ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

Τμήμα Πληροφορικής

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ¨ΠΡΟΗΓΜΕΝΑ ΣΥΣΤΗΜΑΤΑ ΠΛΗΡΟΦΟΡΙΚΗΣ¨

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

«Μηχανική Μάθηση σε Ανομοιογενή Δεδομένα

(Machine Learning in Imbalanced Data Sets)»

«Ανδρικάκης Ανδρέας»

«ΜΠΣΠ14005»

Επιβλέπων: Γ. Τσιχριντζής , Καθηγητής

Πειραιάς, *«Σεπτέμβριος » «2017»*

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Πληροφορικής

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ¨ΠΡΟΗΓΜΕΝΑ ΣΥΣΤΗΜΑΤΑ ΠΛΗΡΟΦΟΡΙΚΗΣ¨

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

«Μηχανική Μάθηση σε Ανομοιογενή Δεδομένα (Machine Learning in Imbalanced Data Sets)»

«Ανδρικάκης Ανδρέας»

«ΜΠΣΠ14005»

Επιβλέπων: Γ. Τσιχριντζής , Καθηγητής

Εγκρίθηκε από την εξεταστική επιτροπή

…… …… ……

Επιτελική Σύνοψη

Η ερευνητική εργασία που εκπονήθηκε με τίτλο ¨Μηχανική Μάθηση σε Ανομοιογενή Δεδομένα ¨ (Machine Learning in Imbalanced Data Sets) αποτελεί μια διατριβή που ολοκληρώθηκε στα πλαίσια του μεταπτυχιακού προγράμματος ¨Προηγμένα Συστήματα Πληροφορικής ¨ του Τμήματος Πληροφορικής του Πανεπιστήμιου Πειραιώς.

Η εργασία δίνει ιδιαίτερη βάση στην επίλυση του προβλήματος αποδοτικής χρήσης αλγορίθμων ταξινόμησης μηχανικής μάθησης για ανομοιογενή δεδομένα. Περιγράφει και αναλύει βασικούς αλγορίθμους μηχανικής μάθησης καθώς και αναφέρει τις μεθόδους αντιμετώπισης του προβλήματος των ανομοιογενών δεδομένων. Παρουσιάζονται εκτενέστερα αλγοριθμικές τεχνικές που διαχειρίζονται ανομοιογενή δεδομένα όπως οι αλγόριθμοι AdaCost, Cost Senistive Boosting, Metacost και άλλοι.

Περιλαμβάνει και παρουσιάζει συνοπτικά την βιβλιογραφία γύρω από τα θέματα και την αξιολόγηση των αλγορίθμων μηχανικής μάθησης και δίνει έμφαση στην κατανόηση και τον τρόπο ταξινόμησης των δεδομένων.

Στο τελευταίο μέρος της ερευνητικής εργασίας εξετάζονται μετρικές αξιολογήσεις των μεθόδων Μηχανικής Μάθησης σε ανομοιογενή δεδομένα όπως είναι οι καμπύλες διαχείρισης λειτουργικών χαρακτηριστικών (ROC curves), καμπύλες ακριβείας (PR curves) αλλά και οι καμπύλες κόστους.

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ

1.1 Εισαγωγικές Έννοιες

Οι βασικές πρωταρχικές έννοιες της Μάθησης (Learning), Νοημοσύνης (Intelligence) και η Εξόρυξη Δεδομένων(Data Mining) με τις αντίστοιχες σύγχρονες έννοιες της Μηχανικής Μάθησης (Machine Learning), Τεχνικής Νοημοσύνης (Imbalance Data Sets) από Βάσεις Δεδομένων , είναι στενά συνδεδεμένες και αλληλο-επιδρούσες μεταξύ τους όσον αφορά θεωρητικά και εφαρμοσμένα προβλήματα των υπολογιστικών επιστημών(computing sciences) και των πολυποίκιλων εφαρμογών τους.

Η μάθηση μπορεί να αναφέρεται στη απομνημόνευση εμπειριών αλλά και στην δημιουργία πολύπλοκων επιστημονικών θεωριών. Αναφέρονται διάφορες μορφές μάθησης, ταξινομήσεις, κατηγοριοποιήσεις , όπως επαγωγική(Jevons, 1874; Plotkin, 1971; Quilan, 1990), υπολογιστική συλλογική (Kolmogorov, 1965; Kearns and Vazirani,1994; Antony and Bartlett, 1999), ενισχυτική μάθηση , μάθηση δένδρων αποφάσεων (Friejenbaum,1961;Breimann et al.,1984;Breiman,1996), στατιστικές μέθοδοι μάθησης (DeGroot,1970;Hastie et al.,2001;Domingos and Pazzani,1997;Elkan,1997,Cristianni and Scholkopf,2002), σχέσης γνώσης και μάθησης (Fikes et al.,1972;Mitchel et al.,1986;Minton 1988; Dietterich,1990).

