοχή R2 επιλέγει την κλάση ω2. Βέβαια, τα σφάλματα στην ταξινόμηση δεν είναι δυνατόν να αποφευχθούν.

Θα υπάρξει σφάλμα ταξινόμησης στην περίπτωση που υπάρχει μια πεπερασμένη πιθανότητα για ένα διάνυσμα x να ανήκει στην κλάση ω1. Το ίδιο συμβαίνει και για σημεία της περιοχής R1 με κλάση ω2. Η πιθανότητα σφάλματος στην περίπτωση που έχουμε δύο ισοπίθανες κλάσεις υπολογίζεται από την παρακάτω εξίσωση:

$$P_e = \frac{1}{2}\int_{-\infty}^{x_0} p(x|\omega 2)dx + \frac{1}{2}\int_{x_0}^{+\infty} p(x|\omega 1)dx$$

Η πιθανότητα αυτή ισούται με τη συνολική γραμμοσκιασμένη περιοχή που φαίνεται στην παραπάνω εικόνα [13].

2.3 Απλοικός Μπευζιανός Ταξινομητής

Έχοντας ένα σύνολο δεδομένων εκπαίδευσης, θα πρέπει να προσδιορίσουμε τις συναρτήσεις πυκνότητας πιθανότητας $p(x|\omega i), i = 1,2, \dots, M$, για να εφαρμοστεί ο κανόνας ταξινόμησης του Bayes. Συμπεραίνουμε ότι το πλήθος των δειγμάτων εκπαίδευσης Ν καλύτερα είναι να είναι μεγάλο, ώστε οι εκτιμήσεις των συναρτήσεων πυκνότητας πιθανότητας να είναι ασφαλείς. Παρ' όλα αυτά, η ανάγκη για επιπλέον δεδομένα εκπαίδευσης αυξάνεται εκθετικά σε σχέση με τη διάσταση l του χώρου των χαρακτηριστικών. Θα χρειαστούν N^l δείγματα εκπαίδευσης για ένα διανυσματικό χώρο l διαστάσεων, αν από N δείγματα εκπαίδευσης μπορούν να προκύψουν επαρκώς ακριβείς εκτιμήσεις μιας συνάρτησης πυκνότητας πιθανότητας για διανυσματικό χώρο μιας διάστασης. Οι μεγάλες τιμές του l δεν επιτρέπουν με μεγάλη ακρίβεια την εκτίμηση της συνάρτησης πυκνότητας πιθανότητας, γιατί είναι πολύ δύσκολο να συγκεντρωθεί ένας πολύ μεγάλος αριθμός δεδομένων εκπαίδευσης [13].

Συχνά χρησιμοποιούμε την προσέγγιση ότι τα χαρακτηριστικά $x_j, j = 1,2, \dots, l$ του προβλήματος είναι στατικώς ανεξάρτητα. Άρα ισχύει:

$$p(x|\omega i) = \prod_{j=1}^{l} p(x_j|\omega i), \qquad i = 1,2, \dots, M$$

Στη συγκεκριμένη περίπτωση, χρειάζονται lN δείγματα εκπαίδευσης και όχι N^l, για τον υπολογισμό l συναρτήσεων πυκνότητας πιθανότητας μίας διάταξης. Ο Απλοικός Μπευζιανός Ταξινομητής (Naïve Bayes Classifier), χρησιμοποιώντας την παραπάνω παραδοχή, αναθέτει ένα άγνωστο δείγμα με διάνυσμα χαρακτηριστικών $x = [x1, x2, \dots, x_l]^T$ στην κλάση:

$$\omega_m = \arg\max \prod_{j=1}^{l} p(x_j|\omega_i), \qquad i = 1, 2, \dots, M$$

2.4 Μπεϋζιανά δίκτυα

Ένας τρόπος αντιμετώπισης της ανάγκης για μεγάλο πλήθος δεδομένων εκπαίδευσης είναι ο Απλοϊκός Μπεϋζιανός Ταξινομητής. Ουσιαστικά εκμεταλλεύεται πιο αποδοτικά τα διαθέσιμα δεδομένα εκπαίδευσης, παρ' όλα αυτά είναι μία ακραία θεώρηση, γιατί υποθέτει ότι όλα τα χαρακτηριστικά είναι εντελώς ανεξάρτητα μεταξύ τους. Άρα, θα πρέπει να βρεθεί μια λύση που θεωρεί την ανεξαρτησία των χαρακτηριστικών όπου αυτή χρειάζεται.

Τα Μπεϋζιανά Δίκτυα (Bayesian Networks) εισάγουν μία μεθοδολογία δημιουργίας μοντέλων, που μπορούν να εκφράσουν υποθέσεις ανεξαρτησίας για συγκεκριμένα

χαρακτηριστικά $x_i, i = 1,2, \dots, l$. Χρησιμοποιείται ο κανόνας της αλυσίδας ο οποίος είναι ο εξής:

$$p(x1, x2, \dots, x_l) = p(x_l|x_{l-1}, \dots, x1)p(x_{l-1}|x_{l-2}, \dots, x_l) \dots p(x_2|x_1)p(x1)$$

Ο κανόνας αυτός δεν εξαρτάται από τη σειρά με την οποία παρουσιάζονται τα χαρακτηριστικά, και εφαρμόζεται πάντα, δηλώντας ότι η από κοινού συνάρτηση πυκνότητας πιθανότητας μπορεί να εκφραστεί ως γινόμενο συναρτήσεων πυκνότητας υπό συνθήκη πιθανότητας [14]. Έτσι, αξιοποιώντας τον παραπάνω κανόνα η υπό συνθήκη εξάρτηση για κάθε χαρακτηριστικό x_i θα περιοριστεί σε ένα υποσύνολο χαρακτηριστικών τα οποία θα εμφανίζονται σε κάθε όρο των γινομένων της αλυσίδας. Η εξίσωση λοιπόν μπορεί να γραφτεί ως εξής:

$$p(x) = p(x1)\prod_{i-2}^{l} p(x_i|A_i)$$
$$A_i \subseteq \{x_{i-1}, x_{i-2}, \dots, x_1\}$$

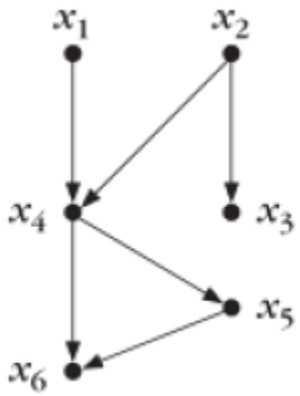
Ας δώσουμε ένα παράδειγμα, για $l = 6$. Έχουμε:

$$p(x_6|x_5, \dots, x_1) = p(x_6|x_5, x_4)$$
$$p(x_5|x_4, \dots, x_1) = p(x_5|x_4)$$
$$p(x_4|x_3, x_2, x_1) = p(x_4|x_2, x_1)$$
$$p(x_3|x_2, x_1) = p(x_3|x_2)$$
$$p(x_2|\, x_1) = p(x_2)$$

Άρα,

$$A_6 = \{x_5, x_4\}, A_5 = \{x_4\}, A_4 = \{x_2, x_1\}, A_3 = \{x_2\}, A_2 \neq \emptyset$$

Οι παραπάνω εξαρτήσεις απεικονίζονται γραφικά στο παρακάτω σχήμα:



Εικόνα 15 Γραφικό Μοντέλο απεικόνισης των εξαρτήσεων υπό συνθήκη μεταξύ των χαρακτηριστικών

Κάθε κόμβος στο γράφημα αντιστοιχεί σε ένα χαρακτηριστικό. Οι γονείς ενός χαρακτηριστικού x_i, είναι τα χαρακτηριστικά εκείνα που συνδέονται άμεσα μαζί του και ανήκουν στο σύνολο A_i. Άρα, το χαρακτηριστικό x_i, αν πάρουμε υπόψιν τους γονείς του, είναι υπό συνθήκη ανεξάρτητο από οποιοδήποτε συνδυασμό χαρακτηριστικών που δεν είναι απόγονοί του. Από την πρόταση $p(x_3|x_2, x_1) = p(x_3|x_2)$ δεν μπορούμε να βγάλουμε το συμπέρασμα ότι τα x_1 και x_3 είναι ανεξάρτητα.

Η εκτίμηση της κοινής συνάρτησης πυκνότητας πιθανότητας προκύπτει από το γινόμενο απλούστερων όρων, με βάση τις παραπάνω παραδοχές. Σε γενικές γραμμές, κάθε όρος περιλαμβάνει μικρότερο αριθμό χαρακτηριστικών σε σχέση με απλούστερους όρους, όπως στις παραπάνω εξισώσεις, στις οποίες δεν περιλαμβάνονται πάνω από τρία χαρακτηριστικά. Η ανάγκη για περισσότερα δεδομένα εκπαίδευσης μειώνεται, καθώς ο υπολογισμός κάθε όρου της συνάρτησης πυκνότητας πιθανότητας γίνεται σε διανυσματικό χώρο χαμηλότερων διαστάσεων. Παρακάτω, ορίζεται η έννοια του Μπεϋζιανού Δικτύου, με τις συναρτήσεις πυκνότητας πιθανότητας να γίνονται απλές πιθανότητες, αφού θεωρούμε ότι τα χαρακτηριστικά παίρνουν διακριτές τιμές [15].

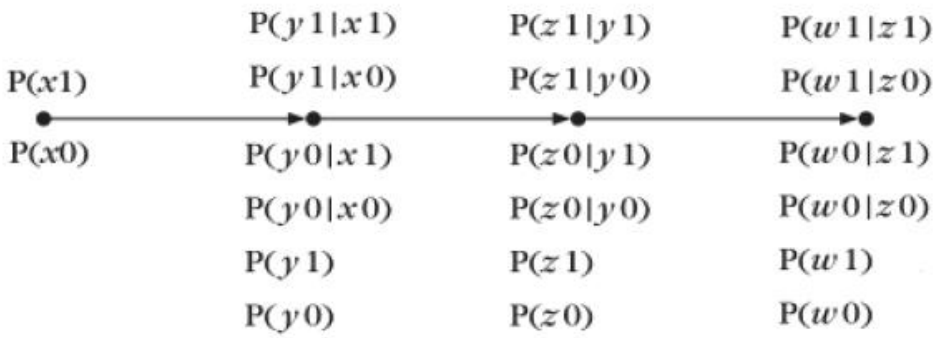
Ως Μπεϋζιανό Δίκτυο (Bayesian Network) ορίζεται ένα κατευθυνόμενο ακυκλικό γράφημα (directed acyclic graph- DAG) όπου ο κάθε κόμβος αναφέρεται σε μία τυχαία μεταβλητή (χαρακτηριστικό). Η τιμή της τυχαίας μεταβλητής, υπόκειται σε διακυμάνσεις λόγω τύχης. Μία τυχαία μεταβλητή μπορεί να πάρει ένα σύνολο δυνατών τιμών, σε κάθε μία από τις οποίες αντιστοιχεί μία πιθανότητα, αν μιλάμε για διακριτές τυχαίες μεταβλητές ή μία πυκνότητα πιθανότητας, αν μιλάμε για συνεχείς τυχαίες μεταβλητές. Γενικότερα, όλες οι πιθανές τιμές μιας τυχαίας μεταβλητής μπορεί να αντιπροσωπεύουν τα πιθανά αποτελέσματα ενός πειράματος που η πραγματοποίηση του είναι σε εξέλιξη ή έχει ήδη πραγματοποιηθεί, αλλά το αποτέλεσμα του είναι αβέβαιο.

Οι τυχαίες μεταβλητές, μπορεί να είναι διακριτές, δηλαδή να έχουν αριθμήσιμο πλήθος δυνατών τιμών ή συνεχείς, δηλαδή να μπορούν να πάρουν οποιαδήποτε τιμή σε ένα διάστημα αριθμών ή σε ένωση διαστημάτων.

Κάθε κόμβος του Μπεϋζιανού δικτύου λοιπόν, σχετίζεται με ένα σύνολο υπο συνθήκη πιθανοτήτων του $P(x_i|A_i)$, όπου x_i η μεταβλητή που αναπαριστά ο κόμβος και A_i το σύνολο των γονέων του κόμβου στο γράφημα [29].

Για να γίνει πλήρης προσδιορισμός ενός Μπεϋζιανού Δικτύου θα πρέπει να προσδιοριστούν:
Οι περιθώριες πιθανότητες (marginal probabilities) των ριζών του δικτύου, δηλαδή των κόμβων που έχουν γονέα.
Οι υπο συνθήκη πιθανότητες για τους κόμβους που δεν είναι ρίζες, για όλους τους δυνατούς συνδυασμούς των τιμών τους, δεδομένου των γονέων τους.

Αν πολλαπλασιάσουμε όλες τις υπο συνθήκη πιθανότητες με τις περιθώριες πιθανότητες των ριζών, υπολογίζουμε την από κοινού πιθανότητα κάθε μεταβλητής. Για να γίνει αυτό, πραγματοποιούμε αρχικά τοπολογική ταξινόμηση (topological sorting) των τυχαίων μεταβλητών, δηλαδή διάταξη των μεταβλητών έτσι ώστε κάθε μεταβλητή να έχει θέση πριν τους απογόνους της. Έχουμε για παράδειγμα ένα απλό Μπεϋζιανό Δίκτυο, που παρουσιάζεται στην παρακάτω εικόνα, το οποίο περιέχει τις δυαδικές μεταβλητές x, y, w, z και η από κοινού πιθανότητα για τη μεταβλητή y μπορεί να υπολογιστεί ως εξής:

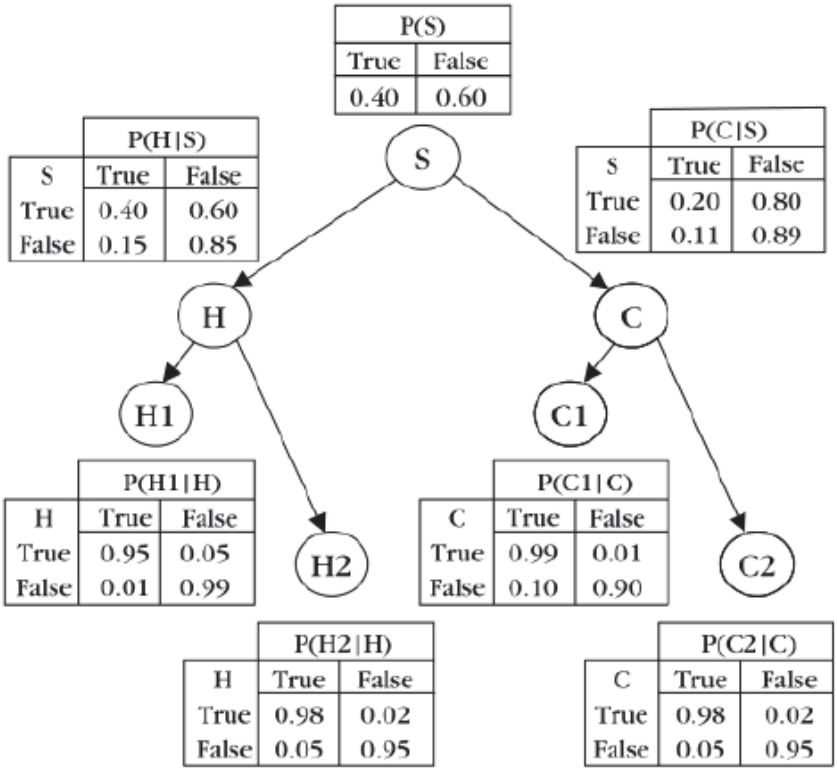


Εικόνα 16 Οι υπό συνθήκη εξαρτήσεις έχουν περιοριστεί σε μια μόνο μεταβλητή

$$P(y1) = P(y1|x1)P + P(y1|x0)P(x0)$$
$$P(y0|x1) = 1 - P(y1|x1)$$

Όπου με y1, x1 συμβολίζονται τα $y = 1$, $x = 1$ και με y0, x0 συμβολίζονται τα $y = 0$, $x = 0$.

Παρακάτω παρουσιάζεται άλλο ένα παράδειγμα Μπεϋζιανού Δικτύου. Είναι ένα παράδειγμα ιατρικής διάγνωσης. Με S συμβολίζεται η μεταβλητή η οποία αναφέρεται στους καπνιστές, με C η μεταβλητή η οποία αναφέρεται στον καρκίνο του πνεύμονα και με H στην καρδιοπάθεια. Οι C1, C2 είναι ιατρικά τεστ για τη διάγνωση καρκίνου, και οι H1, H2 είναι ιατρικά τεστ για τη διάγνωση της καρδιοπάθειας [15].

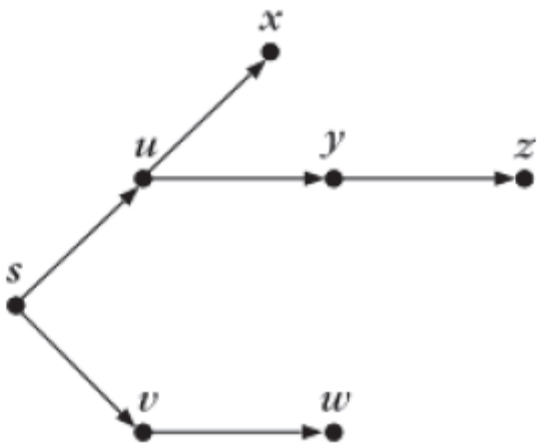


Εικόνα 17 Παράδειγμα Μπεϋζιανού Δικτύου

Οι καπνιστές του πληθυσμού φαίνονται στον πίνακα που βρίσκεται στον κόμβο- ρίζα, ενώ οι αντίστοιχες υπό συνθήκη πιθανότητες είναι οι πίνακες στους κόμβους του δέντρου. Για παράδειγμα, ένας καπνιστής μπορεί να αναπτύξει καρκίνο με πιθανότητα $P(C: True|S: True)$.

Με δεδομένο το γεγονός ότι κάποιοι από τους κόμβους έχουν παρατηρηθεί, άρα μπορούν να εξαχθούν συμπεράσματα σχετικά με τις πιθανότητες (probability inference), το Δίκτυο δίνει τη δυνατότητα για αποδοτικό υπολογισμό της υπό συνθήκης πιθανότητας οποιουδήποτε κόμβου στο γράφημα [52].

Όσων αφορά τα Μπεϋζιανά Δίκτυα που έχουν δομή δέντρου, τα συμπεράσματα σχετικά με τις πιθανότητες μπορεί να εξαχθούν με ένα συνδυασμό υπολογισμών προς τα κάτω (top-down) και προς τα πάνω (bottom-up) στο δέντρο. Διάφοροι αλγόριθμοι έχουν προταθεί με βάση αυτή την ιδέα. Για την περίπτωση γραφημάτων που δεν έχουν πάνω από ένα μονοπάτι μεταξύ δύο κόμβων, δηλαδή απλά συνδεδεμένων γραφημάτων, οι αλγόριθμοι έχουν γραμμική πολυπλοκότητα ως προς τον αριθμό των κόμβων [27] [28].
Παρακάτω φαίνεται ένα κατευθυνόμενο ακυκλικό γράφημα, οι κόμβοι του οποίου αντιστοιχούν στις μεταβλητές s, u, v, x, y, w, z με από κοινού πιθανότητα $P(s, u, v, x, y, w, z)$ [16].



Εικόνα 18 Παράδειγμα Μπεϋζιανού Δικτύου με δενδρική δομή

Η πιθανότητα $P(s,u,v,x,y,w,z)$ προκύπτει αν πολλαπλασιάσουμε όλες τις υπό συνθήκη πιθανότητες που καθορίζουν το δίκτυο. Αν $z = z_0$ είναι η πληροφορία που έχει παρατηρηθεί, θα πρέπει να υπολογιστεί η πιθανότητα $P(s|z = z_0)$. Με βάση τον κανόνα του Bayes έχουμε:

$$P(s|z = z_0) = \frac{P(s|z = z_0)}{P(z = z_0)} = \frac{P(s|z = z_0)}{\sum_s P(s|z = z_0)}$$

Η από κοινού πιθανότητα θα πρέπει να «περιθωριοποιηθεί» ως προς όλες τις δυνατές τιμές των u, v, x, y, w για να βρεθεί η $P(s|z = z_0)$. Δηλαδή:

$$P(s|z = z_0) = \sum_{u,v,x,y,w} P(s,u,v,x,y,w,z = z_0)$$

Για τον υπολογισμό της $P(s|z = z_0)$, θα πρέπει να γίνουν L^5 πράξεις, αν υποθέσουμε ότι κάθε μία από τις διακριτές μεταβλητές μπορεί να πάρει L τιμές [26]. Εάν έχουμε περισσότερες μεταβλητές και μεγάλο σύνολο τιμών L, ο αριθμός των υπολογισμών γίνεται απαγορευτικά μεγάλος. Εάν αξιοποιηθεί η δομή του Μπεϋζιανού Δικτύου, μπορεί να ελαττωθεί το υπολογιστικό κόστος. Με τη χρήση του κανόνα της αλυσίδας έχουμε:

$$\sum_{u,v,x,y,w} P(s,u,v,x,y,w,z = z_0) =$$

$$\sum_{u,v,x,y,w} P(s)P(u|s)\ P(v|s)P(w|v)P(x|u)\ P(y|u)P(z = z_0|y) =$$

$$P(s)\sum_{u,v} P(u|s)\ P(v|s)\sum_{w} P(w|v)\sum_{x} P(x|u)\sum_{y} P(y|u)P(z = z_0|y)$$

Όπου:

$$\sum_{w} P(w|v) = v$$

$$\sum_{x} P(x|u) = u$$

$$\sum_{y} P(y|u)P(z = z_0|y) = u$$

$$\sum_{u,v} P(u|s)\ P(v|s) \sum_{w} P(w|v) \sum_{x} P(x|u) \sum_{y} P(y|u)P(z = z_0|y)\ = s$$

ή αλλιώς:

$$\sum_{u,v,x,y,w} P(s,u,v,x,y,w,z = z_0) =$$

$$P(s) \sum_{u,v,x,y,w} P(u|s)\ P(v|s)\varphi_1(v)\varphi_2(u)\ \varphi_3(u)$$

Όπου οι ορισμοί των συναρτήσεων $\varphi_i(.), i = 1,2,3$ είναι οι αντίστοιχοι όροι που έχουν σημειωθεί παραπάνω. Για να προκύψει η $\varphi_3(u)$ για μία τιμή της u, θα πρέπει να γίνουν L πράξεις. Άρα, θα χρειαστούν L^2 πράξεις για να υπολογιστεί η $\varphi_3(u)$ για όλες τις τιμές της u. Το ίδιο ισχύει και για τις συναρτήσεις $\varphi_1(v)$, $\varphi_2(u)$. Γενικότερα, το πρόβλημα εξαγωγής συμπερασμάτων πιθανότητας για πολλαπλά δίκτυα είναι NP-δύσκολο, και η προσπάθεια επικεντρώνεται στην εύρεση προσεγγιστικών λύσεων [16].

Για να εκπαιδευτεί ένα Μπεϋζιανό δίκτυο ακολουθούνται δύο στάδια. Η εκμάθηση της τοπολογίας του δικτύου είναι το πρώτο στάδιο, που μπορεί να προκύψει από κάποιον ειδικό στο συγκεκριμένο πεδίο γνώσης ή από τεχνικές βελτιστοποίησης με χρήση δεδομένων εκπαίδευσης ή από τεχνικές εκπαίδευσης. Το δεύτερο στάδιο είναι ο υπολογισμός άγνωστων παραμέτρων, από τα διαθέσιμα δεδομένα εκπαίδευσης [51].