Ο όρος Τεχνητή Νοημοσύνη (ΤΝ) αναφαίρετα σε σχετικά νέο επιστημονικό πεδίο που περιλαμβάνει ένα ευρύ φάσμα απλών και πολύπλοκων διαδικασιών όπως π.χ. κατανόηση ανθρώπινης σκέψης , κατασκευή νοημόνων οντοτήτων, γενικές διαδικασίες μάθησης και αντίληψης , συστηματοποίηση και αυτοματοποίηση διανοητικών εργασιών και δραστηριοτήτων ,διαδικασίες σκέψης ,συλλογιστική συμπεριφορά, νοημοσύνη ορθολογικότητα, αναπαράσταση γνώσης ,μηχανική μάθηση κ.α. Συνοπτικά παρουσιάζονται διάφοροι ορισμοί της ΤΝ, βασικές αρχές , εξέλιξη και σύγχρονες τάσεις και εφαρμογές της στην επιστήμη και τεχνολογία.

1.2 Μάθηση Μηχανών και Εφαρμογές

1.2.1 Εισαγωγή

Η Μηχανική Μάθηση ή Γνωστικές Μηχανές (Machine Learning) ,ένας κλάδος της Τεχνητής Νοημοσύνης , είναι ένα επιστημονικό πεδίο που αναφέρεται στο σχεδιασμό και στην ανάπτυξη αλγορίθμων που δέχονται ως είσοδοι (input), εμπειρικά δεδομένα , όπως εκείνα που προέρχονται από αισθητήρες (sensors) ή βάσεις δεδομένων και δίνει σχέδια ή σχετικές προβλέψεις για τα χαρακτηριστικά των εμπλεκόμενων μηχανισμών που δημιούργησαν δεδομένα. Τα κυρία χαρακτηριστικά των άγνωστων βασικών κατανομών πιθανοτήτων μπορούν να γίνουν γνωστά έτσι ώστε τα δεδομένα να χρησιμοποιηθούν με αποδοτικό τρόπο από ένα εκπαιδευόμενο. Τέτοια δεδομένα μπορούν να θεωρηθούν ως περιπτώσεις πιθανών σχέσεων μεταξύ των παρατηρούμενων μεταβλητών.

Ένα από τα κύρια αντικείμενα της έρευνας μάθησης μηχανών είναι ο σχεδιασμός αλγορίθμων που αναγνωρίζουν πολυσύνθετα σχέδια και να λάβουν νοήμονες αποφάσεις βασισμένες στα δεδομένα εισόδου. Μια βασική δυσκολία είναι ότι ομάδα όλων των δυνατών συμπεριφορών με όλα τα πιθανά δεδομένα εισόδου είναι πολύ μεγάλη για να συμπεριληφθεί σε ένα σύνολο δεδομένων που έχουν παρατηρηθεί (επιλεγμένα δεδομένα[training data]). Με βάση τα προηγούμενα , ο εκπαιδευόμενος (Learner) πρέπει να γενικεύσει από τα δεδομένα παραδείγματα έτσι ώστε να μπορεί να παράγει χρήσιμα συμπεράσματα για νέα προβλήματα.

Η Μάθηση Μηχανών μπορεί εναλλακτικά να ορισθει ως ένα επιστημονικό πεδίο που προσδίδει στους υπολογιστές την ικανότητα να μάθουν χωρίς να έχουν άμεσα προγραμματιστεί για τέτοιο σκοπό (Samuel 1959). Ο A. Samuel ένας Αμερικανός καθηγητής του πανεπιστημίου του Stanford, πρωτοπόρος στα πεδία Τεχνητής Νοημοσύνης και Παιγνίων Υπολογιστών, παρουσίασε το 1959 ένα πρόγραμμα υπολογιστή για το παιχνίδι Ντάμας (Checkers-playing program), που θεωρείται ως το πρώτο πρόγραμμα αυτομαθησης υπολογιστή.Ένας πλέον τυπικός ορισμός σχετικός την Εκμάθηση Μηχανών αναφέρει ότι προγράμματα υπολογιστών μπορούν να μάθουν την εμπειρία E, όσον αφορά μερικές τάξεις έργων T (Tasks) και μέτρων απόδοσης P(Performance

Measures), εάν οι αποδόσεις τους στα έργα T, όπως μετρήθηκαν με P, βελτιώνεται με την εμπειρία E.

Αναφέρεται ότι μια μηχανή μαθαίνει κάθε φορά που αλλάζει την δομή του ,το πρόγραμμα ή τα δεδομένα , που βασίζονται στα δεδομένα εισόδου ή σε ανταπόκριση εξωτερικών πληροφοριών , έτσι ώστε η αναμενόμενη απόδοση θα βελτιωθεί. Τέτοιες αλλαγές όπως η πρόσθεση εγγράφων σε μια βάση δεδομένων , ανήκουν σε δικαιοδοσίες άλλων γνωστικών αντικειμένων και γενικά είναι γνωστές ως μάθηση . Για παράδειγμα , όταν η απόδοση μιας μηχανής αναγνώρισης ομιλίας βελτιώνεται μετά από το άκουσμα διάφορων δειγμάτων της ομιλίας ενός ατόμου, κατανοούμε και μπορούμε να πούμε ότι η μηχανή έχει μάθει. Η Μάθηση Μηχανών συνήθως αναφέρεται σε αλλαγές σε συστήματα που εκτελούν διαδικασίες (αναγνώριση, διάγνωση, σχεδιασμός, έλεγχος ρομπότ, προβλέψεις, κλπ.), που σχετίζονται με την Τεχνητή Νοημοσύνη(AI).

Διάφορα γνωστικά αντικείμενα έχουν χρησιμοποιηθεί στην Μάθηση Μηχανής , όπως

- Στατιστική (Anderson 1958),
- Μοντέλα Εγκεφάλου(McCulloch et al. 1943, Rosenblatt 1958, Sejnowski et al. 1988, Gluk et al. 1989),
- Προσαρμοσμένη Θεωρία Ελέγχου(Bolling et al. 1988, Sutton et al. 1987),
- Τεχνητή Νοημοσύνη (Samuel 1959, Carbonell 1983, Laird et al. 1986, Minton 1988, Kolodner 1993, Quinlan 1990, Muggleton 1991, Etzioni 1993, Lavra et al. 1994),
- Εξελικτικά Μοντέλα [Evolutionary Models](Γενετικοί Αλγόριθμοι(Holland 1975, Γενετικός Προγραμματισμός (Koza 1992-1994)).

1.2.2 Εφαρμογές Μάθησης Μηχανών

Συμφώνα με το έγκυρο διεθνές επιστημονικό σχήμα ταξινόμησης της ACM για τα γνωστικά αντικείμενα της Πληροφορικής , η Μάθηση Μηχανών είναι ένας κλάδος της Τεχνητής Νοημοσύνης (Artificial Intelligence),που αποτελεί μια υποκατηγορία των Μεθοδολογιών Υπολογισμών ,βασικής κατηγορίας της Πληροφορικής.

Το πολυδιάστατο φάσμα εφαρμογών της Μάθησης Μηχανών Περιλαμβάνει τα ακόλουθα σημαντικά γνωστικά αντικείμενα:

Εφαρμογές Προηγμένων Υπολογισμών

- Νόηση Μηχανών(Machine perception)
- Ενόραση Υπολογιστών (Computer vision)
- Επεξεργασία Φυσικών Γλωσσών (Natural language processing)
- Αναγνώριση Σχεδίων(Syntactic pattern recognition)
- Μηχανές Αναζήτησης(Search engines)

Γενικές Εφαρμογές Υπολογιστών

- Μηχανική Λογισμικού (Software engineering)
- Μετακινήσεις Ρομπότ(Robot locomotion)
- Αναγνώριση Ομιλίας και Χειρογράφων (Speech and handwriting recognition)
- Προσαρμοζόμενοι Ισότοποι (Adaptive websites)
- Αναγνώριση Αντικειμένων στην Ενόραση Υπολογιστών (Object recognition in Computer vision)
- Παίξιμο Παιγνίων (Game playing)
- Εξόρυξη Ακολουθιών (Sequence mining)
- Εξόρυξη Γνωμών (Sentiment Analysis or Opinion Mining)
- Συναισθηματικοί Υπολογισμοί (Affective computing)
- Ανάκτηση Πληροφοριών (Information Retrieval)
- Συνιστώμενα Συστήματα (Recommender systems)