Τα Μπεϋζιανά δίκτυα, σε σχέση με άλλες μεθόδους ταξινόμησης όπως είναι τα Δέντρα Απόφασης ή τα Νευρωτικά Δίκτυα, παρουσιάζουν ένα ενδιαφέρον χαρακτηριστικό. Αυτό είναι η δυνατότητα αξιοποίησης της πληροφορίας που υπάρχει εκ των προτέρων για ένα δεδομένο πρόβλημα με τη μοντελοποίηση των συσχετίσεων μεταξύ των χαρακτηριστικών. Η πληροφορία σχετικά με τη δομή του Μπεϋζιανού δικτύου, μπορεί να προέρχεται από την αξιοποίηση της γνώσης ειδικών ή από γνώση του πεδίου του προβλήματος και μπορεί να χρησιμοποιηθεί ως εξής:

- Δήλωση κόμβου ως κόμβο-ρίζα, δηλαδή ότι ο κόμβος δεν έχει γονείς
- Δήλωση κόμβου ως κόμβο-φύλλο, δηλαδή ότι ο κόμβος δεν έχει κόμβους παιδιά
- Δήλωση ότι κάποιος κόμβος είναι άμεσο αίτιο ή άμεσο αποτέλεσμα κάποιου άλλου κόμβου

- Δήλωση ότι κάποιος κόμβος δεν συνδέεται άμεσα με κάποιον άλλο κόμβο
- Δήλωση ότι δύο κόμβοι είναι ανεξάρτητοι, η οποία προκείπτει από ένα σύνολο συνθηκών
- Προσδιορισμός μερικής διάταξης των κόμβων, δηλαδή δήλωση ότι ένας κόμβος εμφανίζεται νωρίτερα από ένα άλλο κόμβο στη διάταξη.

Ένα πρόβλημα των Μπεϋζιανών δικτύων είναι ότι σε σύνολα δεδομένων με πολλά χαρακτηριστικά, δεν είναι κατάλληλα, επειδή θα πρέπει να κατασκευαστεί ένα πολύ μεγάλο δίκτυο, το οποίο δε θα είναι αποδοτικό ως προς το χρόνο εκτέλεσης και το χώρο αποθήκευσης [17].

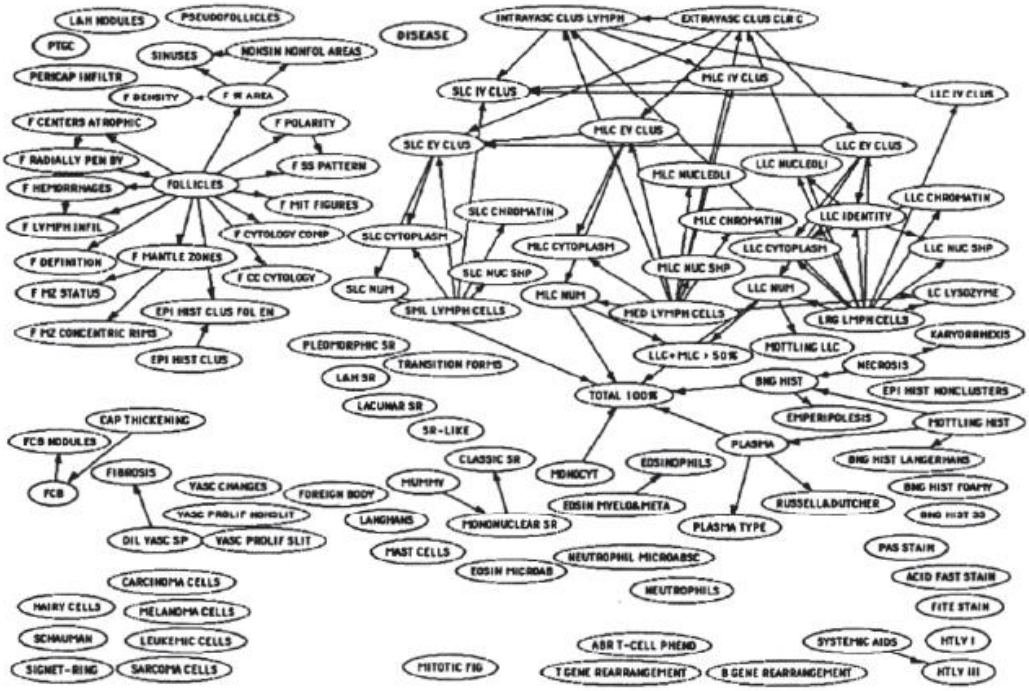
2.5 Εφαρμογές των Μπεϋζιανών Δικτύων στη Βιοπληροφορική

Τα Μπεϋζιανά Δίκτυα είναι κατάλληλα για την έκφραση συλλογιστικών διαδικασιών που εμπεριέχουν την έννοια της αβεβαιότητας. Είναι ιδιαίτερα δημοφιλή στον χειρισμό γνώσης με αβεβαιότητα, η οποία προκύπτει από διαγνώσεις ασθενειών, από τη διαδικασία επιλογής βέλτιστων μεθόδων θεραπείας και από προβλέψεις του αποτελέσματος της θεραπείας, σε διάφορες περιπτώσεις ασθενών. Γενικά, έχουν χρησιμοποιηθεί αρκετά για εφαρμογές που αφορούν την Υγεία. Επίσης, χρησιμοποιούνται σε εφαρμογές που δε σχετίζονται άμεσα με τη διαχείριση δεδομένων για παθήσεις συγκεκριμένων ασθενών. Στην κλινική επιδημιολογία, για την κατασκευή μοντέλων για συγκεκριμένες ασθένειες και για την ερμηνεία δεδομένων από μικροσυστοιχίες γενετικής έκφρασης (microarray gene expression data) χρησιμοποιούνται Μπεϋζιανά Δίκτυα [41].

Τα Μπεϋζιανά Δίκτυα έχουν χρησιμοποιηθεί σε πολλές εφαρμογές σχεδίασης ιατρικών έμπειρων συστημάτων (medical expert systems). Πολύ γνωστό είναι το σύστημα MUNIN για διάγνωση νευρο-μυϊκών διαταραχών, το σύστημα NESTOR για τη διάγνωση ενδοκρινολογικών διαταραχών, και το σύστημα ALARM για την παρακολούθηση ασθενών σε Μονάδες Εντατικής Θεραπείας [25].

Επίσης, υπάρχει το σύστημα PATHFINDER το οποίο έχει χρησιμοποιηθεί για τη διάγνωση ασθενειών του λεμφικού συστήματος και σχεδιάστηκε με σκοπό να βοηθήσει τους παθολόγους στη διάγνωση στο πεδίο της αιματολογίας. Το σύστημα λειτουργεί πολύ απλά, με το χρήστη να εισάγει τιμές για ένα ή περισσότερα χαρακτηριστικά μίας περιοχής ενός λεμφικού κόμβου (lymph-node section). Χρησιμοποιούνται οι εξής κατηγορίες χαρακτηριστικών: ανοσολογία, διακριτικά χαρακτηριστικά, εργαστηριακά τεστ, φλεγμονώδεις συνιστώσες, μεταστατικά κύτταρα, μοριακή βιολογία, μορφολογία, άλλες διαγνώσεις, πρότυπα (patterns), ειδικές κηλίδες, σφαιρικές δομές. Ένα σύνολο χαρακτηριστικών ομαδοποιείται από κάθε μία από τις κατηγορίες αυτές, για το οποίο σύνολο χαρακτηριστικών ο χρήστης μπορεί να εισάγει κάποια πληροφορία. Έχοντας ως δεδομένο αυτά τα ζευγάρια χαρακτηριστικού- τιμής, εμφανίζεται από το σύστημα μία

διαφορική διάγνωση, στην οποία ταξινομούνται οι διάφορες ασθένειες με βάση την πιθανότητα εμφάνισης τους.



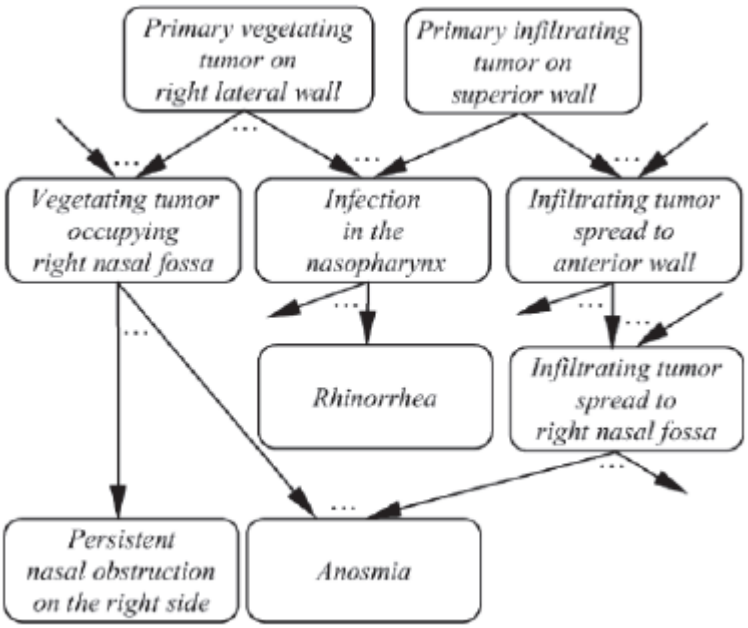
Εικόνα 19 Μπεϋζιανό Δίκτυο του συστήματος PATHFINDER. Ο κόμβος DISEASE περιέχει πάνω από 60 ασθένειες λεμφικών κόμβων.

Στην παραπάνω εικόνα φαίνεται το συνολικό Μπεϋζιανό δίκτυο που χρησιμοποιείται. Ειδικοί αξιολόγησαν την απόδοση του συστήματος, και σε κλίμακα από 0- 10 και ο μέσος όρος απόδοσης ήταν 7.99 (για ανεξάρτητα δεδομένα) και 8.94 (για εξαρτημένα δεδομένα).

Ένα άλλο σύστημα είναι το QMR (Quick Medical Reference), το οποίο χρησιμοποιείται για διάγνωση στην Εσωτερική Παθολογία (Internal Medicine) και βασίζεται στο μοντέλο CPCS (Computer based Patient Case Simulation System). Μία διαφορετική έκδοση αυτού του συστήματος είναι το σύστημα QMR-DT, η οποία εξάγει θεωρητικές αποφάσεις χρησιμοποιώντας αναπαράσταση γνώσης με Μπεϋζιανά Δίκτυα. Το σύστημα αυτό, θεωρεί ένα σύνολο από d_i ασθένειες και ένα σύνολο ευρημάτων F, τα οποία εισάγει ο χρήστης στο σύστημα και στη συνέχεια υπολογίζει τις εκ των υστέρων περιθώριες πιθανότητες $P(d_i^+|F)$, όπου d_i^+ είναι το γεγονός της παρουσίας της ασθένειας d_i. Τα ευρήματα προέρχονται από δεδομένα κλινικών εξετάσεων. Για να περιγράψει καλύτερα την υπό εξέταση περίπτωση, ο χρήστης καλείται να επιλέξει τα κατάλληλα από τα διαθέσιμα ευρήματα. Για να καταγραφεί για παράδειγμα ένα μη φυσιολογικό επίπεδο χοληστερόλης, ο χρήστης θα πρέπει να επιλέξει μεταξύ των δύο δυαδικών ευρημάτων: CHOLESTEROL_BLOOD_INCREASED ή CHOLISTEROL_BLOOD_DECREASED. Μία άλλη ιατρική εφαρμογή που πρέπει να σημειωθεί είναι το σύστημα DIAVAL, το οποίο είναι ένα διαγνωστικό έμπειρο σύστημα για ηχοκαρδιογραφία (echocardiography) [24].

Τα δυναμικά Μπεϋζιανά Δίκτυα (Dynamic Bayesian Networks), είναι μία επέκταση των Μπεϋζιανών Δικτύων ως προς την παράμετρο του χρόνου, συσχετίζοντας μεταβλητές ως προς διαφορετικά χρονικά βήματα, και βρίσκουν και αυτά με τη σειρά τους αρκετές εφαρμογές στην Ιατρική.

Ένα σύστημα σχεδιασμένο για τη διάγνωση και πρόγνωση του ρινοφαρυγγικού καρκίνου (nasopharyngeal cancer) λέγεται NasoNet. Η εξάπλωση του καρκίνου μπορεί να χαρατηριστεί ως μία μη- ντετερμινιστική, δυναμική διεργασία. Το συγκεκριμένο σύστημα χρησιμοποιεί δίκτυα πιθανοτικών γεγονότων σε διακριτό χρόνο (network of probabilistic events in discrete time- NPEDT), δηλαδή Μπεϋζιανά Δίκτυα που χρησιμοποιούνται στη μοντελοποίηση της εξέλιξης μίας διαδικασίας στο χρόνο. Σε κάποιο από τα ρινοφαρυγγικά τείχη μπορεί να υπάρξει ένας αρχικός όγκος, ο οποίος μπορεί να εξαπλωθεί και να εισβάλει σε γειτονικές περιοχές. Ένα σύνολο γεγονότων χαρακτηρίζει τη διαδικασία της εξάπλωσης. Στο σύστημα NasoNet, όλα τα γεγονότα που σχετίζονται με την εξάπλωση του ρινοφαρυγγικού καρκίνου αναπαριστώνται ως κόμβοι ενός Μπεϋζιανού Δικτύου. Το σύστημα προσπαθεί να προβλέψει την κατεύθυνση εξέλιξης του καρκίνου, συλλέγοντας δεδομένα για τα γεγονότα, ώστε να καθοδηγήσει τους ογκολόγους για την εφαρμογή της κατάλληλης θεραπείας του ασθενούς.



Εικόνα 20 Τμήμα Μπεϋζιανού Δικτύου, το οποίο μοντελοποιεί την εξέλιξη του ρινοφαρυγγικού καρκίνου

Στην παραπάνω εικόνα παρουσιάζεται ένα μέρος του Μπεϋζιανού Δικτύου που μοντελοποιεί την εξέλιξη του ρινοφαρυγγικού καρκίνου. Το συγκεκριμένο σύστημα εμφανίζει απόδοση πάνω από 82.7% στο σύνολο ελέγχου που χρησιμοποιήθηκε [18].

Τα δυναμικά Μπεϋζιανά Δίκτυα χρησιμοποιούνται επίσης για τη μοντελοποίηση της δυναμικής συμπεριφοράς κατά την αποτυχία κάποιου οργάνου ασθενή σε Μονάδες Εντατικής Θεραπείας και για την περιγραφή διάφορων κυτταρικών συστημάτων.

Ακόμη, για την πρόγνωση της μελλοντικής κατάστασης των ασθενών και των αποτελεσμάτων από διάφορες θεραπείες, χρησιμοποιείται μία νέα τεχνική στην ιατρική που είναι τα προγνωστικά Μπεϋζιανά Δίκτυα. Οι ερευνητές έχουν ξεκινήσει να αναπτύσσουν Μπεϋζιανά Δίκτυα για αυτό το σκοπό στους τομείς της ογκολογίας και των λοιμωδών νοσημάτων. Στη συγκεκριμένη διαδικασία, γίνεται προσπάθεια ενσωμάτωσης των εννοιών από την παραδοσιακή ανάλυση θνησιμότητας και επιβίωσης σε μεθόδους που χρησιμοποιούν Μπεϋζιανά Δίκτυα [22].

Ένα άλλο σύστημα που στόχος του είναι να παρέχει καινοτόμες ιατρικές υπηρεσίες, ονομάζεται σύστημα CAALYX (Complete Ambient Assisted Living eXperiment). Χρησιμοποιεί σύγχρονες μεθόδους τηλεϊατρικής, με ενίσχυση της δυνατότητας των ηλικιωμένων ομάδων να ζουν αυτόνομα, αξιοποιώντας τη βάση γνώσης των περισσότερο διαδεδομένων διαταραχών υγείας και αλλαγών στα βιόσημα αυτής της ηλικιακής ομάδας. Σε ένα Μπεϋζιανό Δίκτυο για την εκμάθηση από τους συναγερμούς που προκύπτουν κατά την παρακολούθηση των ασθενών και την κατηγοριοποίηση των νέων περιπτώσεων βασίζεται το υποσύστημα τεχνητής νοημοσύνης, που εμπεριέχεται στο παραπάνω σύστημα.

Επίσης, τα Μπεϋζιανά Δίκτυα βρίσκουν εφαρμογή στη διαχείριση των λοιμωδών νοσημάτων στις Μονάδες Εντατικής Θεραπείας. Επίσης, εξαιτίας της αβεβαιότητας που εμπεριέχουν πολλοί παράγοντες σε περιπτώσεις υπηρεσιών υγείας για επείγοντα περιστατικά (emergency medical services), τα Μπεϋζιανά Δίκτυα μπορούν να χρησιμοποιηθούν και για τη μοντελοποίηση τέτοιων υπηρεσιών. Ακόμη, με τη βοήθεια των Μπεϋζιανών Δικτύων, αναπτύσσονται κλινικά μοντέλα για την ταξινόμηση όγκων στις ωοθήκες (ovarian tumors), για περιπτώσεις που περιέχουν πολύ θόρυβο ή τα δεδομένα μπορεί να είναι ελλιπή [23].

Τέλος, τα Μπεϋζιανά Δίκτυα χρησιμοποιούνται στην ταξινόμηση δευτεροταγών δομών πρωτεϊνών, στη μοντελοποίηση των ευρημάτων από μικροσυστοιχίες (microarrays) και στην επικύρωσή τους με τη μέθοδο RT-PCR (real time reverse transcriptase- polymerase chain reaction), στη μοντελοποίηση για τη συνύπαρξη παθολογίας μαστού και στην δημιουργία προγνωστικών μοντέλων για διαχείριση ασθενών [18].

Κεφάλαιο 3ο

3 Μπεϋζιανό Μοντέλο Απόφασης για τη διάγνωση ασθενειών

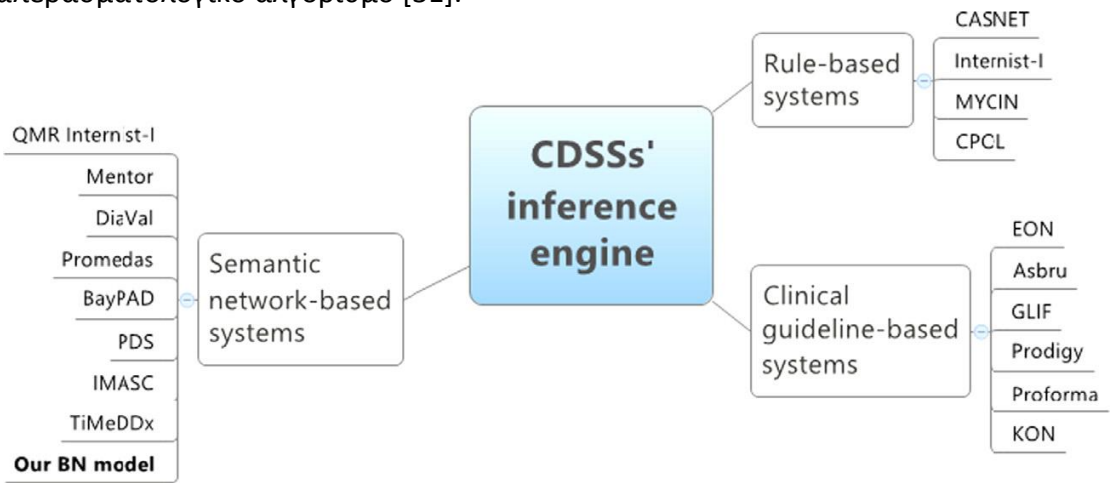
Στόχος του ανθρώπου, είναι να βελτιώσει όσο γίνεται τη διάγνωση ασθενειών με οποιοδήποτε τρόπο. Έτσι, τα διαγνωστικά λάθη είναι μία πολύ σημαντική πηγή πληροφορίας, η οποία μπορεί να αποτρέψει βλάβες της υγεία μας. Ένα διαγνωστικό λάθος μπορεί να θεωρηθεί ως μια διάγνωση η οποία δεν έχει γίνει σωστά ή καθυστέρησε, και η οποία ανιχνεύεται λανθασμένα από κάποιο τεστ. Η γνώση και η κατανόηση ιατρικών λαθών έχει προωθήσει την ασφαλέστερη υγειονομική περίθαλψη μέσω λύσεων πληροφοριακών συστημάτων υγείας. Τα Συστήματα Κλινικών Αποφάσεων (Clinical decision support systems) ή CDSS, αποτελούν μία σημαντική κατηγορία συστημάτων που παρέχουν πληροφορίες για την υγεία μας (health information systems), τα οποία έχουν σχεδιαστεί για να καλυτερέψουν τον τρόπο με τον οποίο παίρνονται οι κλινικές αποφάσεις. Χαρακτηριστικά από μεμωνομένους ασθενείς, ταιριάζουν με χαρακτηριστικά από μία βάση δεδομένων, και ένας αλγόριθμος παράγει εκτιμήσεις και συστάσεις ειδικά για τον ασθενή. Μελέτες έχουν δείξει ότι τα συστήματα CDSS μπορούν να μειώσουν τα διαγνωστικά ποσοστά λάθους [30].

Το Μπεϋζιανό Μοντέλο Απόφασης (BN) μπορεί να να χρησιμοποιηθεί για την κατασκευή των συστημάτων υποστήριξης κλινικών αποφάσεων (CDS) για να βοηθήσει στη διάγνωση των ασθενειών αυτών. Τα BNs είναι καλά προσαρμοσμένα για την εκπροσώπηση της αβεβαιότητας και της αιτιότητας, που είναι και οι δύο παρόντες που υπάρχουν στον κλινικό τομέα.

Η δομή του Δικτύου συνήθως χτίζεται με βάση τα τρέχοντα διαγνωστικά κριτήρια και τη συμβολή από γιατρούς οι οποίοι είναι ειδικοί σε αυτόν τον τομέα. Οι παράμετροι του Δικτύου υπολογίζονται χρησιμοποιώντας έναν αλγόριθμο μάθησης από ένα σύνολο δεδομένων από πραγματικές κλινικές περιπτώσεις.

Τα CDSSs αναλυτικότερα, είναι υπολογιστικά συστήματα που έχουν σχεδιαστεί για την υποστήριξη υψηλού επιπέδου γνωστικών λειτουργιών που αφορούν κλινική διάγνωση, όπως η αιτιολογία, η δημιουργία απόφασης και η μάθηση. Τα συστήματα CDSS μπορούν να σχεδιαστούν για μια σειρά από προφίλ χρηστών, συμπεριλαμβανομένων των παθολόγων, γενικών ιατρών, νοσηλευτών, ή ακόμα και ασθενών που αναζητούν πρόληψη ή άλλες συμπεριφορές που σχετίζονται με την υγεία. Επίσης, τα συστήματα CDSSs ομαδοποιούνται μαζί με άλλα συστήματα, συμπεριλαμβανομένων των συστημάτων που βασίζονται σε κανόνες (rule-based systems), τα συστήματα που βασίζονται σε κλινικές κατευθυντήριες οδηγίες (clinical guideline-based systems) και σημασιολογικά συστήματα (semantic network-based systems). Τα συστήματα που βασίζονται σε κανόνες

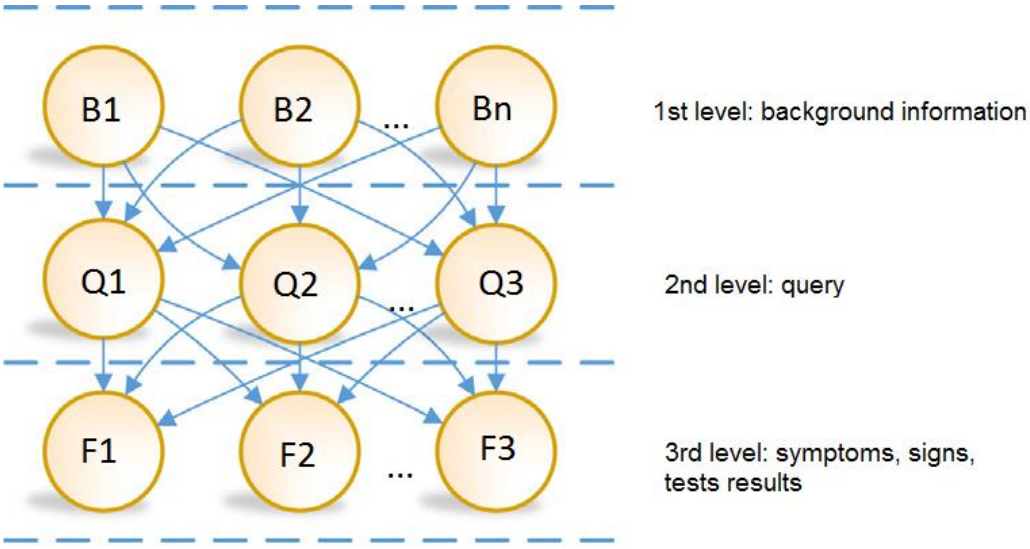
χρησιμοποιούν απλές εκφράσεις (π.χ. if- then- else) για την εξαγωγή των κλινικών αποφάσεων. Τα συστήματα που βασίζονται σε κλινικές κατευθυντήριες οδηγίες δείχνουν την πιο πιθανή κλινική απόφαση ή το μονοπάτι από ένα σύνολο προκαθορισμένων επιλογών, που καθοδηγείται από μια ροή εργασίας που περιγράφει κανόνες διάγνωσης ή θεραπευτικής διαδικασίας. Τα συστήματα που χαρακτηρίζονται ως σημασιολογικά χρησιμοποιούν σημασιολογικές σχέσεις μεταξύ των εννοιών για να εκτελέσει τον συμπερασματολογικό αλγόριθμο [31].



Εικόνα 211 Κλινικά Συστήματα Υποστήριξης Αποφάσεων (CDSSs) που ομαδοποιούνται ανάλογα με το μηχανισμό εξαγωγής συμπερασμάτων τους

Το σύστημα QMR (Quick Medical Reference) / Internist-Iis είναι ένα από τα πρώτα συστήματα που χρησιμοποιούν έναν μηχανισμό εξαγωγής συμπερασμάτων με βάση ένα Μπεϋζιανό Δίκτυο. Είναι δηλαδή ένα πιθανοτικό γραφικό μοντέλο που αντιπροσωπεύει ένα σύνολο τυχαίων μεταβλητών και τις εξαρτήσεις τους σε ένα μη κατευθυνόμενο άκυκλο γράφημα (DAG). Στο QNR χρησιμοποιείται ένα Μπεϋζιανό Δίκτυο με τη δομή δύο επιπέδων που εκπροσωπούν τις πιθανολογικές σχέσεις μεταξύ ασθενειών και συμπτωμάτων. Οι δεσμευμένες πιθανότητες των συμπτωμάτων και ασθενειών εκτιμούνται συνδυάζοντας πιθανότητες από τα προφίλ της νόσου και στατιστικά στοιχεία από τα νοσοκομεία, και οι στατιστικές παράμετροι απορρέουν από τα συστήματα πληροφοριών για την υγεία και την κρίση των ανθρώπινων εμπειρογνωμόνων από τον αντίστοιχο τομέα της νόσου. Επιπλέον, το σύστημα QMR χρησιμοποιεί ένα OR μοντέλο, που περιγράφεται από τον Περλ, για να απλοποιηθεί η δεσμευμένη πιθανοτική εκτίμηση. Ως εκ τούτου, με βάση τα δεδομένα συμπτώματα, το δίκτυο μπορεί να χρησιμοποιηθεί για να υπολογίσει τις πιθανότητες της παρουσίας διαφόρων ασθενειών με μία κατάλληλη μηχανή διεξαγωγής συμπερασμάτων, υποστηρίζοντας τη διαγνωστική διαδικασία του γιατρού. Μια εκδοχή του QMR δικτύου που περιγράφεται από τον Pradhan, περιλαμβάνει back-ground κόμβους, τους οποίους ονόμασαν «προδιαθεσικούς» παράγοντες (predisposing factors), που χρειάζονται εκ των προτέρων πιθανότητες, ενώ οι υπόλοιποι κόμβοι που υπάρχουν, απαιτούν πιθανότητες που να αξιολογούν κάθε μία από τις τιμές τους. Έτσι, ονομάζεται Μπεϋζιανή δομή τριών επιπέδων, μία δομή όπου κάθε επίπεδο αντιπροσωπεύει (1) βασικές πληροφορίες, (2) ασθένειες και (3), τα συμπτώματα, τα σημεία και τα αποτελέσματα νευροψυχολογικών τεστ, αντίστοιχα [32].

Στην παρακάτω εικόνα φαίνεται ένα τέτοιο Δίκτυο.



Εικόνα 22 Παράδειγμα Μπεϋζιανού Δικτύου τριών επιπέδων

Μια δυσκολία στη χρήση μιας τέτοιας δομής μοντέλων είναι η εκτίμηση των δεσμευμένων πιθανοτήτων που αντανακλούν στις πραγματικές κλινικές περιπτώσεις και, ταυτόχρονα, η αντοιστίχηση των τυχαίων μεταβλητών ανεξαρτήτως των OR μοντέλων.
Γενικότερα, υπάρχουν διάφορα CDSS συστήματα. Για παράδειγμα, το Mentor είναι ένα CDSS σύστημα που προβλέπει τη νοητική καθυστέρηση στα νεογέννητα. Το συγκεκριμένο σύστημα βασίζεται σε ένα Μπεϋζιανό Δίκτυο του οποίου η δομή έχει ανακαλυφθεί από δεδομένα που χρησιμοποιούν ένα αλγόριθμο που προτείνεται και επικυρώνεται από ειδικούς του χώρου. Το DiaVal είναι ένα CDSS σύστημα που χρησιμοποιείται για τη διάγνωση των καρδιαγγειακών παθήσεων χρησιμοποιώντας ένα Μπεϋζιανό Δίκτυο. Η δομή του δικτύου χτίστηκε από ειδικούς του χώρου, χρησιμοποιώντας μια αιτιακή αναπαράσταση της καρδιακής παθοφυσιολογίας και, κατόπιν, ενσωματώθηκαν κάποια αποτελέσματα, κυρίως από ηχοκαρδιογραφία.

Ο κ. Σακελλαρόπουλος και ο κ. Νικηφορίδης [64], το 2000, σε εργασία που δημοσίευσαν με τίτλο «Decision support system based on Bayesian networks for the prognosis of patients with head injuries», χρησιμοποίησαν ένα BN με διακριτή κατανομή πιθανότητας για την εκτίμηση της πρόγνωσης μετά από 24 ώρες, για ασθενείς οι οποίοι είχαν τραύματα στο κεφάλι. Η δομή του BN και οι παράμετροι που αντλήθηκαν από τις περιπτώσεις των ασθενών και τα αποτελέσματα πρόγνωσης συγκρίθηκαν με εκείνες που γίνονται από έναν εμπειρογνώμονα. Πέτυχαν ένα ποσοστό επιτυχίας κοντά στο ποσοστό επιτυχίας του εμπειρογνώμονα [64]. Ο Burnside [65] έχτισε ένα BN για υποβοήθηση των ακτινολόγων, στο κομμάτι της λήψης των αποφάσεών τους, ενσωματώνοντας συμπεράσματα μαστογραφίας με χρήση BI-RADS (Breast Imaging Reporting and Data System), ως

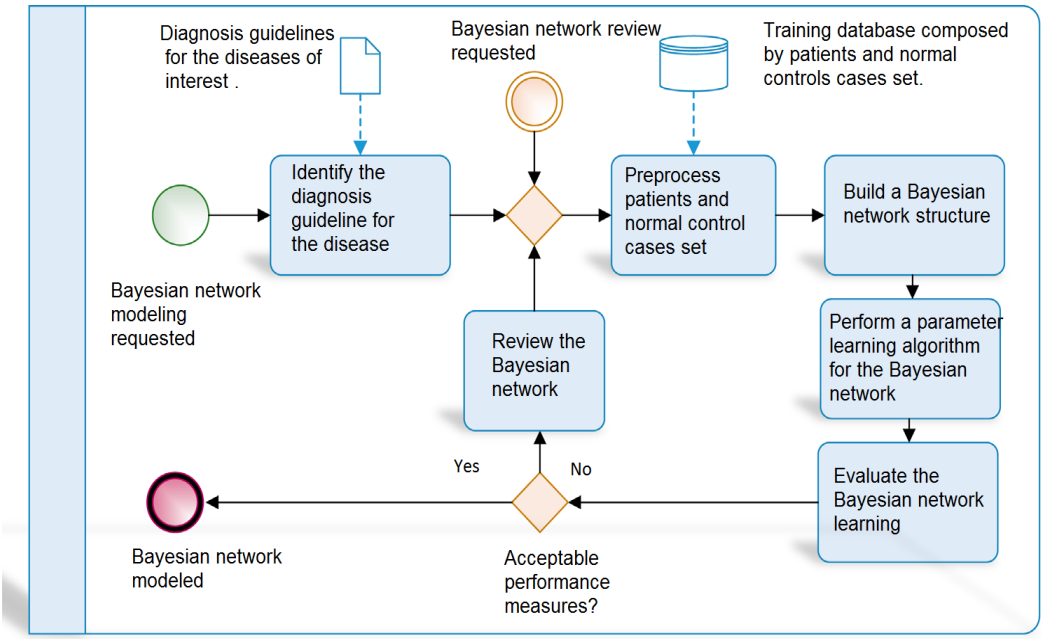
τυποποιημένο λεξιλόγιο που αναπτύχθηκε για μαστογραφία.Στο BN μοντέλο πιθανοτήτων παρέχονται οι πιθανότητες, ο καρκίνος να είναι καλοήθης, κακοήθης και προκακοήθης [65]. Οι Aronsky και Haug [66] έδειξαν τη μοντελοποίηση και την αξιολόγηση ενός BN για τη διάγνωση της πνευμονίας.

Μια άλλη δημοφιλής τεχνική για την κατασκευή μοντέλων απόφασης εφαρμόζει πολλαπλά κριτήρια απόφασης υποβοηθούμενων μεθόδων (MCDA). Οι μέθοδοι MCDA περιλαμβάνουν ένα σύνολο μεθόδων για τη συνάθροιση πολλαπλών κριτηρίων αξιολόγησης, κριτήρια που αφορούν μία ή περισσότερες πιθανές δράσεις. Υπάρχουν εννοιολογικές ομοιότητες μεταξύ των μεθόδων MCDA και των Μπεϋζιανών μεθόδων μάθησης. Και οι δύο προοπτικές εξετάζουν το πρόβλημα της εκμάθησης μιας απόφασης από τα δεδομένα, ως μεγιστοποίηση της εμπειρικής συνάρτησης χρησιμότητας (empirical utility function). Σε μια Μπεϋζιανή προοπτική μάθησης, η συνάρτηση χρησιμότητας μπορεί να μεγιστοποιήσει κάποιο σκορ Μπεϋζιανής εκτίμησης, όπως Μεγίστης Πιθανοφάνειας Εκτίμησης (Maximum-Likelihood Estimation- MLE) [33].

Υπάρχουν έργα τα οποία προτείνουν την ενσωμάτωση προσεγγίσεων που βασίζονται στις μεθόδους MCDA και στις στατιστικές μεθόδους μάθησης, δηλαδή, στην εφαρμογή των MCDA εννοιών σε ένα πλαίσιο στατιστής μάθησης και εκμάθησης υβριδικών μεθοδολογιών. Ο Castro [67] έδειξε μια προσέγγιση των MCDA μεθόδων, η οποία ονομάζεται MAOBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique) με ενσωματωμένο Μπεϋζιανό Δίκτυο. Χρησιμοποίησε ένα εκτεταμένο διακριτό Μπεύζιανό Δίκτυο, με χρησιμότητα και κόμβους απόφασης, το λεγόμενο «Influence» Διάγραμμα. Ένα τέτοιο διάγραμμα είναι μια γενίκευση του Μπεϋζιανού Δικτύου, στην οποία υπάρχουν όχι μόνο προβλήματα πιθανολογικών συμπερασμάτων, αλλά και προβλήματα λήψης αποφάσεων τα οποία μπορούν να διαμορφωθούν και να επιλυθούν. Οι κόμβοι χρησιμότητας αντιπροσωπεύουν μια σειρά από στόχους ή προτιμήσεις οι οποίες ορίζονται από αυτόν που καθορίζει τις αποφάσεις. Σε μια προσέγγιση των MCDA μεθόδων, οι κόμβοι του Μπεϋζιανού Δικτύου αντιπροσωπεύουν τους εσωτερικούς ή εξωτερικούς παράγοντες που μπορούν να επηρεάσουν τα κριτήρια απόφασης. Η Μπεϋζιανή δομή τους, επέτρεψε τον υπολογισμό των δεσμευμένων πιθανοτικών πινάκων (CPTs) των Μπεϋζιανών κόμβων από το σύνολο δεδομένων χωρίς να εφαρμόζει μια σύνθετη στατιστική μέθοδο εκμάθησης. Επιπλέον, οι τιμές των γνωρισμάτων που λείπουν αντιμετωπίστηκαν ως μία κατάσταση, η οποία μπορεί να αντικατασταθεί από μια υποθετική τιμή.

Ο Menezes [68] πρότεινε ένα υβριδικό μοντέλο που συνδυάζει τις μεθόδους MCDA και τα Μπεϋζιανά Δίκτυα για τη διάγνωση του Διαβήτη Τύπου 2. Ο Pinheiro [69] χρησιμοποίησε μια παρόμοια προσέγγιση και παρουσίασε ένα μοντέλο κατάταξης με βάση τις μεθόδους MCDA και τα Μπεϋζιανά Δίκτυα για την βελτίωση της διάγνωσης του AD. Στόχος είναι να εξεταστούν ποια στοιχεία του ασθενή έχουν μεγαλύτερο αντίκτυπο στον καθορισμό της διάγνωσης του AD. Το Μπεϋζιανό Δίκτυο τους χτίστηκε με το χέρι (manually), και οι Μπεϋζιανοί κόμβοι σημασιολογικά σχετίζονται με κάθε στοιχείο αξιολόγησης. Γενικά, το

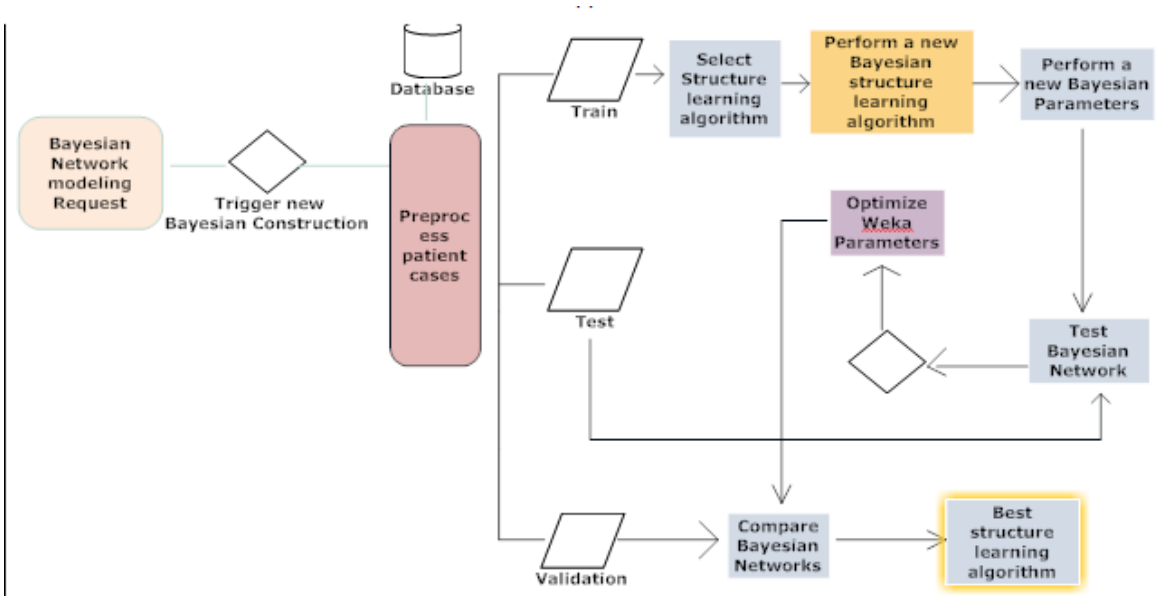
Μπεϋζιανό Δίκτυο χρειάζεται αναθεώρηση κάθε φορά που υπάρχει ένα νέο σύνολο δεδομένων εκπαίδευσης και χρειάζονται αλλαγές στις κατευθυντήριες γραμμές για για γίνει διάγνωση.



Εικόνα 23 Διαδικασία μοντελοποίησης Μπεϋζιανού Δικτύου που περιλαμβάνει εκπαίδευση και evaluate

Με βάση όλα τα παραπάνω λοιπόν, και στη συγκεκριμένη περίπτωση που εξετάζουμε τον καρκίνο του τραχήλου της μήτρας, έχουμε κάποια δεδομένα τα οποία θέλουμε να εξετάσουμε, οπότε αρχικά πρέπει να τα κατηγοριοποιήσουμε ανάλογα με το ποια θα εκπαιδευτούν, θα γίνουν τεστ και θα επικυρωθούν (validate). Αρχικά λοιπόν, χρησιμοποιούμε τους αλγορίθμος εκμάθησης γράφων, για να γίνει train. Στη συνέχεια, και αφού ολοκληρωθεί η εκμάθηση, επιλέγουμε το test set των δεδομένων μας για να κάνουμε τεστ. Σ' αυτή τη φάση, εφαρμόζουμε τον αλγόριθμο που έχουμε επιλέξει στα δεδομένα μας. Ο κάθε αλγόριθμος που εφαρμόζεται στα δεδομένα έχει διάφορες παραμέτρους, τις οποίες επιλέγουμε εμείς κάθε φορά, αφού έχουμε βέβαια μελετήσει τον αλγόριθμο, και ξέρουμε τη συμπεριφορά που θα δούμε να προκύπτει. Ο αλγόριθμος παραμένει ο ίδιος, ενώ οι παράμετροι αλλάζουν κάθε φορά, μέχρι να προκύψουν τα καλύτερα αποτελέσματα. Όπως αναφέρεται παραπάνω, στο τέλος χρησιμοποιούμε ένα διαφορετικό τεστ για το validate, καθώς δεν πρέπει να χρησιμοποιήσουμε το test set. Τελικά, κάθε φορά δημιουργείται ένας διαφορετικός γράφος. Έχοντας επιλέξει έναν συγκεκριμένο αλγόριθμο, και διαφορετικές παραμέτρους, μπορεί να προκύψει ο ίδιος γράφος. Στόχος είναι να βελτιστοποιηθούν οι παράμετροι και να προκύψει ο καλύτερος δυνατός γράφος. Άρα, συγκρίνουμε τους γράφους και κρατάμε τον καλύτερο [34].

Στην παρακάτω εικόνα αναλύεται σχηματικά η διαδικασία που ακολουθούμε.



Εικόνα 24 Διαδικασία Μοντελοποίησης του Μπεϋζιανού Δικτύου, με όλες τις διαδικασίες. Αρχικά γίνεται εκμάθηση του γράφου, στη συνέχεια test και μετά validation. Τελικά προκύπτει ο καλύτερος γράφος.

Στο επόμενο κεφάλαιο αναλύονται οι αλγόριθμοι εκμάθησης γράφων που μελετήθηκαν. Είναι σημαντικό να τονιστεί ό,τι γνωρίζουμε τη συμπεριφορά των αλγορίθμων και με βάση αυτή κάνουμε και την κατάλληλη παραμετροποίηση στο σύστημα.

Κεφάλαιο 4ο

4 Ο αλγόριθμος K2

Ο αλγόριθμος K2 είναι μια τυπική μέθοδος αναζήτησης (search method) και κατάταξης (score method). Ξεκινά με την παραδοχή ότι ένας κόμβος δεν έχει καθόλου γονείς, και σε κάθε βήμα προσθέτει σταδιακά κάθε ένα γονέα, του οποίου η προσθήκη αυξάνει κυρίως τη πιθανότητα της προκύπτουσας δομής. Ο K2 σταματά να προσθέτει γονείς στους κόμβους, όταν η προσθήκη ενός γονέα, δεν μπορεί να αυξήσει την πιθανότητα του δικτύου.
Ο αλγόριθμος K2 χρησιμοποιείται για να μάθουμε την τοπολογία ενός Μπεϋζιανού δικτύου. Για να γίνει κατανοητός ο αλγόριθμος, θα χρησιμοποιήσουμε ένα παράδειγμα. Εξετάζεται λοιπόν μία βάση δεδομένων, η οποία φαίνεται στον παρακάτω πίνακα:

case	x_1	x_2	x_3
1	1	0	0
2	1	1	1
3	0	0	1
4	1	1	1
5	0	0	0
6	0	1	1
7	1	1	1
8	0	0	0
9	1	1	1
10	0	0	0

Εικόνα 25 Παράδειγμα βάσης δεδομένων για την εφαρμογή του αλγορίθμου K2

Στη συγκεκριμένη περίπτωση θεωρούμε ως στόχο κατάταξης (classification target) το $x1$.
Ο αλγόριθμος K2 παρουσίαζεται παρακάτω:

1. Ξεκινάει η διαδικάσια του K2;

2. {Είσοδος: Ένα σύνολο από n κόμβους, μία διάταξη των κόμβων, ένα άνω φράγμα u για τον αριθμό των γονέων a, και μια βάση δεδομένων D που περιλαμβάνει m περιπτώσεις.}
3. {Έξοδος: Για κάθε κόμβο, εμφανίζονται οι γονείς του.}
4. Για i:= 1 έως n κάνουμε
   i. πi := Ø;
   ii. Pold := f(i, πi); {Αυτή η συνάρτηση υπολογίζεται χρησιμοποιώντας την Εξίσωση παρακάτω}
   iii. OKToProceed := true;
5. While OKToProceed and |πi| < u έχουμε
   i. ας είναι ο z ο κόμβος στην Pred(xi) - πi που μεγιστοποιεί την f(i, πi ∪ {z});
   ii. Pnew := f(i, πi ∪ {z});
   iii. αν Pnew > Pold τότε
   iv. Pold := Pnew;
   v. πi := πi ∪ {z};
6. αλλίως OKToProceed := false;
7. end {while};
8. Γράψε (’Κόμβος: ’, xi, ’ Γονιός του xi: ’,πi);
9. end {for};
10. end {K2};

Για τον υπολογισμό της συνάρτησης Pold χρησιμοποιείται ο παρακάτω τύπος:

$$f(i,\pi i) = \prod_{j-1}^{qi} \frac{(ri-1)!}{(Nij+ri-1)!} \prod_{k=1}^{ri} aijk!$$

Όπου

πi: οι γονείς του κόμβου xi

qi: | Øi |

Øi: μία λίστα με όλα τα πιθανά στιγμιότυπα των γονιών του κόμβου xi στη βάση δεδομένων D. Εάν p1,…,ps είναι οι γονείς του xi, τότε το Øi είναι το Καρτεσιανό γινόμενο $\{u_1^{p1}, \dots, u_{rp1}^{p1}\} x \dots x \{u_1^{ps}, \dots, u_{rps}^{ps}\}$, όλων των πιθανών τιμών των στιγμιότυπων p1 έως ps

ri: | vi |

vi: λίστα των πιθανών τιμών των στιγμιότυπων του xi

aijk: ο αριθμός των περιπτώσεων της βάσης δεδομένων D, στις οποίες το χαρακτηριστικό xi αρχικοποιείται με την τιμή της k-στης περίπτωσης, και οι γονείς του xi στο πi, είναι αρχικοποιημένοι με την j-στο στιγμιότυπο του Øi

$Nij = \sum_{k-1}^{ri} aijk$. Αυτό είναι ο αριθμός των περιπτώσεων της βάσης δεδομένων στις οποίες οι γονείς του xi στο πi, έχουν αρχικοποιηθεί με το j-στο στιγμιότυπο του Øi
Η παραπάνω συνάρτηση δηλώνει την πιθανότητα οι γονείς του κόμβου xi, να είναι πi [45] [46] [47].