Προηγμένοι Υπολογισμοί και Οικονομία

- Υπολογιστική Οικονομία (Computational finance)
- Ανάλυση Αγοράς Μετοχών (Stock market analysis)
- Ανίχνευση Απάτης Πιστωτικών Καρτών (Detecting credit card fraud)

Προηγμένοι Υπολογισμοί και Ιατρική

- Βιοπληροφορική (Bioinformatics)
- Μέσα Αλληλεπίδραση Εγκεφάλου-Μηχανών (Brain-machine interfaces)
- Χημειοπληροφορική (Cheminformatics)
- Ιατρικές Διαγνώσεις(Medical diagnosis)
- Δομημένος Έλεγχος Υγείας (Structural health monitoring)
- Ταξινόμηση Ακολουθιών DNA (Classifying DNA sequences)

1.3 Μάθηση: Μορφές Μάθησης

Οι πράκτορες μάθησης μπορεί να περιλαμβάνουν ορισμένα στοιχεία εκτέλεσης, τα οποία αποφασίζουν ποιες ενέργειες θα πραγματοποιηθούν και στοιχεία μάθησης που τροποποιούν τα στοιχεία εκτέλεσης έτσι ώστε να λαμβάνουν καλύτερες αποφάσεις.

Οι ερευνητές μηχανικής μάθησης χρησιμοποιούν μια μεγάλη ποικιλία στοιχείων μάθησης των οποίων η σχεδίαση επηρεάζεται από το περιβάλλον στο οποίο εφαρμόζονται.

Η σχεδίαση αυτή επηρεάζεται από τους ακόλουθους παράγοντες:

i. Ποιες συνιστώσες στοιχείων εκτέλεσης πρέπει να κοινοποιηθούν,

ii. Ποιες αναδράσεις πρόκειται να διατεθούν για τη μάθηση των συνιστωσών αυτών,

iii. Ποιες αναπαραστάσεις χρησιμοποιούνται για τις συνιστώσες.

Για τη δημιουργία στοιχείων εκτέλεσης υπάρχουν διάφοροι τρόποι, ενώ οι συνιστώσες των πρακτόρων περιλαμβάνουν ορισμένες πληροφορίες και στόχους. Για κάθε μια από τις συνιστώσες μπορεί να υπάρξει μάθηση μέσα από κατάλληλη ανάδραση.

Ο τύπος ανάδρασης για τη μάθηση σχετίζεται με τον προσδιορισμό της φύσης μαθησιακών προβλημάτων που αντιμετωπίζουν οι πράκτορες. Η μηχανική μάθηση περιλαμβάνει τρεις διακεκριμένες κατηγορίες μάθησης:

i. επιβλεπόμενη μάθηση (supervised learning):μάθηση συνάρτησης από παραδείγματα εισόδων και εξόδων πρακτόρων,

ii. μη-επιβλεπόμενη μάθηση (unsupervised learning): μάθηση προτύπων εισόδων χωρίς να δίνονται συγκεκριμένες τιμές εξόδων,

iii. ενισχυτική μάθηση : χρήση παρατηρημένων ειδών αναδράσεων (ανταμοιβές ή ενισχύσεις) για μάθηση (σχεδόν) βέλτιστης πολίτικης για το περιβάλλον.

Για τη λειτουργία αλγορίθμων μάθησης χρειάζονται αναπαραστάσεις γνωστών πληροφοριών, οι οποίες αντιστοιχούν στις συνιστώσες των πρακτόρων, π.χ. αλγόριθμος μάθησης

για πιθανοτικές περιγραφές (δίκτυα Bayes) για συμπερασματικές συνιστώσες πρακτόρων θεωρίας αποφάσεων, αλγόριθμος μάθησης για γραμμικά σταθμισμένα πολυώνυμα για συναρτήσεις χρησιμότητας σε προγράμματα παιχνιδιών κ.α.

Για το σχεδιασμό μαθησιακών συστημάτων χρειάζεται η διαθεσιμότητα προηγμένων γνώσεων. Το μεγαλύτερο μέρος της ανθρώπινης μάθησης προκύπτει μέσα στα πλαίσια μεγάλου όγκου σχετικών γνώσεων και πληροφοριών, ευρίσκονται στο Διαδίκτυο(Internet) και σε σχετικές βάσεις δεδομένων(Data bases).

1.3.1 Επαγωγική Μάθηση

Η επαγωγική μάθηση αφορά τη μάθηση συναρτήσεων από παραδείγματα εισόδων και εξόδων. Η μάθηση με διακριτές τιμές ονομάζεται ταξινόμηση, ενώ η μάθηση συνέχων συναρτήσεων καλείται παλινδρόμηση.