Κεφάλαιο 5ο

5 Γενετικοί αλγόριθμοι

Οι γενετικοί αλγόριθμοι είναι μία προσαρμοσμένη τεχνική αναζήτησης. Η συγκεκριμένη τεχνική είναι μία στοχαστική μέθοδος αναζήτησης, με την οποία βρίσκουμε βέλτιστες λύσεις σε μεγάλους πολύπλοκους διανυσματικούς χώρους αναζήτησης. Ένας γενετικός αλγόριθμος εφαρμόζεται σε ένα σύνολο από στιγμιότυπα, που ονομάζονται πληθυσμός. Με τη σειρά του, το κάθε στιγμιότυπο, αποτελεί μία κωδικοποίηση των δεδομένων εισόδου, και ονομάζεται χρωμόσωμα. Τα γονίδια, αποτελούν το χρωμόσωμα, και κάθε ένα από αυτά έχει μία δυαδική τιμή που δείχνει εάν υπάρχει ή όχι ένα συγκεκριμένο στοιχείο. Τα άτομα, αναπαρίστανται με τη μορφή συμβολοσειρών, οι οποίες παίζουν το ρόλο των χρωμοσωμάτων των οργανισμών. Οι αλγόριθμοι αυτοί, συνδυάζουν την επιβίωση της ισχυρότερης συμβολοσειράς, με μία τυχαία ανταλλαγή δομημένης δομής [48].

Οι γενετικοί αλγόριθμοι, βασίζονται σε φυσικές λειτουργίες των ζωντανών οργανισμών, και προσπαθούν να τις προσομοιώσουν, δημιουργώντας αλγορίθμους, που μπορούν να χρησιμοποιηθούν για την επίλυση διαφόρων προβλημάτων. Ανήκουν στην ευρύτερη κατηγορία των εξελικτικών αλγορίθμων [37].

Με λίγα λόγια, ο αλγόριθμος θέλει να δημιουργήσει νέες γενιές ατόμων, με το καλύτερο δυνατό «γενετικό υλικό», που μπορεί να είναι κάποια συγκεκριμένα χαρακτηριστικά, τα οποία θέλουμε να επικρατήσουν για κάποιο συγκεκριμένο λόγο [49].
Η «ποιότητα» των ατόμων, δηλαδή η αναζήτηση για την καλύτερη λύση, μετράτε με τη βοήθεια μιας συνάρτησης, που ονομάζεται συνάρτηση καταλληλότητας. Οι λύσεις που έχουν χαμηλές τιμές συνάρτησης καταλληλότητας, σιγά σιγά εξαλείφονται, ενώ οι λύσεις που έχουν υψηλές τιμές συνάρτησης καταλληλότητας, έχουν την ικανότητα να αναπαράγουν νέες λύσεις. Η συνάρτηση καταλληλότητας παρέχει τα κριτήρια για την αξιολόγηση των υποψηφίων προς επιλογή στιγμιοτύπων, και γενικά ελέγχει την επιλογή των καλύτερων λύσεων [53].

Κάθε επανάληψη του αλγορίθμου, οδηγεί στη δημιουργία μίας καινούριας γενιάς ατόμων. Ο αλγόριθμος έχει τρεις βασικούς τελεστές (operators), οι οποίοι τον βοηθούν να μην ανέρχεται σε κορεσμό αποτελεσμάτων γρήγορα και λόγω της στασιμότητας του πληθυσμού. Αυτοί οι τελεστές είναι: επιλογή, διασταύρωση και μετάλλαξη των χρωμοσωμάτων. Ο πληθυσμός, δημιουργείται από ένα σύνολο από τυχαία επιλεγμένα στιγμιότυπα. Στη συνέχεια, τα στιγμιότυπα αξιολογούνται με το κριτήριο της συνάρτησης καταλληλότητας. Δύο στιγμιότυπα επιλέγονται για την επόμενη γενιά, σύμφωνα με την καταλληλότητά τους [50].

Η διασταύρωση (crossover), είναι μία διαδικασία η οποία παράγει ανα-συνδιασμούς ακολουθιών bits, οι οποίοι προκύπτουν με ανταλλαγή τμημάτων μεταξύ ζευγών χρωμοσωμάτων για τη δημιουργία νέων στιγμιότυπων. Για να εξασφαλιστεί ότι όλα τα πιθανά χρωμοσώματα είναι δυνατόν να προκύψουν ή ότι έχει περάσει συγκεκριμένος αριθμός γενεών, υπάρχει η διαδικασία της μετάλλαξης (mutation).

Η γενική μορφή του αλγορίθμου είναι η εξής:

Γενετικός_Αλγόριθμος (Καταλληλότητα, Όριο_Καταλληλότητας, N, r, m)

1. Επιλέγεται ο αρχικός πληθισμός που θα εξεταστεί, Π, με N υποψήφιες λύσεις.
2. Προσδιορίζεται η «ποιότητα»- καταλληλότητα κάθε λύσης.
3. Όσο δεν ισχύει κάποια συνθήκη τερματισμού:

i. Επανέλαβε $N/2$ φορές τα ακόλουθα βήματα:

(i) Επέλεξε δύο λύσεις από τον πληθυσμό Π.

(ii) Συνδύασε τις δύο λύσεις για να βγάλεις δύο απογόνους.

(iii) Υπολόγισε την καταλληλότητα των δύο απογόνων.

ii. Δημιούργησε το νέο πληθυσμό Π', έχοντας υπόψη όλους τους νέους απογόνους που προέκυψαν, και θέσε $\Pi = \Pi'$.

4. Τα παιδιά με πιθανότητα κοντά στο 0 «μεταλλάσονται».
5. Κάποια άτομα αφαιρούνται από τον πληθυσμό, με βάση κάποιο κριτήριο αφαίρεσης- επιλογής [48].

Κεφάλαιο 6ο

6 Ο αλγόριθμος Hill climbing

Στην επιστήμη των υπολογιστών, ο αλγόριθμος Hill climbing είναι μια μαθηματική τεχνική βελτιστοποίησης που ανήκει στην οικογένεια των αλγορίθμων αναζήτησης (local search). Είναι ένας επαναληπτικός αλγόριθμος που ξεκινά με μία αυθαίρετη λύση σε ένα πρόβλημα, και στη συνέχεια προσπαθεί να βρει μια καλύτερη λύση με το να αλλάζει ένα μόνο στοιχείο της λύσης κάθε φορά. Αν η αλλαγή δημιουργεί μια καλύτερη λύση, μία σταδιακή αλλαγή γίνεται στη νέα αυτή λύση, και η διαδικασία επαναλαμβάνεται μέχρι να βρεθούν περαιτέρω βελτιώσεις.

Ο αλγόριθμος Hill climbing είναι καλός στο να βρίσκει μία τοπική βέλτιστη λύση, δηλαδή μία λύση που δεν μπορεί να βελτιωθεί άλλο σε μία γειτονική περιοχή, αλλά δεν είναι απαραίτητο ότι η συγκεκριμένη λύση είναι η καλύτερη δυνατή (the global optimum) από όλες τις πιθανές λύσεις στο χώρο αναζήτησης [58].

Η τοπική αναζήτηση αφορά προβλήματα στόχου. Σε τέτοια προβλήματα μας ενδιαφέρει ο στόχος και όχι η διαδρομή προς το στόχο. Οι αλγόριθμοι τοπικής αναζήτησης παρακολουθούν μόνο μία τρέχουσα κατάσταση, γενικά μετακινούνται μόνο σε γειτονικές καταστάσεις και δεν κάνουν συστηματική αναζήτηση. Τα πλεονεκτήματα τους είναι ότι χρησιμοποιούν ελάχιστη (σταθερή) μνήμη και βρίσκουν καλές λύσεις σε πολύ μεγάλους ή άπειρους χώρους. Ουσιαστικά ο συγκεκριμένος αλγόριθμος χρησιμοποιείται σε προβλήματα όπου πρέπει να βρεθεί μία λύση πολύ γρήγορα, έστω και αν αυτή δεν είναι η καλύτερη [60].

Τα βασικά βήματα του αλγορίθμου είναι:

1. Η αρχική κατάσταση είναι η τρέχουσα κατάσταση.
2. Αν η κατάσταση είναι μία τελική τότε ανέφερε τη λύση και σταμάτησε.
3. Εφάρμοσε τους τελεστές μετάβασης για να βρεις τις καταστάσεις- παιδιά.
4. Βρες την καλύτερη κατάσταση σύμφωνα με την ευριστική συνάρτηση.
5. Η καλύτερη κατάσταση γίνεται η τρέχουσα κατάσταση.
6. Πήγαινε στο βήμα 2.

Ο ψευδοκώδικας για τον αλγόριθμο είναι [59]:

1. algorithm hc (InitialState, FinalState)
2. begin
3. CurrentState← InitialState;

4. While CurrentState≠ FinalState κάνε
    (i) Children←Expand (CurrentState);
    (ii) If Children=Ø τότε επέστρεψε failure;
    (iii) EvaluatedChildren←Heuristic (Children);
    (iv) bestChild←best (EvaluatedChildren);
    (v) if hValue (CurrentState) ≥ hValue (bestChild)
    a. then return failure;
    b. else CurrentState←bestChild;
    (vi) endif;
5. endwhile;
6. return success;
7. end

Κεφάλαιο 7ο

7 Ο Αλγόριθμος Tabu Search

Οι τρέχουσες εφαρμογές του TS (Tabu Search) εκτείνονται σε διάφορους τομείς όπως είναι η διαχείριση πόρων, οι τηλεπικοινωνίες, η σχεδίαση συστημάτων VLSI, η οικονομική ανάλυση, ο σχεδιασμός (scheduling), ο σχεδιασμός χώρου (space planning), η κατανομή ενέργειας, η μοριακή εφαρμοσμένη μηχανική, τα λογιστικά, η ταξινόμηση σχεδίων (pattern classification), η εύκαμπτη κατασκευή (flexible manufacturing), η διαχείριση αποβλήτων, η εξερεύνηση ορυκτών (mineral exploration), η βιοϊατρική ανάλυση, η περιβαλλοντική συντήρηση κ.ά [54].

Ο αλγόριθμος TS αναπτύχθηκε για να λύσει συνδυαστικά προβλήματα βελτιστοποίησης. Χαρακτηρίζεται από τη χρήση μιας ευπροσάρμοστης μνήμης, είναι ικανός να εξαλείψει το τοπικό ελάχιστο και να ψάξει περιοχές πέρα από ένα τοπικό ελάχιστο και ουσιαστικά είναι ένα είδος επαναληπτικής αναζήτησης. Με λίγα λόγια, έχει την ικανότητα να βρει το συνολικό ελάχιστο σε έναν πολύμορφο (multimodal) χώρο αναζήτησης.

Η διαδικασία με την οποία γίνεται η αναζήτηση tabu βασίζεται με μια συνάρτηση εκτίμησης που επιλέγει την υψηλότερη εκτίμηση λύσης σε κάθε επανάληψη [44].

Η συνάρτηση εκτίμησης με τη σειρά της, επιλέγει την κίνηση που παράγει την καλύτερη εκτίμηση ή την ελάχιστη χειροτέρευση στην αντικειμενική συνάρτηση. Χρησιμοποιείται μια λίστα tabu για την αποθήκευση των χαρακτηριστικών των αποδεκτών κινήσεων έτσι ώστε αυτά τα χαρακτηριστικά να μπορούν να χρησιμοποιηθούν για την ταξινόμηση ορισμένων κινήσεων ως tabu (π.χ. για να αποφευχθούν) σε επόμενες επαναλήψεις. Με άλλα λόγια, η λίστα tabu καθορίζει ποιες λύσεις μπορούν να προσεγγιστούν με μια κίνηση από την τρέχουσα λύση. Αφού οι κινήσεις που δεν οδηγούν σε βελτιώσεις είναι αποδεκτές στην αναζήτηση tabu, είναι πιθανό να επιστρέψει σε λύσεις που έχουν ήδη προσπελαστεί. Η λίστα tabu χρησιμοποιείται για να αντιμετωπίσει αυτό το πρόβλημα. Μια στρατηγική που ονομάζεται στρατηγική απαγόρευσης (forbidding strategy) χρησιμοποιείται για τον έλεγχο και την ενημέρωση της λίστας tabu. Χρησιμοποιώντας την στρατηγική απαγόρευσης, ένα μονοπάτι που έχει επισκεφθεί (προσπελαστεί) προηγούμενα αποφεύγεται και νέες περιοχές του χώρου αναζήτησης εξερευνώνται [55].

Πριν περιγραφεί η μέθοδος του TS, θα γίνει μία παρουσίαση της τοπικής αναζήτησης καθόδου (LS) που είναι μία ευριστική μέθοδος (είναι επίσης γνωστή ως hill climbing, η οποία αναλύθηκε στο κεφάλαιο 6). Ο αλγόριθμος καθόδου LS ξεκινά από μια αρχική (ίσως, τυχαία παραγόμενη) λύση s. Η διαδικασία αναζήτησης συνεχίζεται με την εκτέλεση μερικών διαδοχικών μετασχηματισμών των λύσεων. Μια κίνηση εφαρμόζεται στην

τρέχουσα λύση s προκειμένου να αποκτηθεί μια νέα λύση s' από τη γειτονιά της τρέχουσας λύσης $\Theta(s)$. Εάν η απόφαση είναι «θετική», τότε η τρέχουσα λύση αντικαθίσταται από την γειτονική, η οποία θα χρησιμοποιηθεί ως «αφετηρία» για τις επόμενες δοκιμές, διαφορετικά, η αναζήτηση συνεχίζεται με την τρέχουσα λύση. Η όλη διαδικασία ολοκληρώνεται όταν η τρέχουσα λύση γίνεται τοπικά βέλτιστη.

Ο αλγόριθμος TS προέρχεται κατά κάποιο τρόπο από την πολιτική που περιγράφεται παραπάνω. Πιο συγκεκριμένα όμως, ο TS υπερβαίνει αυτή την μέθοδο. Σε αντίθεση με τον κλασσικό LS (που περιορίζεται να βρει μια τοπικά βέλτιστη λύση μόνο), οι αλγόριθμοι που βασίζονται στον TS συνεχίζουν την αναζήτηση ακόμα κι αν μια τοπικά βέλτιστη λύση βρεθεί. Με συντομία, ο TS είναι μια διαδικασία διαδοχικών κινήσεων από το ένα τοπικό βέλτιστο στο άλλο. Το καλύτερο τοπικό βέλτιστο που θα βρεθεί κατά τη διάρκεια αυτής της διαδικασίας είναι η προκύπτουσα λύση του TS. Κατά συνέπεια, ο TS είναι μια εκτεταμένη τοπική αναζήτηση καθόδου. Ο TS επιτρέπει τη διαφυγή από τα τοπικά βέλτιστα. Συνεπώς, ερευνά πολύ μεγαλύτερο μέρος του διαστήματος λύσης σε σύγκριση με το LS. Ως εκ τούτου, ο TS προσφέρει περισσότερες ευκαιρίες για την ανακάλυψη υψηλής ποιότητας λύσεων σε σχέση με τον παραδοσιακό LS [56].

Η κεντρική ιδέα της μεθόδου TS, είναι ότι επιτρέπει τις κινήσεις ανόδου (climbing) όταν δεν υπάρχει καμία βελτιωμένη γειτονική λύση. Φυσικά, η επιστροφή στις τοπικά βέλτιστες λύσεις που έχουν επισκεφτεί προηγουμένως απαγορεύεται προκειμένου να αποφευχθεί η ανακύκλωση (cycling) της αναζήτησης. Ο TS βασίζεται, σε μια μεθοδολογία απαγορεύσεων: μερικές κινήσεις είναι «παγωμένες» (γίνονται «tabu») κατά διαστήματα [57].

Τα βήματα του αλγορίθμου φαίνονται παρακάτω:

1. **function** *tabu_search*(*s*); //*input*: *s* – η αρχική λύση, *output*: s^* – η καλύτερη λύση που έχει βρεθεί//
2. s^* := *s*; //αρχικοποίησε την λίστα *tabu* ***T***;
3. **repeat** // συνέχισε τον κύριο κύκλο (*cycle*) του TS // με δεδομένα τη συνάρτηση γειτονιάς ***Θ*** και την λίστα tabu ***T*** βρες την καλύτερη δυνατή λύση $\boldsymbol{s' \in \Theta'(s) \subseteq \Theta(s)}$, όπου το $\boldsymbol{\Theta'(s)}$ αποτελείται από λύσεις (ή τα "χαρακτηριστικά" τους) που δεν είναι στην λίστα *tabu* ***T*** ή ικανοποιούν το κριτήριο φιλοδοξίας;
4. *s* := *s*'; // αντικατέστησε την τρέχουσα λύση με την καινούρια // βάλε την λύση *s* (ή το "χαρακτηριστικό" του) στην λίστα *tabu* ***T***;

   i. **if** $f(s) < f(s^*)$ **then** s^* := *s*; // σώσε την καλύτερη λύση μέχρι στιγμής //

ii. ενημέρωσε την λίστα *tabu* ***T***

iii. **until** να ικανοποιηθεί το κριτήριο τερματισμού **return** *s.*[*]

Κεφάλαιο 8ο

8 Ο αλγόριθμος Simulated annealing

Ο αλγόριθμος Simulated annealing είναι ένας αλγόριθμος που χρησιμοποιείται κυρίως για την επίλυση προβλημάτων βελτιστοποίησης, όπως είναι η βελτιστοποίηση κάποιων συναρτήσεων που έχουν πολλές μεταβλητές. Γενικά, συνδυάζει ιδέες από διαφορετικές μεθοδολογικές προσεγγίσεις, εμπλουτίζοντάς τες με ορισμένα πρωτότυπα στοιχεία.
Το όνομα του προέρχεται από την ανόπτηση (annealing) στη μεταλλουργία. Η συγκεκριμένη διαδικασία χρησιμοποιείται στη μεταλλουργία για να μαλακώσουμε ή να σκληρύνουμε μέταλλα και γυαλί θερμαίνοντας τα σε υψηλή θερμοκρασία και στη συνέχεια ψύχοντας τα σταδιακά, επιτρέποντας έτσι στο υλικό να στερεοποιηθεί σε μία κρυσταλλική κατάσταση χαμηλής ενέργειας.

Η ανακάλυψη του αλγορίθμου simulated annealing είναι ένα παράδειγμα της χρήσης των ιδεών της στατικής μηχανικής (statistical mechanics), μία περιοχή της φυσικής συμπυκνωμένης ύλης, για τη χρήση μεγάλων και πολύπλοκων προβλημάτων βελτιστοποίησης.

Η βασική ιδέα είναι ότι σε κάθε επανάληψη, επιλέγεται μία τυχαία κίνηση. Αν βελτιώνει την κατάσταση τότε η κίνηση είναι αποδεκτή, αλλιώς γίνεται αποδεκτή με κάποια πιθανότητα μικρότερη από 1. Η πιθανότητα μειώνεται εκθετικά ως προς την ακαταλληλότητα της κίνησης. Επίσης, μειώνεται σύμφωνα με μία παράμετρο θερμοκρασίας Τ. Αρχικά ξεκινά με μια μεγάλη τιμή της Τ και στη συνέχεια η Τ μειώνεται σταδιακά. Σε μεγάλες τιμές της Τ, ο αλγόριθμος μοιάζει με καθαρή τυχαία αναζήτηση. Κατά το τέλος του αλγορίθμου, όταν οι τιμές της Τ είναι αρκετά μικρές, ο αλγόριθμος μοιάζει με Climbing Hill.

Ο ψευδοκώδικας για το συγκεκριμένο αλγόριθμο παρουσιάζεται παρακάτω:

1. συνάρτηση Simulated-Annealing(problem; schedule)
2. επιστρέφει μια κατάσταση

3. είσοδοι: problem, a problem
- schedule, a mapping from time to «temperature»

4. τοπικές μεταβλητές: current, a node next, a node T, the temperature
5. current $\leftarrow$ MakeNode(RandomState[problem])
6. για t $\leftarrow$ 1 to ∞ κάνε
- T $\leftarrow$ schedule[t]
- εάν T = 0 τότε επέστρεψε current
- next $\leftarrow$ ένα τυχαία επιλεγμένο successor of current
- ΔE $\leftarrow$ Value[next]- Value[current]
- εάν ΔE > 0 τότε current $\leftarrow$ next
- αλλιώς current $\leftarrow$ next only με πιθανότητα $e^{\frac{\Delta E}{T}}$

Η παράσταση $e^{\frac{\Delta E}{T}}$ προέρχεται από τη θεωρία της στατιστικής μηχανικής:
Η πιθανότητα να βρούμε ένα φυσικό σύστημα σε μία κατάσταση ενέργειας E είναι ανάλογη της συνάρτησης των Gibbs- Boltzmann $e^{\frac{-E}{(kT)}}$, όπου $T > 0$ είναι η θερμοκρασία και $k > 0$ είναι μια σταθερά.

Ο αλγόριθμος Simulated annealing βρίσκει ένα ολικό μέγιστο, με πιθανότητα που τείνει στο 1 αν το χρονοδιάγραμμα (schedule) μειώνει τη θερμοκρασία T αρκετά αργά. Το ακριβές όριο για την παράμετρο T και το χρονοδιάγραμμα T εξαρτάται συνήθως από το πρόβλημα [63].

Κεφάλαιο 9^{ο}

9 Ιατρικά δεδομένα που χρησιμοποιήθηκαν για τη μελέτη και διεξαγωγή αποτελεσμάτων

Η βάση δεδομένων που χρησιμοποιήσαμε είναι ένα δείγμα 700 εξεταζόμενων γυναικών από το Αττικό Νοσοκομείο και από το Νοσοκομείο Ιωαννίνων. Ως reference test (για την εξακρίβωση της ορθότητας και εγκυρότητας των αποτελεσμάτων) χρησιμοποιήθηκε η ιστολογική βιοψία. Το κυτταρολογικό επίχρισμα λήφθηκε σε LBC μορφή (Liquid Based Cytology – Κυτταρολογία Υγρής Φάσης). Σύμφωνα με τα αποτελέσματα των εξετάσεων έχουμε τυποποίηση 35 HPV τύπων (υψηλού και χαμηλού κινδύνου), ανίχνευση του mRNA των τύπων 16, 18, 31, 33, 45, ανίχνευση της πρωτεΐνης P16 που υπερεκφράζεται στον καρκίνο του τραχήλου και κυτταρομετρία ροής για ανίχνευση του mRNA των HPV τύπων υψηλού κινδύνου.

Όπως έχουμε αναφέρει και παραπάνω, ο σκοπός υλοποίησης των εναλλακτικών σεναρίων επί του Μπεϋζιανού γράφου είναι να αποτελέσουν ένα μέσο επιβεβαίωσης των ιατρικών δεδομένων που διαθέτουμε και αφετέρου μέσω των σεναρίων αυτών επιδιώκουμε να προσδιορίσουμε όσο καλύτερα γίνεται το αποτέλεσμα του test Pap όταν αυτό προκύπτει ASCUS (atypical squamous cells of undetermined significance). Αυτό που έχει την μεγαλύτερη δυσκολία στην αντιστοίχιση του με το αποτέλεσμα της βιοψίας είναι το αποτέλεσμα ASCUS. Αυτό είναι που προβληματίζει εδώ και χρόνια την ιατρική κοινότητα [3].

Επίσης, υπενθυμίζουμε ότι ένα σημαντικό στοιχείο που αφορά τη διεξαγωγή συμπερασμάτων είναι η ομαδοποίηση των αποτελεσμάτων της βιοψίας σε δύο σημαντικές κατηγορίες. Αυτές οι κατηγορίες είναι οι εξής:

1.1 CIN^{-}

Σε αυτή την κατηγορία περιλαμβάνονται τα εξής αποτελέσματα :

- *NEGATIVE*
- *CIN1*

2.1 CIN^{+}

Σε αυτή την κατηγορία περιλαμβάνονται τα εξής αποτελέσματα :

- *CIN2*

- *CA*

Η συγκεκριμένη κατηγοριοποίηση είναι πολύ χρήσιμη για τους κλινικούς ιατρούς. Με βάση αυτήν, κρίνουν εάν θα παραπέμψουν τον ασθενή σε επανάληψη του test Pap (κατηγορία CIN-) ή θα προχωρήσουν σε επέμβαση (κατηγορία CIN+).

9.1 Εργαλεία εφαρμογής αλγορίθμων εκμάθησης

Το WEKA είναι ένα εργαλείο της Μηχανικής Μάθησης (machine learning) που στόχο έχει να βοηθήσει την εφαρμογή των τεχνικών μηχανικής μάθησης σε μία ποικιλία από προβλήματα πραγματικού κόσμου. Σε αντίθεση με άλλα εργαλεία μηχανικής μάθησης, η έμφαση είναι στο να παρέχει ένα περιβάλλον εργασίας για τον ειδικό και όχι τον εμπειρογνώμονα της μηχανικής μάθησης. Υπάρχει μεγάλη ανάγκη παροχής διαδραστικών εργαλείων για το χειρισμό των δεδομένων, την οπτικοποίηση των αποτελεσμάτων, τη σύνδεση με τη βάση δεδομένων, τη διασταυρωμένη επικύρωση και σύγκριση των συνόλων των αποτελεσμάτων.
Είναι ένα πολύ δημοφιλές λογισμικό, το οποίο είναι γραμμένο σε Java, και αναπτύχθηκε στο Πανεπιστήμιο του Waikato, στη Νέα Ζηλανδία. Περιέχει μία συλλογή από εργαλεία οπτικοποίησης και αλγορίθμων για την ανάλυση δεδομένων και μοντέλων πρόβλεψης με γραφικές εφαρμογές για εύκολη πρόσβαση σε αυτή τη λειτουργία [35].
Το WEKA υποστηρίζει αρκετές τυπικές εργασίες όπως είναι η εξόρυξη δεδομένων (data mining) και πιο συγκεκριμένα η προεπεξεργασία δεδομένων, η ομαδοποίηση, η ταξινόμηση (classification), η οπισθοδρόμηση (regression) και η επιλογή χαρακτηριστικών. Όλες οι τεχνικές Weka στηρίζονται στην υπόθεση ότι τα δεδομένα είναι διαθέσιμα ως ένα ενιαίο επίπεδο αρχείο ή σχέση, όπου κάθε σημείο δεδομένων περιγράφεται από ένα σταθερό αριθμό από χαρακτηριστικά (συνήθως, αριθμητικό ή ονομαστικό χαρακτηριστικό).
Το Weka παρέχει πρόσβαση σε βάσεις δεδομένων SQL χρησιμοποιώντας Java Database Connectivity και μπορεί να επεξεργαστεί το αποτέλεσμα που επιστρέφεται από ένα ερώτημα βάσης δεδομένων. Δεν είναι σε θέση να κάνει πολυ-σχεσιακή εξόρυξη δεδομένων (multi- relational data mining), αλλά υπάρχει ξεχωριστό λογισμικό για τη μετατροπή μιας συλλογής από πίνακες της βάσης δεδομένων που συνδέονται σε ένα ενιαίο πίνακα που είναι κατάλληλος για επεξεργασία χρησιμοποιώντας το Weka. Ένας άλλος σημαντικός τομέας που σήμερα δεν καλύπτεται από τους αλγορίθμους που περιλαμβάνονται στο Weka είναι η μοντελοποίηση της ακολουθίας (sequence modeling) [36].
Αφού επιλέξουμε ποιον αλγόριθμο θέλουμε να εφαρμόσουμε στα δεδομένα, στη συνέχεια επιλέγουμε κάποια συγκεκριμένα χαρακτηριστικά ταξινόμησης (score metrics). Οι αλγόριθμοι ταξινόμησης έχουν κοινά τα παρακάτω χαρακτηριστικά:

- **initAsNaiveBayes** αν είναι επιλεγμένο ως αληθές (true by default), η αρχική δομή του δικτύου που χρησιμοποιείται για την εκκίνηση της διάσχισης του χώρου αναζήτησης είναι μια Μπευζιανή δομή δικτύου. Δηλαδή είναι, μια δομή με βέλη από τη μεταβλητή τάξη σε κάθε μία από τις μεταβλητές ιδιότητες. Αν

οριστεί ψευδής (φαλσε), μια κενή δομή του δικτύου θα πρέπει να χρησιμοποιηθεί (δηλαδή δεν υπάρχουν καθόλου βέλη).

- **markovBlanketClassifier** ορίζεται ψευδές από προεπιλογή (False by default). Αν οριστεί ως αληθές, στο τέλος της διάσχισης του χώρου αναζήτησης, μια ευρετική (heuristic) χρησιμοποιείται για να εξασφαλίσει κάθε ένα από τα χαρακτηριστικά που βρίσκονται στην «κουβέρτα Markov» (Markov blanket) του κόμβου ταξινόμησης. Αν ένας κόμβος είναι ήδη στην «κουβέρτα Markov» (δηλαδή, είναι ένας γονέας, παιδί του αδελφού του ταξινομητή) δεν συμβαίνει τίποτα, αλλιώς ένα βέλος προστίθεται. Αν οριστεί ψευδές δεν προστίθενται κανένα τέτοιο βέλος.

- **scoreType** Καθορίζει το μετρικό σκορ που θα χρησιμοποιηθεί.

- **maxNrOfParents** είναι ένα άνω όριο για τον αριθμό των γονέων κάθε κόμβου στο δίκτυο.

9.2 Παραμετροποίηση, εφαρμογή και αξιολόγηση των αλγορίθμων εκμάθησης

9.2.1 Γενετικός Αλγόριθμος

Ο πρώτος αλγόριθμος στον οποίο μελετάται η παραμετροποίηση του και αξιολογείται αντίστοιχα, είναι ο Γενετικός Αλγόριθμος. Οι διαθέσιμες παράμετροι είναι οι παρακάτω:

- **populationSize** : είναι το μέγεθος του πληθυσμού που επιλέγεται σε κάθε γενιά.
- **descendantPopulationSize**: είναι ο αριθμός των απογόνων που παράγονται σε μία γενιά (generation).
- **runs**: είναι ο αριθμός παραγόμενων γενιών.
- **seed:** είναι η τιμή αρχικοποίησης για την γεννήτρια τυχαίων αριθμών.
- **useMutation:** είναι η παράμετρος που χρησιμοποιείται για να αναφέρει εάν θα πρέπει να χρησιμοποιηθεί μετάλλαξη ή όχι. Η μετάλλαξη εφαρμόζεται από τυχαία προσθήκη ή διαγραφή ενός ενιαίου τόξου.
- **useCrossOver:** είναι η παράμετρος που χρησιμοποιείται για να αναφέρει εάν θα πρέπει να χρησιμοποιούνται διαστραυρώσεις (crossovers). Η διασταύρωση (crossover) εφαρμόζεται με τυχαία επιλογή ενός δείκτη k και επιλέγοντας τα πρώτα k bits από το ένα, και το υπόλοιπο από άλλη δομή του δικτύου στον πληθυσμό. Τουλάχιστον ένα από τα useMutation και useCrossOver θα πρέπει να οριστεί σε true.
- **useTournamentSelection:** όταν είναι ψευδές (False), οι καλύτερες επιδόσεις των δικτύων επιλέγονται από τους απογόνους για να σχηματίσουν τον

πληθυσμό της επόμενης γενιάς. Όταν η τιμή του είναι αληθής (true), χρησιμοποιείται η επιλογή τουρνουά (tournament selection). Η τουρνουά επιλογή επιλέγει τυχαία δύο άτομα από τους απογόνους και επιλέγει εκείνη που αποδίδει καλύτερα [61].

Τονίζουμε ότι πραγματοποιήθηκε **Trial and Error Method**, δηλαδή αφού εφαρμόσαμε τα δεδομένα για διάφορες τιμές των παραμέτρων των αλγορίθμων, καταλήξαμε στις πιο χαρακτηριστικές και στους γράφους με την καλύτερη απόδοση [62]. Αρχικά πραγματοποιείται η εκμάθηση του γράφου και στη συνέχεια. Η ευαισθησία και η ειδικότητα που υπολογίζονται για κάθε μία περίπτωση, αφορούν τις περιπτώσεις με CIN- και CIN+.

Μόλις επιλέξουμε το Γενετικό αλγόριθμο, και συγκεκριμένες παραμέτρους θα υπολογίσουμε το γράφο.

Στη συνέχεια γίνεται Εκμάθηση του Γράφου και βλέπουμε τα ποσοστά ορθής και λανθασμένης ταξινόμησης των αποτελεσμάτων και συγκεκριμένα 79.0503% του ποσοστού των αποτελεσμάτων είναι ταξινομημένο σωστά και 20.9497% λανθανσμένα, με βάση την αναλυτική παραμετροποίηση που φαίνεται στο παράρτημα.

Στον πίνακα, φαίνονται και τ’ αποτελέσματα μετά το re-envaluation, με βάση το test και το validation. Ταξινομούνται σωστά 113 και λανθασμένα 41, ενώ φαίνεται το αντίστοιχο ποσοστό ορθής και λανθασμένης ταξινόμησης.

```
Correctly Classified Instances         283               79.0503 %
Incorrectly Classified Instances        75               20.9497 %
Kappa statistic                          0.6749
Mean absolute error                      0.1528
Root mean squared error                  0.2735
Relative absolute error                 46.8059 %
Root relative squared error             67.7616 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.87      0.147      0.817     0.87      0.843      0.93     Negative
                 0.74      0.132      0.764     0.74      0.752      0.895    CIN1
                 0.75      0.05       0.75      0.75      0.75       0.955    CIN2/3
                 0.538     0          1         0.538     0.7        0.959    Cancer
Weighted Avg.    0.791     0.12       0.793     0.791     0.789      0.922

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 134  19   1   0 |   a = Negative
  26  97   8   0 |   b = CIN1
   4  11  45   0 |   c = CIN2/3
   0   0   6   7 |   d = Cancer
```

Εικόνα 26 Εκμάθηση Γράφου και ποσοστά ορθής και λανθασμένης ταξινόμησης των αποτελεσμάτων και συγκεκριμένα 79.0503% του ποσοστού των αποτελεσμάτων είναι ταξινομημένο σωστά και 20.9497% λανθανσμένα.

```
Correctly Classified Instances         113               73.3766 %
Incorrectly Classified Instances        41               26.6234 %
Kappa statistic                          0.5961
Mean absolute error                      0.1723
Root mean squared error                  0.3062
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.731     0.069      0.891     0.731     0.803      0.93     Negative
                 0.8       0.288      0.571     0.8       0.667      0.782    CIN1
                 0.667     0.041      0.815     0.667     0.733      0.945    CIN2/3
                 0.5       0          1         0.5       0.667      0.871    Cancer
Weighted Avg.    0.734     0.133      0.774     0.734     0.74       0.884

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 49 18  0  0 |  a = Negative
  6 40  4  0 |  b = CIN1
  0 11 22  0 |  c = CIN2/3
  0  1  1  2 |  d = Cancer
```

Εικόνα 27 Re-evaluation (επαναξιολόγηση) αποτελεσμάτων, σωστή ταξινόμηση 113 αποτελεσμάτων και λανθασμένη ταξινόμηση 41 αποτελεσμάτων, νέο confusion matrix.

Όπως έχουμε αναφέρει παραπάνω, τα αποτελέσματα των εξετάσεων τα αποτελέσματα του τεστ για CIN- και CIN+, μπορεί να συμπίπτουν με την αληθινή κατάσταση του ατόμου αλλά μπορεί και όχι. Τότε μπορούν να υπάρξουν τα παρακάτω αποτελέσματα:

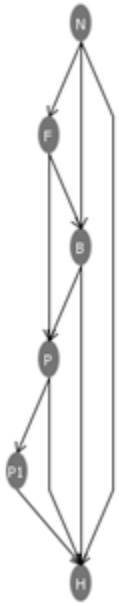
- Αληθώς θετικό (true positive- TP): Ο ασθενής αναγνωρίστηκε σωστά ως ασθενής.
- Ψευδώς θετικό (false positive- FP): Το υγιές άτομο αναγνωρίστηκε λανθασμένα ως ασθενής.
- Αληθώς αρνητικό (true negative- TN):Το υγιές άτομο αναγνωρίστηκε σωστά ως υγιές.
- Ψευδώς αρνητικό (false negative- FN): Ο ασθενής αναγνωρίστηκε λανθασμένα ως υγιής.

Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,675676
- Ειδικότητα= $\frac{d}{b+d}$= 0,965812

Από τα παραπάνω 113 ταξινομήθηκαν ως TN, τα 4 ως FP, τα 12 ως FN και τα 25 ως TP.

Ο γράφος που προκύπτει είναι ο παρακάτω:



Εικόνα 28 Γράφος με καλύτερη απόδοση με βάση την πρώτη παραμετροποίηση του Γενετικού αλγορίθμου

Γενικά, παρατηρήθηκε ότι ο αριθμός των γενιών, δηλαδή η παράμετρος runs, παίζει σημαντικό ρόλο στη διαμόρφωση των αποτελεσμάτων, καθώς όσο πιο λίγες γενιές έχουμε, τόσο περισσότερα αποτελέσματα ταξινομούνται λανθασμένα. Η παραμετροποίηση που ακολουθεί, φαίνεται πάλι στο παράρτημα και έχουμε διπλασσιάσει το descendatPopulationSize, δηλαδή το πόσοι απόγονοι παράγονται. Έχουμε λοιπόν τα παρακάτω αποτελέσματα.

Αριστερά του πίνακα φαίνεται η εκμάθηση του γράφου και η ταξινόμηση των αποτελεσμάτων, και δεξιά, φαίνονται τα αποτελέσματα μετά το re-evaluation.

```
Correctly Classified Instances         282               78.7709 %
Incorrectly Classified Instances        76               21.2291 %
Kappa statistic                          0.6719
Mean absolute error                      0.1576
Root mean squared error                  0.2736
Relative absolute error                 48.2942 %
Root relative squared error             67.7956 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.87      0.147      0.817     0.87      0.843      0.933    Negative
                 0.718     0.123      0.77      0.718     0.743      0.892    CIN1
                 0.783     0.06       0.723     0.783     0.752      0.954    CIN2/3
                 0.538     0          1         0.538     0.7        0.963    Cancer
Weighted Avg.    0.788     0.119      0.791     0.788     0.786      0.923

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 134  19   1   0 |   a = Negative
  26  94  11   0 |   b = CIN1
   4   9  47   0 |   c = CIN2/3
   0   0   6   7 |   d = Cancer
```

Εικόνα 29 Εκμάθηση αλγορίθμου με αύξηση της παραμέτρου runs, confusion matrix, αποτελέσματα ταξινόμησης αποτελεσμάτων και ποσοστά σωστής και λανθασμένης ταξινόμησης

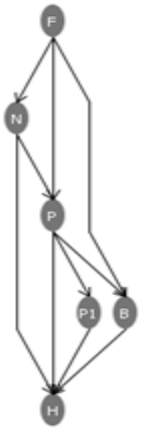
```
Correctly Classified Instances         112               72.7273 %
Incorrectly Classified Instances        42               27.2727 %
Kappa statistic                          0.5867
Mean absolute error                      0.1808
Root mean squared error                  0.3141
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.731     0.069      0.891     0.731     0.803      0.927    Negative
                 0.78      0.288      0.565     0.78      0.655      0.777    CIN1
                 0.667     0.05       0.786     0.667     0.721      0.92     CIN2/3
                 0.5       0          1         0.5       0.667      0.789    Cancer
Weighted Avg.    0.727     0.134      0.765     0.727     0.734      0.873

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 49 18  0  0 |  a = Negative
  6 39  5  0 |  b = CIN1
  0 11 22  0 |  c = CIN2/3
  0  1  1  2 |  d = Cancer
```

Εικόνα 30 Re-evaluation αποτελεσμάτων, νέο confusion matrix και ποσοστά σωστής και λανθασμένης ταξινόμησης, 112 αποτελέσματα ταξινομήθηκαν ως TN, τα 5 ως FP, τα 12 ως FN και τα 25 ως TP

Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,675676
- Ειδικότητα= $\frac{d}{b+d}$= 0,957265

Ο γράφος που προκύπτει είναι ο εξής:



Εικόνα 31 Γράφος με την καλύτερη απόδοση για την παραπάνω παραμετροποίηση του αλγορίθμου

Επίσης, εάν δεν υπάρχουν διαυστραυρώσεις παρατηρείται ότι δε γίνεται σωστή ταξινόμηση, οπότε εάν η συγκεκριμένη παράμετρος useCrossOver γίνει True, προκύπτουν διαφορετικά αποτελέσματα. Η αναλυτική παραμετροποίηση φαίνεται στο παράρτημα, όπου μπορείτε να ανατρέξετε.

```
Correctly Classified Instances         280               78.2123 %
Incorrectly Classified Instances        78               21.7877 %
Kappa statistic                          0.6618
Mean absolute error                      0.1631
Root mean squared error                  0.2759
Relative absolute error                 49.9739 %
Root relative squared error             68.3439 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.864     0.157      0.806     0.864     0.834      0.933    Negative
                 0.725     0.137      0.754     0.725     0.739      0.888    CIN1
                 0.75      0.05       0.75      0.75      0.75       0.95     CIN2/3
                 0.538     0          1         0.538     0.7        0.965    Cancer
Weighted Avg.    0.782     0.126      0.785     0.782     0.78       0.921

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 133  19   2   0 |   a = Negative
  28  95   8   0 |   b = CIN1
   4  11  45   0 |   c = CIN2/3
   0   1   5   7 |   d = Cancer
```

Εικόνα 32 Εκμάθηση γράφου, Confusion Matrix για την παραπάνω παραμετροποίηση και ποσοστά σωστών και λανθασμένων ταξινομημένων αποτελεσμάτων

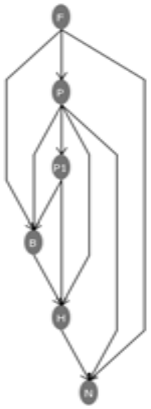
```
Correctly Classified Instances         109               70.7792 %
Incorrectly Classified Instances        45               29.2208 %
Kappa statistic                          0.5591
Mean absolute error                      0.1906
Root mean squared error                  0.3172
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.731     0.092      0.86      0.731     0.79       0.914    Negative
                 0.68      0.26       0.557     0.68      0.613      0.777    CIN1
                 0.727     0.083      0.706     0.727     0.716      0.941    CIN2/3
                 0.5       0          1         0.5       0.667      0.891    Cancer
Weighted Avg.    0.708     0.142      0.732     0.708     0.714      0.874

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
  49  18   0   0 |   a = Negative
   7  34   9   0 |   b = CIN1
   1   8  24   0 |   c = CIN2/3
   0   1   1   2 |   d = Cancer
```

Εικόνα 33 Re-evaluation αποτελεσμέτων, και νέα confusion matrix, ταξινομούνται σωστά 109 και λανθασμένα 45. Τα 108 ταξινομούνται ως TN, τα 9 ως FP, τα 10 ως FN και τα 27 ως TP

Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,72973
- Ειδικότητα= $\frac{d}{b+d}$= 0,923077

Ο γράφος που προκύπτει είναι ο παρακάτω:



Εικόνα 34 Γράφος με την καλύτερη απόδοση για την παραπάνω παραμετροποίηση του αλγορίθμου

Στη συνέχεια αυξάνουμε τον αριθμό των απογόνων που παράγονται σε μια γενιά. Η αναλυτική παραμετροποίηση φαίνεται στο παράρτημα.
Αριστερά στον παρακάτω πίνακα φαίνονται τα αποτελέσματα από την εκμάθηση του γράφου και το confusion matrix, και δεξιά τα αποτελέσματα μετά το re-evaluation.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         283               79.0503 %
Incorrectly Classified Instances        75               20.9497 %
Kappa statistic                          0.6749
Mean absolute error                      0.1528
Root mean squared error                  0.2735
Relative absolute error                 46.8059 %
Root relative squared error             67.7616 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.87      0.147      0.817     0.87      0.843      0.93     Negative
                 0.74      0.132      0.764     0.74      0.752      0.895    CIN1
                 0.75      0.05       0.75      0.75      0.75       0.955    CIN2/3
                 0.538     0          1         0.538     0.7        0.959    Cancer
Weighted Avg.    0.791     0.12       0.793     0.791     0.789      0.922

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 134  19   1   0 |   a = Negative
  26  97   8   0 |   b = CIN1
   4  11  45   0 |   c = CIN2/3
   0   0   6   7 |   d = Cancer
```

Εικόνα 35 Εκμάθηση αποτελεσμάτων και ποσοστά λανθασμένων και μη ταξινομημένων αποτελεσμάτων με βάση την παραπάνω παραμετροποίηση

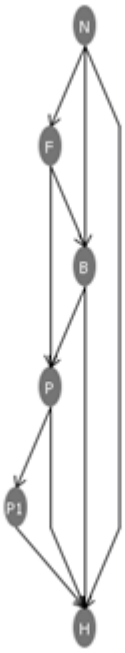
```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         113               73.3766 %
Incorrectly Classified Instances        41               26.6234 %
Kappa statistic                          0.5961
Mean absolute error                      0.1723
Root mean squared error                  0.3062
Relative absolute error                 52.3601 %
Root relative squared error             75.238  %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.731     0.069      0.891     0.731     0.803      0.93     Negative
                 0.8       0.288      0.571     0.8       0.667      0.782    CIN1
                 0.667     0.041      0.815     0.667     0.733      0.945    CIN2/3
                 0.5       0          1         0.5       0.667      0.871    Cancer
Weighted Avg.    0.734     0.133      0.774     0.734     0.74       0.884

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 49 18  0  0 |   a = Negative
  6 40  4  0 |   b = CIN1
  0 11 22  0 |   c = CIN2/3
  0  1  1  2 |   d = Cancer
```

Εικόνα 36 Re-evaluation αποτελεσμάτων και νέα confusion matrix

Μετά το re-evaluation, όπως φαίνεται και παραπάνω, ταξινομούνται σωστά τα 113 και λανθασμένα τα 41.
Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,72973
- Ειδικότητα= $\frac{d}{b+d}$= 0,923077

Ο γράφος που προκύπτει με την καλύτερη απόδοση, φαίνεται στο παρακάτω σχήμα.



Εικόνα 37 Γράφος με την καλύτερη απόδοση με βάση την παραπάνω παραμετροποίηση

Στη συνέχεια αυξάνουμε τις παραμέτρους runs και seed. Η αναλυτική παραμετροποίηση φαίνεται στο παράρτημα.
Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα μετά την εκμάθηση του γράφου και μετά το re-evaluation.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         279               77.933  %
Incorrectly Classified Instances        79               22.067  %
Kappa statistic                          0.6588
Mean absolute error                      0.1613
Root mean squared error                  0.2769
Relative absolute error                 49.4379 %
Root relative squared error             68.5983 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.87      0.157      0.807     0.87      0.838      0.932    Negative
                 0.702     0.123      0.767     0.702     0.733      0.886    CIN1
                 0.767     0.064      0.708     0.767     0.736      0.948    CIN2/3
                 0.538     0          1         0.538     0.7        0.964    Cancer
Weighted Avg.    0.779     0.123      0.783     0.779     0.777      0.919

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 134  18   2   0 |   a = Negative
  28  92  11   0 |   b = CIN1
   4  10  46   0 |   c = CIN2/3
   0   0   6   7 |   d = Cancer
```

Εικόνα 38 Εκμάθηση αλγορίθμου, confusion matrix με βάση την παραπάνω παραμετροποίηση, δηλαδή με αύξηση των παραμέτρων runs και seed

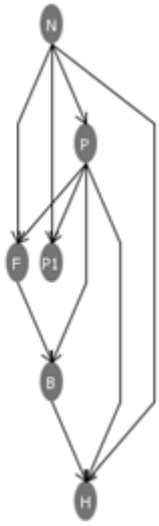
```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         112               72.7273 %
Incorrectly Classified Instances        42               27.2727 %
Kappa statistic                          0.5862
Mean absolute error                      0.1813
Root mean squared error                  0.315
Relative absolute error                 55.0959 %
Root relative squared error             77.4058 %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.731     0.08       0.875     0.731     0.797      0.928    Negative
                 0.78      0.279      0.574     0.78      0.661      0.775    CIN1
                 0.667     0.05       0.786     0.667     0.721      0.916    CIN2/3
                 0.5       0          1         0.5       0.667      0.808    Cancer
Weighted Avg.    0.727     0.136      0.761     0.727     0.733      0.873

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 49 18  0  0 |  a = Negative
  6 39  5  0 |  b = CIN1
  1 10 22  0 |  c = CIN2/3
  0  1  1  2 |  d = Cancer
```

Εικόνα 39 Re-evaluation, νέο confusion matrix, αποτελέσματα σωστής και λανθασμένης ταξινόμησης

Όπως φαίνεται στα δεξιά του πίνακα, μετά το re-evaluation ταξινομούνται 112 σωστά και 42 λανθασμένα. Τα 112 ταξινομήθηκαν ως TN, τα 5 ως FN, τα 12 ως FN και τα 25 ως TP.
Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,675676
- Ειδικότητα= $\frac{d}{b+d}$= 0,957265

Παρακάτω φαίνεται ο γράφος με την καλύτερη απόδοση για τη συγκεκριμένη παραμετροποίηση.