Η επαγωγική μάθηση περιλαμβάνει τη εύρεση μια συνεπούς υπόθεσης που συμφωνεί με τα παραδείγματα. Η δυνατότητα εύρεσης μια απλής συνεπούς υπόθεσης εξαρτάται από την επιλογή του χώρου υποθέσεων. Το πρόβλημα μάθησης χαρακτηρίζεται ως εφικτό αν ο χώρος υποθέσεων περιλαμβάνει τη πραγματική συνάρτησης ,διαφορετικά το πρόβλημα μάθησης χαρακτηρίζεται ως ανέφικτο (Jevons, 1874; Plotkin, 1971; Quilan, 1990).

1.3.2 Υπολογιστική Μάθηση

Η θεωρία υπολογιστικής μάθησης (computational learning theory) είναι ένα πεδίο που ανήκει στη τομή των τριών συνόλων: ΤΝ, στατιστική και θεωρητική επιστήμη υπολογιστών. Οι αλγόριθμοι μάθησης που επιστρέφουν υποθέσεις που είναι πιθανώς προσεγγιστικά σωστές ονομάζονται αλγόριθμοι μάθησης PAC (probably approximately correct).

Μια υπόθεση h θεωρείται προσεγγιστικά σωστή αν το αντίστοιχο σφάλμα e(h) είναι μικρότερο ή ίσο μια μικρής σταθεράς ε, δηλαδή $e(h) \leq \varepsilon$. Χαρακτηριστικό παράδειγμα για την επιλογή της απλούστερης συνεπούς υπόθεσης αποτελεί το αποκαλούμενο ξυράφι Ockham ,δηλαδή επιλέξτε την απλούστερη υπόθεση που συμφωνεί με τα δεδομένα.

Η θεωρία υπολογιστικής μάθησης αναλύει την πολυπλοκότητα δείγματος και την υπολογιστική δυσκολία της επαγωγικής μάθησης (Plotkin, 1971;, Shavlik and Dietterich, 1990; Weiss and Kulikowski, 1991; Kearns and Vazirani, 1994; Antony and Bartlett, 1999).

1.3.3 Στατιστικές Μέθοδοι Μάθησης

Οι ***μέθοδοι στατιστικής μάθησης*** μπορούν να χρησιμοποιηθούν για κατασκευή πολύπλοκων μοντέλων, όπως τα δίκτυα Bayes και τα νευρωνικά δίκτυα. Η μάθηση μπορεί να θεωρηθεί ως μια μορφή αβέβαιης συλλογιστικής από παρατηρήσεις. Η διεργασία μάθησης μπορεί να πάρει τη μορφή διαδικασίας θανατικού συμπεράσματος (τύπου Bayes), οπότε παρέχει γενικές λύσεις σε προβλήματα υπερπροσαρμογής, θορύβου και βέλτιστης πρόβλεψης.

Η ***μάθηση κατά Bayes*** (Bayesian Learning) υπολογίζει την πιθανότητα κάθε υπόθεσης με βάση τα δεδομένα και κάνει προβλέψεις σε αυτή τη βάση. Οι προβλέψεις γίνονται με χρήση όλων των υποθέσεων, που είναι σταθμισμένες σύμφωνα με τις πιθανότητες τους, αντί με χρήση μιας μόνο βέλτιστης υπόθεσης (DeGroot, 1970; Berger, 1985; Gelman et al., 1995; Hastie et al., 2001). Βασικές ποσότητες στην προσέγγιση Bayes είναι (i) η εκ των προτέρων πιθανότητα κάθε υπόθεσης και (ii) η πιθανοφάνεια των δεδομένων κάτω από κάθε υπόθεση.

Οι μέθοδοι στατιστικής μάθησης περιλαμβάνουν τη ¨*μάθηση παραμέτρων με πλήρη δεδομένα*¨ (complete data parameter learning) , που υπολογίζει αριθμητικές παραμέτρους για ένα μοντέλο πιθανοτήτων με σταθερή δομή. Το πρόβλημα μάθησης παραμέτρων μέγιστης πιθανοφάνειας με πλήρη δεδομένα για ένα δίκτυο Bayes αναλύεται σε ανεξάρτητα προβλήματα μάθησης με καθένα να αντιστοιχεί σε παράμετρο.