Εικόνα 40 Γράφος με την καλύτερη απόδοση με βάση την παραπάνω παραμετροποίηση

Παραπάνω φαίνονται μερικές από τις καλύτερες ταξινομήσεις αποτελεσμάτων. Γενικότερα παρατηρήθηκε ότι η αυξομείωση του αριθμού των ατόμων του πληθυσμού δεν επηρέασε δραματικά την ταξινόμηση των αποτελεσμάτων. Αντίθετα, το πόσα άτομα παράγονται ήταν σημαντικό, και μάλιστα όσο περισσότερα παράγονταν τόσο καλύτερα γινόταν η ταξινόμηση. Επίσης, όταν υπήρχαν διασταυρώσεις τα αποτελέσματα ταξινομούνταν καλύτερα.

Ο καλύτερος γράφος προέκυψε στην τρίτη παραμετροποίηση, όπου δεν υπάρχουν διασταυρώσεις. Δηλαδή, χωρίς διασταθρώσεις προέκυψε ο καλύτερος γράφος.

9.2.2 Hill Climbing

Ο δεύτερος αλγόριθμος στον οποίο μελετάται η παραμετροποίηση του και αξιολογείται αντίστοιχα, είναι ο Hill Climbing. Οι διαθέσιμες παράμετροι έχουν αναλυθεί παραπάνω, και επιπλέον υπάρχει η παρακάτω:

- useArcReversal: εάν είναι αληθής (true), τα βήματα καθορίζονται από το επόμενο βήμα.

Η πρώτη παραμετροποίηση του αλγορίθμου, φαίνεται στο παράρτημα. Η παράμετρος useArcReversal επιλέγεται ως False, και το maxNrOfParets ισούται με 1.
Παρακάτω φαίνονται τα αποτελέσματα ταξινόμησης των αποτελεσμάτων μετά την εκμάθηση του γράφου και μετά από το re-evaluation.

```
Correctly Classified Instances         275               76.8156 %
Incorrectly Classified Instances        83               23.1844 %
Kappa statistic                          0.6448
Mean absolute error                      0.1538
Root mean squared error                  0.2942
Relative absolute error                 47.117  %
Root relative squared error             72.894  %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.89      0.176      0.792     0.89      0.838      0.925    Negative
                 0.611     0.088      0.8       0.611     0.693      0.863    CIN1
                 0.85      0.091      0.654     0.85      0.739      0.932    CIN2/3
                 0.538     0          1         0.538     0.7        0.962    Cancer
Weighted Avg.    0.768     0.123      0.779     0.768     0.763      0.905

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 137  15   2   0 |   a = Negative
  32  80  19   0 |   b = CIN1
   4   5  51   0 |   c = CIN2/3
   0   0   6   7 |   d = Cancer
```

Εικόνα 41 Εκμάθηση Γράφου με βάση την παραπάνω παραμετροποίηση, με maxNrOfParets= 1.

```
Correctly Classified Instances         118               76.6234 %
Incorrectly Classified Instances        36               23.3766 %
Kappa statistic                          0.648
Mean absolute error                      0.1551
Root mean squared error                  0.3098
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.896     0.092      0.882     0.896     0.889      0.935    Negative
                 0.52      0.106      0.703     0.52      0.598      0.769    CIN1
                 0.909     0.14       0.638     0.909     0.75       0.935    CIN2/3
                 0.5       0          1         0.5       0.667      0.878    Cancer
Weighted Avg.    0.766     0.104      0.775     0.766     0.759      0.88

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 60  7  0  0 |  a = Negative
  8 26 16  0 |  b = CIN1
  0  3 30  0 |  c = CIN2/3
  0  1  1  2 |  d = Cancer
```

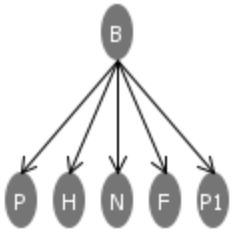
Εικόνα 42 Re-evaluation γράφου και νέο confusion matrix, καθώς και νέα ποσοστά ταξινόμησης των αποτελεσμάτων

Όπως φαίνεται στα δεξιά του παραπάνω πίνακα, μετά το re-evaluation ταξινομεί 118 αποτελέσματα σωστά και 36 λανθασμένα. Τα 101 ταξινομήθηκαν ως TN, τα 16 ως FP, τα 4 ως FN και τα 33 ως TP.

Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,891892
- Ειδικότητα= $\frac{d}{b+d}$= 0,863248

Ο γράφος με την καλύτερη απόδοση που προκύπτει φαίνεται παρακάτω και είναι Naive:



Εικόνα 43 Γράφος με την καλύτερη απόδοση με βάση την παραπάνω παραμετροποίηση

Παρατηρήσαμε ότι αυξάνοντας τον αριθμό των γονέων η ταξινόμηση γίνεται λιγότερο σωστά, οπότε αλλάξαμε και την παράμετρο markovBlanketClassifier σε αληθή. Έτσι, τα αποτέλεσματα είναι καλύτερα.

Αλλάζοντας λοιπόν την παράμετρο αυτή προκύπτουν τα παρακάτω αποτεελέσματα. Η αναλυτική παραμετροποίηση φαίνεται στο παράρτημα. Στη συγκεκριμένη παραμετροποίηση, γίνεται αύξηση της παραμέτρου maxNrOfParents και αλλαγή των παραμέτρων initAsNaiveBayes, markovBlanketClassifier και useArcReversal.

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα μετά την εκμάθηση του γράφου και μετά το re-evaluation.

```
Correctly Classified Instances         278               77.6536 %
Incorrectly Classified Instances        80               22.3464 %
Kappa statistic                          0.6532
Mean absolute error                      0.1537
Root mean squared error                  0.2769
Relative absolute error                 47.1082 %
Root relative squared error             68.6017 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.864     0.152      0.811     0.864     0.836      0.932    Negative
                 0.718     0.145      0.74      0.718     0.729      0.883    CIN1
                 0.733     0.054      0.733     0.733     0.733      0.949    CIN2/3
                 0.538     0          1         0.538     0.7        0.957    Cancer
Weighted Avg.    0.777     0.128      0.779     0.777     0.775      0.918

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 133  19   2   0 |   a = Negative
  27  94  10   0 |   b = CIN1
   4  12  44   0 |   c = CIN2/3
   0   2   4   7 |   d = Cancer
```

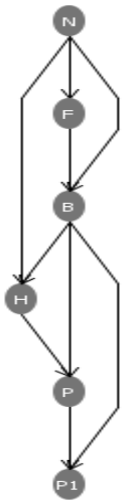
Εικόνα 44 Εκμάθηση Γράφου, αύξηση της παραμέτρου maxNrOfParents και αλλαγή παραμέτρων initAsNaiveBayes, markovBlanketClassifier και useArcReversal

```
Correctly Classified Instances         108               70.1299 %
Incorrectly Classified Instances        46               29.8701 %
Kappa statistic                          0.553
Mean absolute error                      0.1736
Root mean squared error                  0.3135
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.731     0.069      0.891     0.731     0.803      0.932    Negative
                 0.66      0.25       0.559     0.66      0.606      0.795    CIN1
                 0.727     0.107      0.649     0.727     0.686      0.925    CIN2/3
                 0.5       0.007      0.667     0.5       0.571      0.878    Cancer
Weighted Avg.    0.701     0.134      0.726     0.701     0.708      0.884

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 49 18  0  0 |  a = Negative
  6 33 11  0 |  b = CIN1
  0  8 24  1 |  c = CIN2/3
  0  0  2  2 |  d = Cancer
```

Εικόνα 45 Re-evaluation γράφου με βάση την παραπάνω παραμετροποίηση

Όπως φαίνεται στον παραπάνω πίνακα, μετά το re-evaluation τα 100 ταξινομούνται σωστά και τα 54 λανθασμένα. Από αυτά τα 102 ταξινομούνται ως TN, τα 15 ως FP, τα 13 ως FN και τα 24 ως TP.

Ο γράφος με την καλύτερη απόδοση είναι ο παρακάτω:



Εικόνα 46 Γράφος με την καλύτερη απόδοση με βάση την παραπάνω παραμετροποίηση

Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,648649
- Ειδικότητα= $\frac{d}{b+d}$= 0,871795

Στη συνέχεια, κάνοντας την παράμετρο useArcReversal αληθή, προκύπτουν τα παρακάτω αποτελέσματα, μετά την εκμάθηση του γράφου και μετά το re-evaluation.

```
Correctly Classified Instances         275               76.8156 %
Incorrectly Classified Instances        83               23.1844 %
Kappa statistic                          0.6448
Mean absolute error                      0.1539
Root mean squared error                  0.2942
Relative absolute error                 47.1534 %
Root relative squared error             72.887  %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.89      0.176      0.792     0.89      0.838      0.925    Negative
                 0.611     0.088      0.8       0.611     0.693      0.863    CIN1
                 0.85      0.091      0.654     0.85      0.739      0.932    CIN2/3
                 0.538     0          1         0.538     0.7        0.962    Cancer
Weighted Avg.    0.768     0.123      0.779     0.768     0.763      0.905

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 137  15   2   0 |   a = Negative
  32  80  19   0 |   b = CIN1
   4   5  51   0 |   c = CIN2/3
   0   0   6   7 |   d = Cancer
```

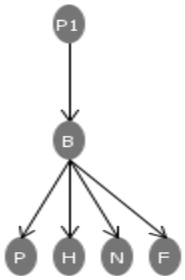
Εικόνα 47 Εκμάθηση Γράφου, για useArcReversal= True, Confusion Matrix και αποτελέσματα σωστής και λανθασμένης ταξινόμησης των αποτελεσμάτων

```
Correctly Classified Instances         118               76.6234 %
Incorrectly Classified Instances        36               23.3766 %
Kappa statistic                          0.648
Mean absolute error                      0.1553
Root mean squared error                  0.3098
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.896     0.092      0.882     0.896     0.889      0.935    Negative
                 0.52      0.106      0.703     0.52      0.598      0.769    CIN1
                 0.909     0.14       0.638     0.909     0.75       0.936    CIN2/3
                 0.5       0          1         0.5       0.667      0.878    Cancer
Weighted Avg.    0.766     0.104      0.775     0.766     0.759      0.88

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 60  7  0  0 |  a = Negative
  8 26 16  0 |  b = CIN1
  0  3 30  0 |  c = CIN2/3
  0  1  1  2 |  d = Cancer
```

Εικόνα 48 Re-evaluation γράφου, νέο Confusion Matrix και νέα αποτελέσματα σωστής και λανθασμένης ταξινόμησης των αποτελεσμάτων

Μετά το re-evaluation ταξινομούνται 118 σωστά και 36 λανθασμένα. Από αυτά τα 101 ταξινομούνται ως TN, τα 16 ως FP, τα 4 ως FN και τα 33 ως TP.

Ο γράφος με την καλύτερη απόδοση είναι ο παρακάτω και είναι Naive:



Εικόνα 49 Γράφος με την καλύτερη απόδοση για τη συγκεκριμένη παραμετροποίηση

Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,891892
- Ειδικότητα= $\frac{d}{b+d}$= 0,863248

Γενικότερα με την παράμετρο useArcReversal αληθή και με αύξηση του αριθμού των γονέων τα αποτελέσματα ταξινομούνται καλύτερα. Η παραμετροποίηση φαίνεται αναλυτικά στο παράρτημα.

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα μετά την εκμάθηση του γράφου και μετά το re-evaluation.

```
Correctly Classified Instances         276               77.095 %
Incorrectly Classified Instances        82               22.905 %
Kappa statistic                          0.6417
Mean absolute error                      0.1561
Root mean squared error                  0.2787
Relative absolute error                 47.8223 %
Root relative squared error             69.0414 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.864     0.152      0.811     0.864     0.836      0.93     Negative
                 0.74      0.172      0.713     0.74      0.727      0.875    CIN1
                 0.65      0.04       0.765     0.65      0.703      0.943    CIN2/3
                 0.538     0          1         0.538     0.7        0.956    Cancer
Weighted Avg.    0.771     0.135      0.774     0.771     0.769      0.913

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 133  20   1   0 |   a = Negative
  27  97   7   0 |   b = CIN1
   4  17  39   0 |   c = CIN2/3
   0   2   4   7 |   d = Cancer
```

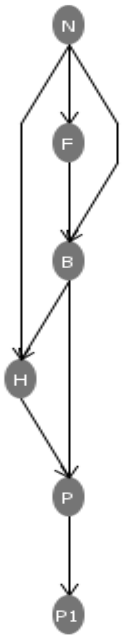
Εικόνα 50 Εκμάθηση Γράφου, αύξηση του maxNrOfParents

```
Correctly Classified Instances         111               72.0779 %
Incorrectly Classified Instances        43               27.9221 %
Kappa statistic                          0.5773
Mean absolute error                      0.1758
Root mean squared error                  0.3127
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.731     0.069      0.891     0.731     0.803      0.932    Negative
                 0.76      0.288      0.559     0.76      0.644      0.766    CIN1
                 0.667     0.058      0.759     0.667     0.71       0.928    CIN2/3
                 0.5       0          1         0.5       0.667      0.839    Cancer
Weighted Avg.    0.721     0.136      0.758     0.721     0.728      0.875

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 49 18  0  0 |  a = Negative
  6 38  6  0 |  b = CIN1
  0 11 22  0 |  c = CIN2/3
  0  1  1  2 |  d = Cancer
```

Εικόνα 51 Re-evaluation με βάση τη συγκεκριμένη παραμετροποίηση, νέο confunion matrix

Μετά το re-evaluation 111 αποτελέσματα ταξινομούνται σωστά και 43 λανθασμένα. Από αυτά τα 111 ταξινομούνται ως TN, τα 6 ως FP, τα 12 ως FN και τα 25 ως TP.

Ο γράφος με την καλύτερη απόδοση φαίνεται παρακάτω:



Εικόνα 52 Γράφος με την καλύτερη απόδοση

Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,675676
- Ειδικότητα= $\frac{d}{b+d}$= 0,948718

Οι δύο Naive γράφοι που προέκυψαν είναι αυτοί με την καλύτερη ευαισθησία και ειδικότητα.

9.2.3 Simulated annealing

Ο τρίτος αλγόριθμος στον οποίο μελετάται η παραμετροποίηση του και αξιολογείται αντίστοιχα, είναι ο Simulated annealing. Οι διαθέσιμες παράμετροι έχουν αναλυθεί παραπάνω, και επιπλέον υπάρχει η παρακάτω:

- TStart: Δείχνει από ποια τιμή θα ξεκινήσει ο αλγόριθμος Simulated Search.
- delta: καθορίζει την πιθανότητα αποδοχής των βημάτων προς τη λάθος κατεύθυνση στο χώρο αναζήτησης) η ποιά μειώνεται σε κάθε επανάληψη.

Η παραμετροποίηση του αλγορίθμου φαίνεται στο παράρτημα στο οποίο μπορείτε να ανατρέξετε.

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα μετά την εκμάθηση του γράφου και μετά το re-evaluation.

```
Correctly Classified Instances         282               78.7709 %
Incorrectly Classified Instances        76               21.2291 %
Kappa statistic                          0.6693
Mean absolute error                      0.1514
Root mean squared error                  0.2737
Relative absolute error                 46.4032 %
Root relative squared error             67.8071 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.864     0.142      0.821     0.864      0.842      0.935    Negative
                 0.756     0.154      0.739     0.756      0.747      0.893    CIN1
                 0.717     0.04       0.782     0.717      0.748      0.956    CIN2/3
                 0.538     0          1         0.538      0.7        0.969    Cancer
Weighted Avg.    0.788     0.124      0.791     0.788      0.786      0.924

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 133  20   1   0 |   a = Negative
  25  99   7   0 |   b = CIN1
   4  13  43   0 |   c = CIN2/3
   0   2   4   7 |   d = Cancer
```

Εικόνα 53 Εκμάθηση Γράφου με βάση τη συγκεκριμένη παραμετροποίηση, confusion matrix και αποτελέσματα σωστής και λανθασμένης παραμετροποίησης αποτελεσμάτων

```
Correctly Classified Instances         110               71.4286 %
Incorrectly Classified Instances        44               28.5714 %
Kappa statistic                          0.5651
Mean absolute error                      0.1741
Root mean squared error                  0.3122
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.731     0.115      0.831     0.731      0.778      0.911    Negative
                 0.74      0.269      0.569     0.74       0.643      0.783    CIN1
                 0.667     0.05       0.786     0.667      0.721      0.937    CIN2/3
                 0.5       0          1         0.5        0.667      0.859    Cancer
Weighted Avg.    0.714     0.148      0.74      0.714      0.719      0.874

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 49 18  0  0 |  a = Negative
  8 37  5  0 |  b = CIN1
  1 10 22  0 |  c = CIN2/3
  1  0  1  2 |  d = Cancer
```

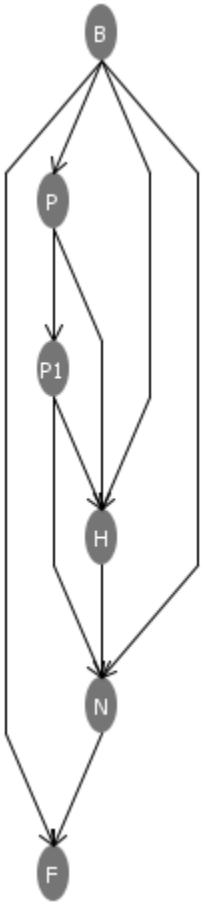
Εικόνα 54 Re-evaluation με βάση τη συγκεκριμένη παραμετροποίηση, νέο confusion matrix και νέα αποτελέσματα σωστής και λανθασμένης παραμετροποίησης αποτελεσμάτων

Μετά το re-evaluation, τα 110 αποτελέσματα ταξινομήθηκαν σωστά και τα 44 λανθασμένα. Από αυτά τα 112 ταξινομήθηκαν ως TN, τα 5 ως FP, τα 12 ως FN και τα 25 ως TP.
Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,675676

- Ειδικότητα= $\frac{d}{b+d}$= 0,957265

Ο γράφος με την καλύτερη απόδοση φαίνεται στο παρακάτω σχήμα.



Εικόνα 55 Γράφος με την καλύτερη απόδοση με βάση την παραμετροποίηση του αλγορίθμου

Καθώς αυξάνουμε την τιμή της παραμέτρου TStart, παρατηρούμε ότι δεν αλλάζει και πολύ το classification. Προς το καλύτερο γίνεται η ταξινόμηση των αποτελεσμάτων όταν αυξήσουμε την παράμετρο delta.
Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα μετά την εκμάθηση του γράφου και μετά το re-evaluation.

```
Correctly Classified Instances         265               74.0223 %
Incorrectly Classified Instances        93               25.9777 %
Kappa statistic                          0.5913
Mean absolute error                      0.1754
Root mean squared error                  0.2927
Relative absolute error                 53.7421 %
Root relative squared error             72.5069 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.922     0.27       0.721     0.922     0.809      0.907    Negative
                 0.557     0.115      0.737     0.557     0.635      0.829    CIN1
                 0.717     0.04       0.782     0.717     0.748      0.955    CIN2/3
                 0.538     0          1         0.538     0.7        0.962    Cancer
Weighted Avg.    0.74      0.165      0.747     0.74      0.731      0.888

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 142  11   1   0 |   a = Negative
  51  73   7   0 |   b = CIN1
   4  13  43   0 |   c = CIN2/3
   0   2   4   7 |   d = Cancer
```

Εικόνα 56 Εκμάθηση Γράφου, με αύξηση της τιμή της παραμέτρου TStart, confusion matrix και αποτελέσματα σωστής και λανθασμένης παραμετροποίησης αποτελεσμάτων

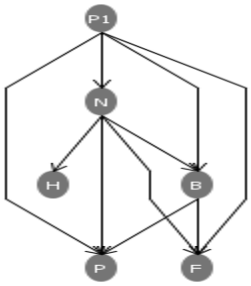
```
Correctly Classified Instances         120               77.9221 %
Incorrectly Classified Instances        34               22.0779 %
Kappa statistic                          0.6614
Mean absolute error                      0.1746
Root mean squared error                  0.3038
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.896     0.126      0.845     0.896     0.87       0.922    Negative
                 0.66      0.144      0.688     0.66      0.673      0.809    CIN1
                 0.758     0.066      0.758     0.758     0.758      0.945    CIN2/3
                 0.5       0          1         0.5       0.667      0.889    Cancer
Weighted Avg.    0.779     0.116      0.779     0.779     0.777      0.89

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 60  7  0  0 |  a = Negative
 10 33  7  0 |  b = CIN1
  0  8 25  0 |  c = CIN2/3
  1  0  1  2 |  d = Cancer
```

Εικόνα 57 Re-evaluation, νέο confusion matrix και νέα αποτελέσματα σωστής και λανθασμένης παραμετροποίησης αποτελεσμάτων

Μετά το re-evaluation, σωστά ταξινομούνται 120 αποτελέσματα ενώ λανθασμένα τα 34. Από αυτά τα 110 ταξινομούνται ως TN, τα 7 ως FP, τα 9 ως FN και τα 28 ως TP. Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,756757
- Ειδικότητα= $\frac{d}{b+d}$= 0,940171

Ο γράφος με την καλύτερη απόδοση είναι ο παρακάτω:



Εικόνα 58 Γράφος με την καλύτερη απόδοση για τη συγκεκριμένη παραμετροποίηση

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα μετά την εκμάθηση του γράφου και μετά το re-evaluation. Αυξήθηκε η παράμετρος TStart, η παραμετροποίηση φαίνεται στο παράρτημα.

```
Correctly Classified Instances         279               77.933  %
Incorrectly Classified Instances        79               22.067  %
Kappa statistic                          0.6549
Mean absolute error                      0.1675
Root mean squared error                  0.284
Relative absolute error                 51.3129 %
Root relative squared error             70.3516 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.857    0.172      0.79      0.857     0.822      0.917    Negative
                 0.756    0.145      0.75      0.756     0.753      0.888    CIN1
                 0.683    0.037      0.788     0.683     0.732      0.913    CIN2/3
                 0.538    0          1         0.538     0.7        0.969    Cancer
Weighted Avg.    0.779    0.133      0.783     0.779     0.777      0.908

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 132  20   2   0 |   a = Negative
  28  99   4   0 |   b = CIN1
   7  12  41   0 |   c = CIN2/3
   0   1   5   7 |   d = Cancer
```

Εικόνα 59 Εκμάθηση γράφου, confusion matrix και αποτελέσματα σωστής και λανθασμένης παραμετροποίησης αποτελεσμάτων

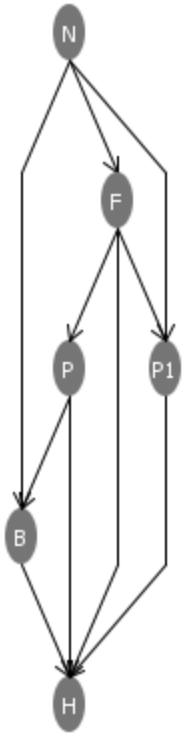
```
Correctly Classified Instances         112               72.7273 %
Incorrectly Classified Instances        42               27.2727 %
Kappa statistic                          0.5858
Mean absolute error                      0.1748
Root mean squared error                  0.3083
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.746    0.08       0.877     0.746     0.806      0.929    Negative
                 0.76     0.279      0.567     0.76      0.65       0.786    CIN1
                 0.667    0.05       0.786     0.667     0.721      0.934    CIN2/3
                 0.5      0          1         0.5       0.667      0.904    Cancer
Weighted Avg.    0.727    0.136      0.76      0.727     0.734      0.883

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 50 17  0  0 |  a = Negative
  7 38  5  0 |  b = CIN1
  0 11 22  0 |  c = CIN2/3
  0  1  1  2 |  d = Cancer
```

Εικόνα 60 Re-evaluation, νέο confusion matrix και νέα αποτελέσματα σωστής και λανθασμένης παραμετροποίησης αποτελεσμάτων

Μετά το re-evaluation τα 112 αποτελέσματα ταξινομούνται σωστά και τα 42 λανθασμένα. Από αυτά τα 112 ταξινομούνται ως TN, τα 5 ως FP, τα 12 ως FN και τα 25 ως TP.
Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,675676
- Ειδικότητα= $\frac{d}{b+d}$= 0,957265

Ο γράφος με την καλύτερη απόδοση φαίνεται στο παρακάτω σχήμα.



Εικόνα 61 Γράφος με την καλύτερη απόδοση για τη συγκεκριμένη παραμετροποίηση

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα μετά την εκμάθηση του γράφου και μετά το re-evaluation. Αυξήθηκε η παράμετρος TStart, η παραμετροποίηση φαίνεται στο παράρτημα.

```
Correctly Classified Instances         259               72.3464 %
Incorrectly Classified Instances        99               27.6536 %
Kappa statistic                          0.5645
Mean absolute error                      0.1837
Root mean squared error                  0.3006
Relative absolute error                 56.2821 %
Root relative squared error             74.4694 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.909     0.299      0.697     0.909     0.789      0.874    Negative
                 0.542     0.106      0.747     0.542     0.628      0.837    CIN1
                 0.683     0.047      0.745     0.683     0.713      0.918    CIN2/3
                 0.538     0          1         0.538     0.7        0.972    Cancer
Weighted Avg.    0.723     0.175      0.734     0.723     0.714      0.872

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 140  11   3   0 |   a = Negative
  54  71   6   0 |   b = CIN1
   7  12  41   0 |   c = CIN2/3
   0   1   5   7 |   d = Cancer
```

Εικόνα 62 Εκμάθηση γράφου, αύξηση της παραμέτρου delta, confusion matrix και αποτελέσματα σωστής και λανθασμένης παραμετροποίησης αποτελεσμάτων

```
Correctly Classified Instances         121               78.5714 %
Incorrectly Classified Instances        33               21.4286 %
Kappa statistic                          0.6706
Mean absolute error                      0.1604
Root mean squared error                  0.2921
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.896     0.092      0.882     0.896     0.889      0.916    Negative
                 0.74      0.183      0.661     0.74      0.698      0.825    CIN1
                 0.667     0.05       0.786     0.667     0.721      0.928    CIN2/3
                 0.5       0          1         0.5       0.667      0.924    Cancer
Weighted Avg.    0.786     0.11       0.793     0.786     0.785      0.889

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 60  7  0  0 |  a = Negative
  8 37  5  0 |  b = CIN1
  0 11 22  0 |  c = CIN2/3
  0  1  1  2 |  d = Cancer
```

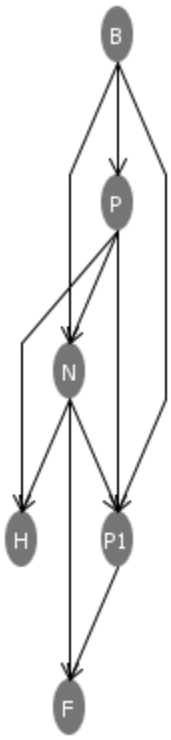
Εικόνα 63 Re-evaluation, νέο confusion matrix και νέα αποτελέσματα σωστής και λανθασμένης παραμετροποίησης αποτελεσμάτων

Στον παραπάνω πίνακα φαίνονται τα αποτελέσματα μετά την εκμάθηση του γράφου και μετά το re-evaluation.
Μετά το re-evaluation τα 121 αποτελέσματα ταξινομούνται σωστά, και τα 33 λανθασμένα. Από αυτά τα 110 ταξινομούνται ως TN, τα 7 ως FP, τα 9 ως FN και τα 28 ως TP.

Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,756757
- Ειδικότητα= $\frac{d}{b+d}$= 0,940171

Ο γράφος με την καλύτερη απόδοση είναι ο παρακάτω:



Εικόνα 64 Γράφος με την καλύτερη απόδοση για τη συγκεκριμένη παραμετροποίηση

Παρατηρήσαμε ότι όταν αυξάνεται η παράμετρος TStart και delta, προκύπτουν οι καλύτεροι γράφοι, με την καλύτερη ευαισθησία και ειδικότητα. Άρα παίζει ρόλο η τιμή εκκίνησης του αλγορίθμου καθώς και η πιθανότητα αποδοχής των βημάτων προς τη λάθος κατεύθυνση στο χώρο αναζήτησης) η ποιά μειώνεται σε κάθε επανάληψη. Η δεύτερη και η τέταρτη παραμετροποίηση έδωσαν τα καλύτερα αποτελέσματα.

9.2.4 K2 Αλγόριθμος

Ο τελευταίος αλγόριθμος στον οποίο μελετάται η παραμετροποίηση του και αξιολογείται αντίστοιχα, είναι ο K2. Οι διαθέσιμες παράμετροι έχουν αναλυθεί παραπάνω, και επιπλέον υπάρχει η παρακάτω:

- **randomOrder**: Όταν αυτή η παράμετρος είναι αληθής (true), η σειρά των κόμβων του δικτύου είναι τυχαία.

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα μετά την εκμάθηση του γράφου και μετά το re-evaluation.

```
Correctly Classified Instances         275               76.8156 %
Incorrectly Classified Instances        83               23.1844 %
Kappa statistic                          0.6448
Mean absolute error                      0.1538
Root mean squared error                  0.2942
Relative absolute error                 47.117  %
Root relative squared error             72.894  %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.89      0.176      0.792     0.89      0.838      0.925    Negative
                 0.611     0.088      0.8       0.611     0.693      0.863    CIN1
                 0.85      0.091      0.654     0.85      0.739      0.932    CIN2/3
                 0.538     0          1         0.538     0.7        0.962    Cancer
Weighted Avg.    0.768     0.123      0.779     0.768     0.763      0.905

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 137  15   2   0 |   a = Negative
  32  80  19   0 |   b = CIN1
   4   5  51   0 |   c = CIN2/3
   0   0   6   7 |   d = Cancer
```

Εικόνα 65 Εκμάθηση γράφου, confution matrix και ποσοστά επιτυχημένης και μη επιτυχημένης ταξινόμησης

```
Correctly Classified Instances         118               76.6234 %
Incorrectly Classified Instances        36               23.3766 %
Kappa statistic                          0.648
Mean absolute error                      0.1551
Root mean squared error                  0.3098
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.896     0.092      0.882     0.896     0.889      0.935    Negative
                 0.52      0.106      0.703     0.52      0.598      0.769    CIN1
                 0.909     0.14       0.638     0.909     0.75       0.935    CIN2/3
                 0.5       0          1         0.5       0.667      0.878    Cancer
Weighted Avg.    0.766     0.104      0.775     0.766     0.759      0.88

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 60  7  0  0 |  a = Negative
  8 26 16  0 |  b = CIN1
  0  3 30  0 |  c = CIN2/3
  0  1  1  2 |  d = Cancer
```

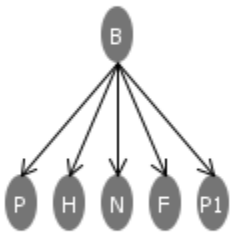
Εικόνα 66 Re-evaluation, νέο confution matrix και ποσοστά επιτυχημένης και μη επιτυχημένης ταξινόμησης

Μετά το re-evaluation τα 118 αποτελέσματα ταξινομήθηκαν σωστά και τα 36 λανθασμένα. Από αυτά τα 112 ταξινομήθηκαν ως TN, τα 5 ως FP, τα 12 ως FN και τα 25 ως TP.

Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,675676
- Ειδικότητα= $\frac{d}{b+d}$= 0,957265

Όπως βλέπουμε παρακάτω, ο καλύτερος γράφος που προκύπτει είναι Naive.



Εικόνα 67 Γράφος με την καλύτερη απόδοση

Με initAsNaiveBayes False έχουμε καλύτερη ταξινόμηση.

```
Correctly Classified Instances         272               75.9777 %
Incorrectly Classified Instances        86               24.0223 %
Kappa statistic                          0.6299
Mean absolute error                      0.157
Root mean squared error                  0.2973
Relative absolute error                 48.1003 %
Root relative squared error             73.6641 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.903    0.191      0.781     0.903      0.837      0.924    Negative
                 0.595    0.097      0.78      0.595      0.675      0.855    CIN1
                 0.8      0.084      0.658     0.8        0.722      0.935    CIN2/3
                 0.538    0          1         0.538      0.7        0.955    Cancer
Weighted Avg.    0.76     0.132      0.768     0.76       0.754      0.902

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 139  13   2   0 |   a = Negative
  35  78  18   0 |   b = CIN1
   4   8  48   0 |   c = CIN2/3
   0   1   5   7 |   d = Cancer
```

Εικόνα 68 Εκμάθηση

```
Correctly Classified Instances         120               77.9221 %
Incorrectly Classified Instances        34               22.0779 %
Kappa statistic                          0.6669
Mean absolute error                      0.1561
Root mean squared error                  0.3077
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.896    0.092      0.882     0.896      0.889      0.934    Negative
                 0.56     0.106      0.718     0.56       0.629      0.759    CIN1
                 0.909    0.124      0.667     0.909      0.769      0.947    CIN2/3
                 0.5      0          1         0.5        0.667      0.861    Cancer
Weighted Avg.    0.779    0.101      0.786     0.779      0.773      0.878

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 60  7  0  0 |  a = Negative
  8 28 14  0 |  b = CIN1
  0  3 30  0 |  c = CIN2/3
  0  1  1  2 |  d = Cancer
```

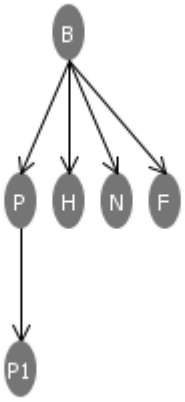
Εικόνα 69 Re-evaluation

Μετά το re-evaluation τα 120 αποτελέσματα ταξινομούνται σωστά, και τα 34 λανθασμένα. Από αυτά τα 103 ταξινομούνται ως TN, τα 14 ως FP, τα 4 ως FN και τα 33 ως TP.

Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,891892
- Ειδικότητα= $\frac{d}{b+d}$= 0,880342

Ο γράφος με την καλύτερη απόδοση είναι:



Εικόνα 70 Γράφος με την καλύτερη απόδοση

```
Correctly Classified Instances         276               77.095  %
Incorrectly Classified Instances        82               22.905  %
Kappa statistic                          0.6417
Mean absolute error                      0.1543
Root mean squared error                  0.2783
Relative absolute error                 47.2722 %
Root relative squared error             68.9528 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.864     0.152      0.811     0.864     0.836      0.93     Negative
                 0.74      0.172      0.713     0.74      0.727      0.877    CIN1
                 0.65      0.04       0.765     0.65      0.703      0.943    CIN2/3
                 0.538     0          1         0.538     0.7        0.957    Cancer
Weighted Avg.    0.771     0.135      0.774     0.771     0.769      0.913

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 133  20   1   0 |   a = Negative
  27  97   7   0 |   b = CIN1
   4  17  39   0 |   c = CIN2/3
   0   2   4   7 |   d = Cancer
```

Εικόνα 71 Εκμάθηση

```
Correctly Classified Instances         111               72.0779 %
Incorrectly Classified Instances        43               27.9221 %
Kappa statistic                          0.5773
Mean absolute error                      0.1714
Root mean squared error                  0.31
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.731     0.069      0.891     0.731     0.803      0.929    Negative
                 0.76      0.288      0.559     0.76      0.644      0.767    CIN1
                 0.667     0.058      0.759     0.667     0.71       0.94     CIN2/3
                 0.5       0          1         0.5       0.667      0.844    Cancer
Weighted Avg.    0.721     0.136      0.758     0.721     0.728      0.877

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 49 18  0  0 |  a = Negative
  6 38  6  0 |  b = CIN1
  0 11 22  0 |  c = CIN2/3
  0  1  1  2 |  d = Cancer
```

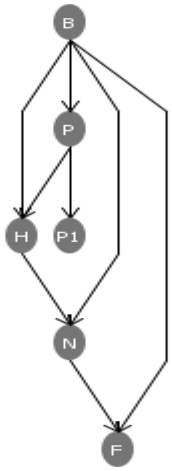
Εικόνα 72 Re-evaluation

Μετά το re-evaluation τα 111 αποτελέσματα ταξινομούνται σωστά, και τα 43 λανθασμένα. Από αυτά τα 111 ταξινομούνται ως TN, τα 6 ως FP, τα 12ως FN και τα 25 ως TP.

Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,675676
- Ειδικότητα= $\frac{d}{b+d}$= 0,948718

Ο γράφος με την καλύτερη απόδοση είναι ο παρακάτω:



Εικόνα 73 Γράφος με την καλύτερη απόδοση

```
Correctly Classified Instances         278               77.6536 %
Incorrectly Classified Instances        80               22.3464 %
Kappa statistic                          0.6526
Mean absolute error                      0.1521
Root mean squared error                  0.2766
Relative absolute error                 46.5923 %
Root relative squared error             68.5164 %
Total Number of Instances              358

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.864     0.152      0.811     0.864     0.836      0.932    Negative
                 0.725     0.15       0.736     0.725     0.731      0.883    CIN1
                 0.717     0.05       0.741     0.717     0.729      0.948    CIN2/3
                 0.538     0          1         0.538     0.7        0.959    Cancer
Weighted Avg.    0.777     0.129      0.779     0.777     0.775      0.918

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 133  19   2   0 |   a = Negative
  27  95   9   0 |   b = CIN1
   4  13  43   0 |   c = CIN2/3
   0   2   4   7 |   d = Cancer
```

Εικόνα 74 Εκμάθηση

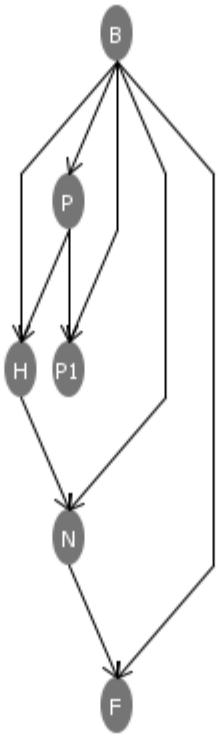
```
Correctly Classified Instances         110               71.4286 %
Incorrectly Classified Instances        44               28.5714 %
Kappa statistic                          0.5703
Mean absolute error                      0.1693
Root mean squared error                  0.3106
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.731     0.069      0.891     0.731     0.803      0.93     Negative
                 0.7       0.26       0.565     0.7       0.625      0.798    CIN1
                 0.727     0.091      0.686     0.727     0.706      0.925    CIN2/3
                 0.5       0          1         0.5       0.667      0.881    Cancer
Weighted Avg.    0.714     0.134      0.744     0.714     0.721      0.885

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 49 18  0  0 |  a = Negative
  6 35  9  0 |  b = CIN1
  0  9 24  0 |  c = CIN2/3
  0  0  2  2 |  d = Cancer
```

Εικόνα 75 Re-evaluation

Μετά το re-evaluation τα 110 αποτελέσματα ταξινομούνται σωστά, και τα 44 λανθασμένα. Από αυτά τα 108 ταξινομούνται ως TN, τα 9 ως FP, τα 9 ως FN και τα 28 ως TP.
Με βάση τους παραπάνω τύπους καθώς και τα αποτελέσματα της μήτρας που προκύπτει έχουμε:

- Ευαισθησία= $\frac{a}{a+c}$= 0,756757
- Ειδικότητα= $\frac{d}{b+d}$= 0,923077

Ο γράφος με την καλύτερη απόδοση είναι:



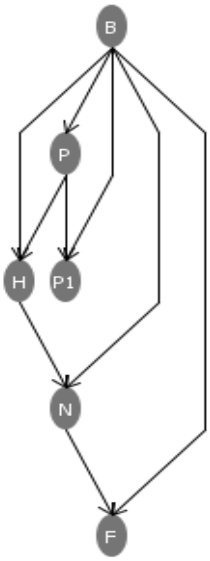
Εικόνα 76 Γράφος με την καλύτερη απόδοση

Ο καλύτερος γράφος που προέκυψε ήταν ο Naive και ήταν αυτός με την καλύτερη ευαισθησία και ειδικότητα.

Αφού ολοκληρώθηκε το βήμα με το test set δεδομένων, γίνεται το validation, ώστε να προκύψε ο καλύτερος γράφος από όλους..

Αυτό που παρατηρήθηκε γενικά είναι ότι ο Γενετικός Αλγόριθμος λόγω του ότι δημιουργεί πολλές γενιές απογόνων είναι αρκετά αργός και οι γράφοι που προέκυψαν από αυτόν δεν έχουν πολύ καλή ευαισθησία και ειδικότητα.

Μετά το validation, ο καλύτερος γράφος που προέκυψε φαίνεται παρακάτω:



Εικόνα 77 Ο καλύτερος γράφος μετά το validation

Κεφάλαιο 10ο

10 Συμπεράσματα

Ο καρκίνος του τραχήλου της μήτρας έχει αποτελέσει έναν από τους βασικούς λόγους θνησιμότητας των γυναικών και γι' αυτό είναι πολύ σημαντική η πρόωρη ανίχνευση του, ώστε η θεραπεία του να γίνει έγκαιρα, και τα ποσοστά θνησιμότητας να μειωθούν.

Γενικότερα, η μηχανική μάθηση είναι ένας πολύ χρήσιμος τομέας της βιοϊατρικής, αφού προσφέρει την υπόσχεση για τη βελτίωση της ευαισθησίας και της ακρίβειας της ανίχνευσης του καρκίνου του τραχήλου της μήτρας, καθώς και τη διάγνωση της, η οποία γίνεται με τη διαδικασία λήψης αποφάσεων. Στις μέρες μας συλλέγονται πολλά ιατρικά δεδομένα για μελέτη και αναπτύσσονται συνεχώς νέες μέθοδοι ανίχνευσης και διάγνωσης. Επειδή λοιπόν η πολυπλοκότητα των τύπων των δεδομένων που υπάρχουν είναι μεγάλη, υπάρχει όλο και μεγαλύτερη ανάγκη για χρήση της μηχανικής μάθησης στον τομέα της ιατρικής.

Στη συγκεκριμένη εργασία μελετήθηκαν οι αλγόριθμοι εκμάθησης γράφων και χρησιμοποιούνται από το πρόγραμμα WEKA για το classification των αποτελεσμάτων των εξετάσεων των ασθενών και στη συγκεκριμένη περίπτωση ασθενών με πιθανότητα να έχουν καρκίνο του τραχήλου της μήτρας (**Error! Reference source not found.**).

Τα συμπεράσματα μπορούν να χρησιμοποιηθούν ως επιβεβαίωση για τον κλινικό ιατρό ώστε να μην στείλει τον ασθενή για βιοψία, χωρίς οι πιθανότητες να έχει καρκίνο να είναι μεγάλες.

Σχετικά με τους αλγορίθμος, ο πιο αργός ήταν ο γενετικός, ο οποίος λόγω του ότι δημιουργεί μεγάλο αριθμό γενιών, δε δίνει υψηλή ευαισθησία και ειδικότητα στους γράφους. Τα καλύτερα αποτελέσματα προέκυψαν από τον K2. Σημαντικό ρόλο στον K2, παίζει ο αριθμός των γονέων. Όσο περισσότερους γονείς έχουμε, τόσο καλύτερα αποτελέσματα προκύπτουν.

Όμως, αυτό που προκύπτει εξετάζοντας όλες τις παραπάνω περιπτώσεις είναι ότι εάν η γυναίκα εξετάζεται τακτικά και κάνει το τεστ Παπανικολάου μία φορά το χρόνο, δεν υπάρχει περίπτωση, ακόμα και αν αρρωστήσει, να μην προλάβει να θεραπευτεί. Ο καρκίνος εξελίσσεται με πολύ γρήγορους ρυθμούς, αλλά όχι τόσο γρήγορους ώστε αν εντοπιστεί σε αρχικό στάδιο να μην μπορεί να καταπολεμιστεί.

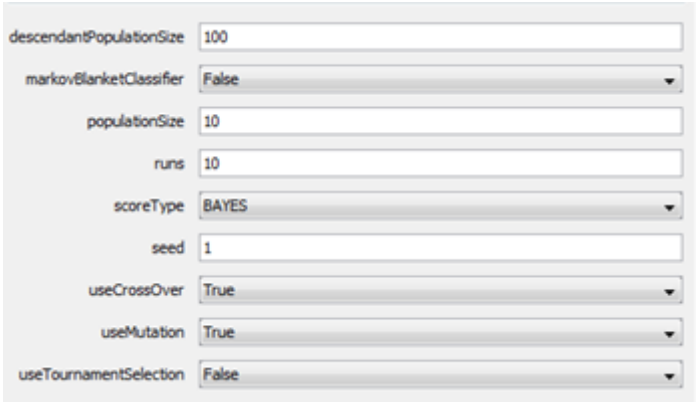
Τέλος, όπως βλέπουμε δε σημαίνει ότι εάν ένα τεστ Παπανικολάου βγει θετικό, ο ασθενής έχει σίγουρα καρκίνου του τραχήλου της μήτρας.

11 Βιβλιογραφικές Πηγές

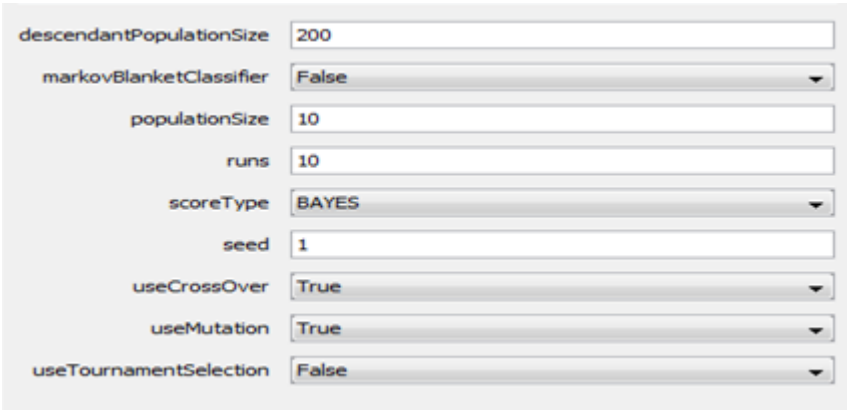
1. (Α.Παπακωνσταντίνου, 2012)
2. (Monograph on Human Papillomavirus)
3. (Wikipedia)
4. (P.J.F. Snijders, 2006)
5. (R.Narimatsu, 2005)
6. (C.Malloy, 2000)
7. (Μ.Αλεπάκη)
8. (M.Rahman, 2006)
9. (Λ.Καμπάς)
10. (N.Friedman, 2009)
11. (S. B. Kotsianntis, 2006)
12. (H. Kelly, 2008)
13. (Δημήτριος Κουτσούρης, 2013)
14. (Stuart Russell, 2005)
15. (M.H.DeGroot, 1989)
16. (E.Jaynes, 2003)
17. (George D.Magoulas, 2001)
18. (W.Feller, 1970)
19. (A.Dawid, 1979)
20. (A.Dawid, Conditional independence for statistical operations, 1980)
21. (J.Pearl, 1988)
22. (A.Becker, 2000)
23. (Neapolitan, 2003)
24. (Fradkin Dmitriy, 2006)
25. (Yonghong Peng, 2010)
26. (Peter J.F Lucas, 2004)
27. (M.Korver, 1993)
28. (Giorgio Valentini, 2009)
29. (Lucas, 1996)
30. (Ian H. Witten)
31. (Sajda, 2006)
32. (P.Smyth,1997)
33. (MI. Jordan,2004)
34. (K.Doi, M.L. Giger, RM.Nishikawa, K.Hoffmann, H. MacMahon, 1993)
35. (Jie Cheng, Russell Greiner, 2001)
36. (RE. Bird, 1990)

37. (S. Acid, L. M. de Campos, J. M. Fernandez- Luna, S. Rodriquez, J. M. Rodriquez, J.L. Salcedo, 2004)
38. (Joseph A. Cruz, David S. Wishart, 2006)
39. (J. T. Horng, L. C. Wu, B. J. Liu, J. L. Kuo, W. H. Kuo, J. J. Zhang, 2009)
40. (Ciro Donalek, 2011)
41. (Li Fei- Fei, Rob Fergus, Pietro Perona, 2006)
42. (Prof. Carolina Ruiz, 2006)
43. (Heni Bouhamed, Afif Masmoudi, Thierry Lecroq, Ahmed Rebaï, 2008)
44. (Evelina Lamma, Fabrizio Riguzzi and Sergio Storari, Italy)
45. (Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, 2005)
46. (Ishibuchi, H..Dept. of Ind. Eng., 2000)
47. (Ishibuchi, H. ; Dept. of Ind. Eng., 2003)
48. (Eugeniusz Nowicki, Czesław Smutnicki., 2006)
49. (C. N. Fiechter, 2001)
50. (Gallego, R.A. ; UTP, Pereira, Colombia ; Romero, R. ; Monticelli, A.J., 2004)
51. (Edmund Burke, Patrick De Causmaecker, Greet Vanden Berghe, 1999)
52. (Takumi OHASHI, Zaher AGHBARI, Akifumi MAKINOUCHI., Japan)
53. (Russell Greiner, 2007)
54. (Weidong Xiao, 2001)
55. (Heinz Muhlenbein, 1991)
56. (Davis, L., 1999)
57. (William L. Goffe, Gary D. Ferrier, John Rogers, 2002)
58. (Sakelaropoulos, Nikiforidis, 2000)
59. (E. Burnside, D. Rubin, and R. Shachter, 2000)
60. (Aronsky D, Haug PJ., 2004)
61. (Jesse Davis, Elizabeth Burnside, Ines de Castro Dutra, David Page, Vitor Santos Costa, 1998)
62. (Rudini Menezes Sampaio, Felipe Leal Valentim, Leondro Alves De Souza, Ricardo Martins De Abreu Silva, 2000)
63. (Marco Antônio Pinheiro de Cristoa, Pável Pereira Caladoa, Maria de Lourdes da ilveiraa, Ilmério Silvab, Richard Muntzc, Berthier Ribeiro-Neto, 2003)

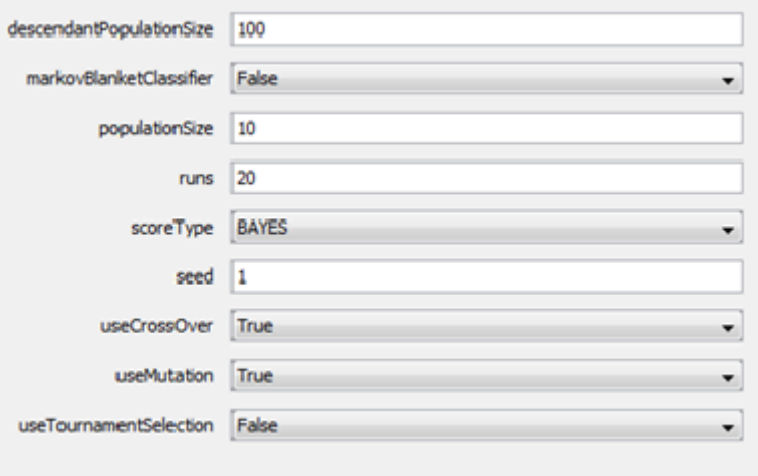
12 Παραμετροποίηση Αλγορίθμων



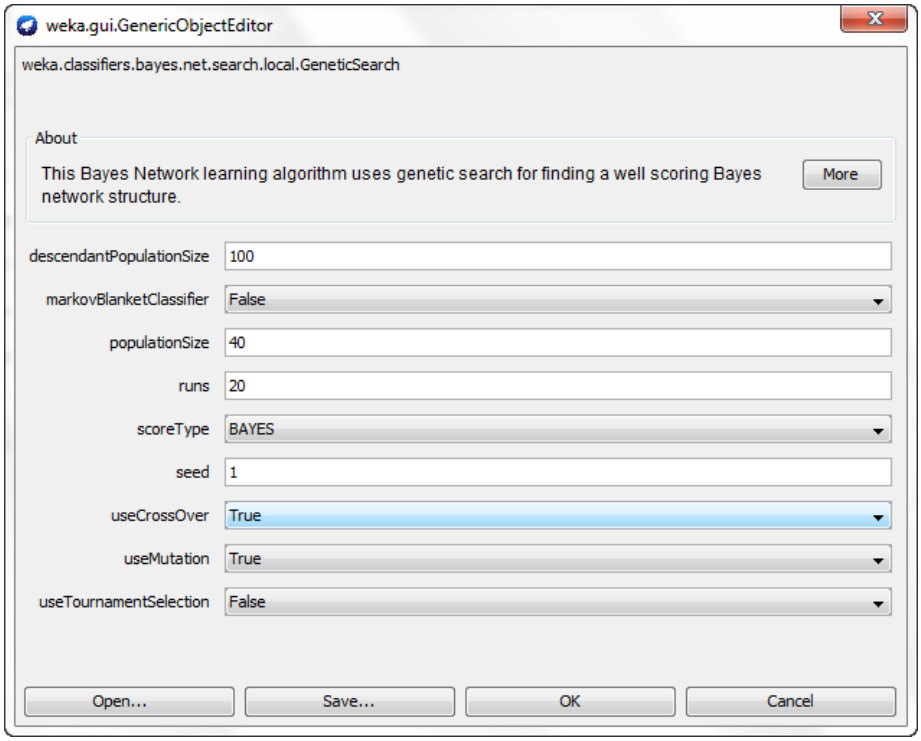
Εικόνα 78 Παραμετροποίηση Γενετικού Αλγορίθμου με descendanntPopulationSize= 100



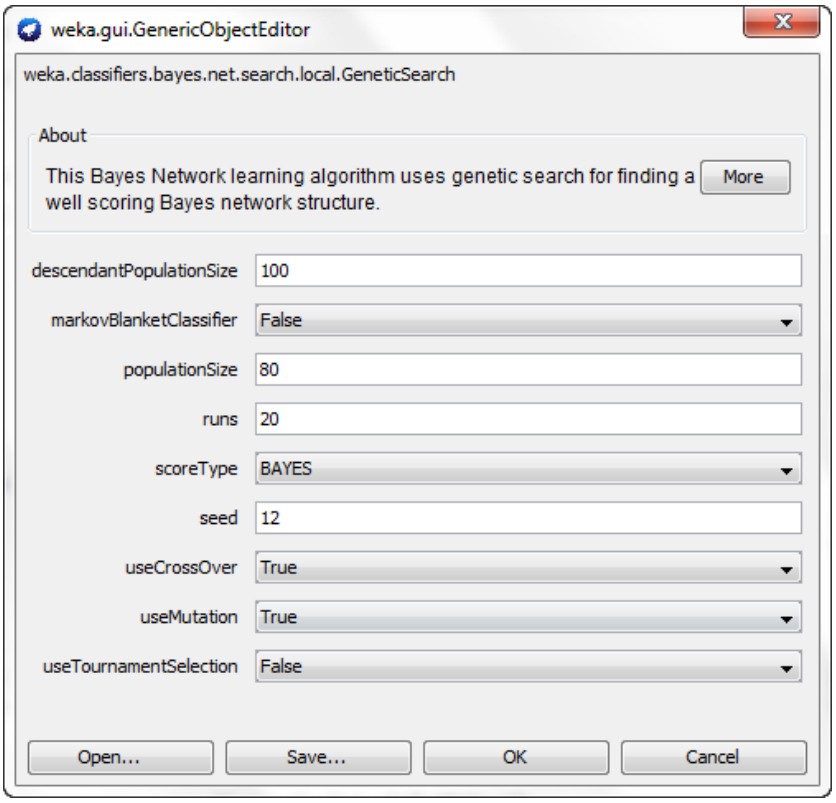
Εικόνα 79 Παραμετροποίηση Γενετικού αλγορίθμου με αύξηση της παραμέτρου descendantPopulationSize, δηλαδή του μεγέθους του πληθυσμού που επιλέγεται σε κάθε γενιά, από 100 σε 200



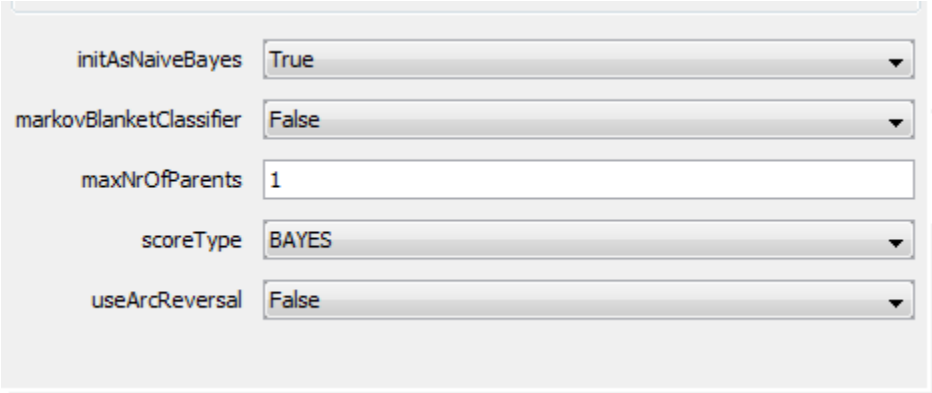
Εικόνα 80 Παραμετροποίηση Γεντικού αλγορίθμου, με αύξηση των παραγόμενων γενιών, δηλαδή της παραμέτρου runs, σε 20



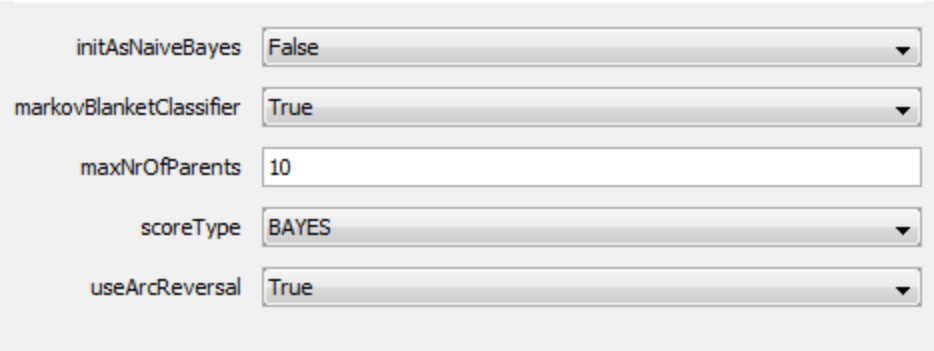
Εικόνα 81 Παραμετροποίηση Γενετικού αλγορίθμου με αύξηση του PopulationSize σε 40



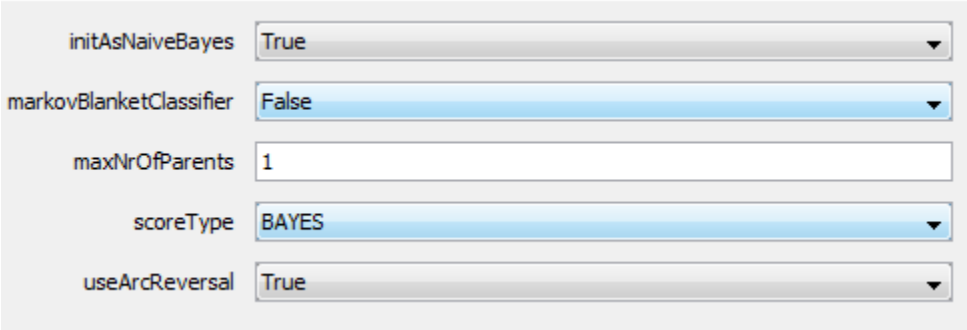
Εικόνα 82 Παραμετροποίηση Γενετικού αλγορίθμου με αύξηση των παραμέτρων runs, seed



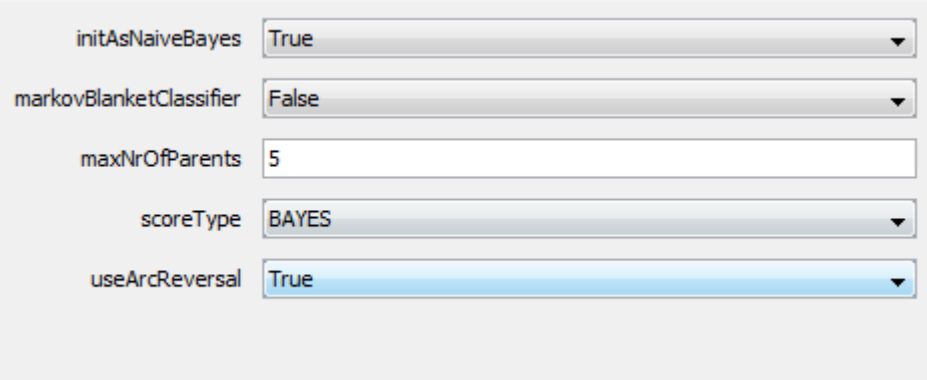
Εικόνα 83 Παραμετροποίηση αλγορίθμου Hill Climbing



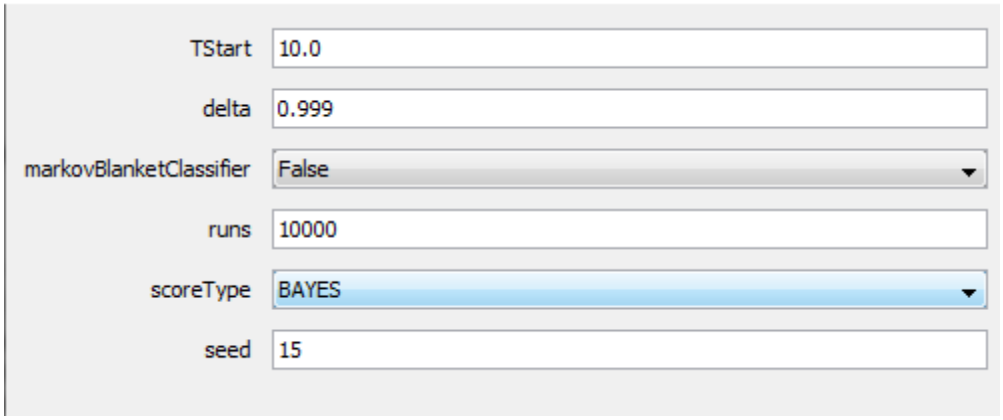
Εικόνα 84 Παραμετροποίηση Hill Climbing, αύξηση της παραμέτρου maxNrOfParents και αλλαγή παραμέτρων initAsNaiveBayes, markovBlanketClassifier και useArcReversal



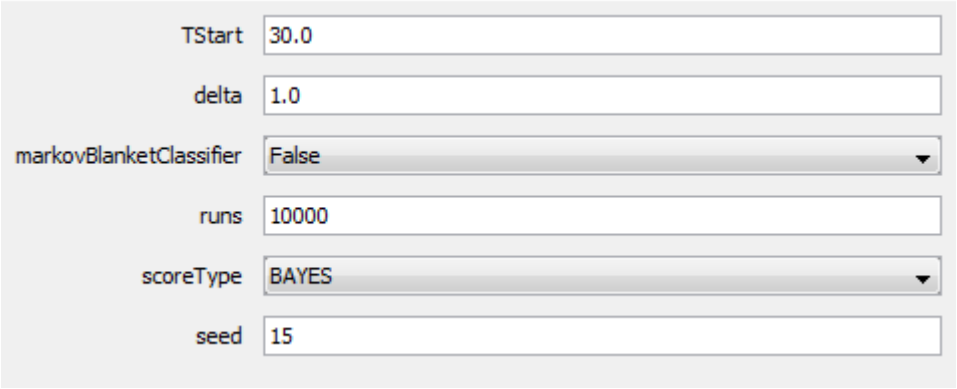
Εικόνα 85 Παραμετροποίηση αλγορίθμου Hill Climbing και αλλαγή παραμέτρου useArcReversal



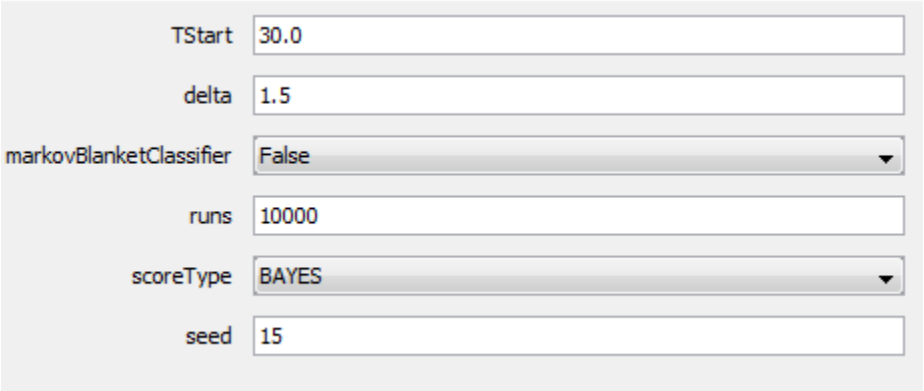
Εικόνα 86 Παραμετροποίηση αλγορίθμου Hill Climbing και αύξηση της παραμέτρου maxNrOfParents



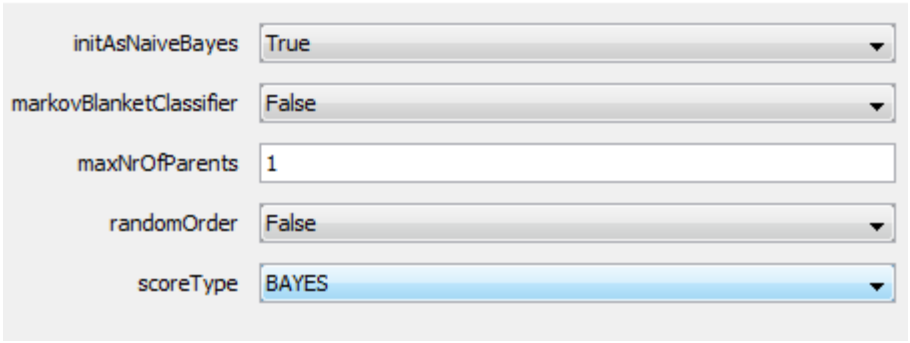
Εικόνα 87 Παραμετροποίηση αλγορίθμου Simulated annealing



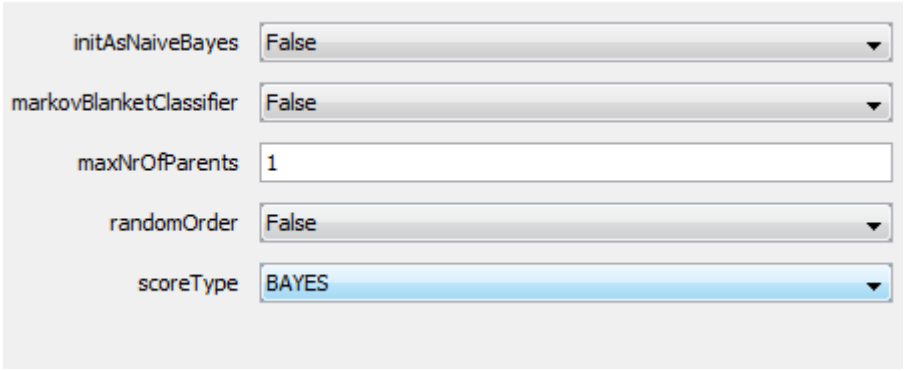
Εικόνα 88 Παραμετροποίηση αλγορίθμου Simulated annealing και αύξηση της παραμέτρου TStart και delta



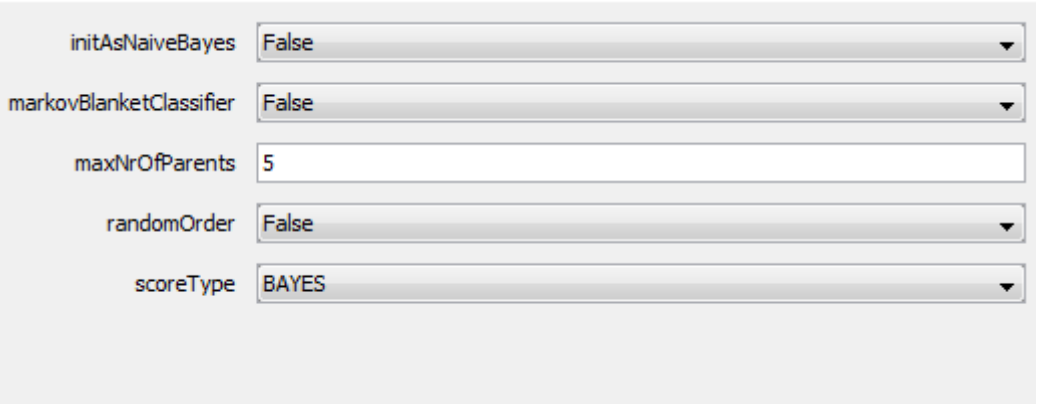
Εικόνα 89 Παραμετροποίηση αλγορίθμου Simulated annealing και αύξηση της παραμέτρου delta



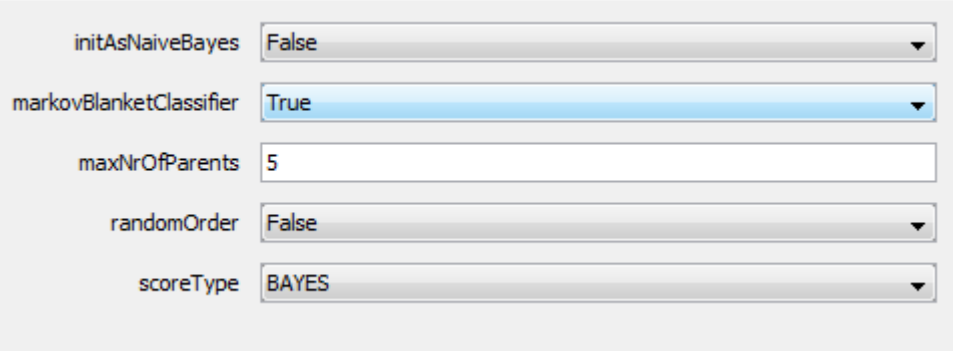
Εικόνα 90 Παραμετροποίηση αλγορίθμου K2, με maxNrOfParents=1



Εικόνα 91 Παραμετροποίηση αλγορίθμου K2, με αλλαγή της παραμέτρου initAsNaiveBayes από True σε False



Εικόνα 92 Παραμετροποίηση αλγορίθμου K2, με αύξηση της παραμέτρου maxNrOfParents



Εικόνα 93 Παραμετροποίηση αλγορίθμου K2, και αλλαγή της παραμέτρου markovBlanketClassifier σε True