Το πλέον σύνηθες μοντέλο δικτύου Bayes που χρησιμοποιείται σε μηχανική μάθηση είναι το αποκαλούμενο ¨απλοϊκό μοντέλο Bayes¨ (naïve Bayes model) που θεωρεί ότι τα χαρακτηριστικά είναι υπό συνθήκη ανεξάρτητα μεταξύ τους με δεδομένη τη κατηγορία (Domingos and Pazzani, 1997; Elkan, 1997).

Τα μοντέλα, που βασίζονται σε στιγμιότυπα, αναπαριστούν κατανομές χρησιμοποιώντας συλλογές στιγμιότυπων εκπαίδευσης ,όπου ο αριθμός παραμέτρων αυξάνεται ανάλογα με το σύνολο εκπαίδευσης. Η μέθοδος ¨*πλησιέστερου γείτονα*¨ εξετάζει στιγμιότυπα που βρίσκονται

πλησιέστερα στο ερευνούμενο σημείο ενώ οι ¨μέθοδοι πυρήνων¨ (μηχανές διανυσμάτων υποστήριξης) σχηματίζουν ένα σταθμισμένο (ως προς την απόσταση) συνδυασμό όλων των στιγμιοτύπων (Aizerman et al., 1964; Boser et al., 1992; Vapnik, 1998; Cristianini and Scholkopf, 2002).

Τα ***νευρωνικά δίκτυα*** είναι πολύπλοκες πολύ-παραμετρικές μη-γραμμικές συναρτήσεις των οποίων οι παράμετροι ,μπορούν να υπολογιστούν από θορυβώδη δεδομένα και έχουν χρησιμοποιηθεί σε πολλές εφαρμογές (Cowan and Sharp, 1988; Bishop, 1995; Ripley, 1996). Τα νευρωνικά δίκτυα πολλών επιπέδων (με προς τα εμπρός τροφοδότηση σήματος) μπορούν να αναπαραστήσουν οποιαδήποτε συνάρτηση, με την προϋπόθεση ότι διαθέτουν αρκετές μονάδες. Η ελαχιστοποίηση του σφάλματος εξόδου μπορεί να γίνει εφαρμόζοντας τον αλγόριθμο ¨*οπισθοδιάδοσης*¨ για τη μέθοδο ¨κατάβασης πλαγιάς¨ στο χώρο παραμέτρων. Ο αποκαλούμενος ¨αισθητήρας¨ (perceptron) είναι ένα νευρωνικό δίκτυο με προς τα εμπρός τροφοδότηση σήματος χωρίς κρυφές μονάδες, που μπορεί να αναπαραστήσει μόνο ¨γραμμικά διαχωρίσιμες συναρτήσεις¨ (Rosenblatt, 1957).

1.3.4 Ενισχυτική Μάθηση

Η ενισχυτική μάθηση (reinforcement learning) χρησιμοποιεί παρατηρημένες ειδικές αναδράσεις (ανταμοιβές ή ενισχύσεις) για τη μάθηση (σχεδόν) βέλτιστων πολιτικών για το περιβάλλον και σε πολλά πολύπλοκα πεδία θεωρείται ότι αποτελεί το μόνο εφικτό τρόπο για την εκπαίδευση προγραμμάτων έτσι ώστε να επιτυγχάνεται υψηλά επίπεδα αποδόσεων. Βέλτιστη πολιτική είναι η πολιτική που μεγιστοποιεί την αναμενόμενη συνολικά ανταμοιβή. Η ενισχυτική μάθηση θεωρείται ένας μικρόκοσμος του συνολικού προβλήματος τη ΤΝ, που μελετάται σε διάφορα απλοποιημένα περιβάλλοντα για να διευκολύνεται η πρόοδος.

Η *παθητική μάθηση* (passive learning) αναφέρεται στη μάθηση χρησιμότητας καταστάσεων ή ζευγών καταστάσεων-ενεργειών , συμπεριλαμβανομένων και της μάθησης μοντέλων περιβάλλοντος. Η **ενεργητική μάθηση** (active learning) αναφέρεται στη εξερεύνηση (exploration), όπου οι πράκτορες πρέπει να μάθουν τι κάνουν με σκοπό να μάθουν πώς να συμπεριφέρονται μέσα σε αυτό.

Η εκμάθηση του τρόπου με τον οποίο συνδέονται καταστάσεις και η πληροφόρηση των περιορισμών μεταξύ των καταστάσεων μπορεί να χρησιμοποιηθεί ο ¨*προσαρμόσιμος δυναμικός προγραμματισμός*¨ (adaptive dynamic programming), που μαθαίνει το μοντέλο μετάβασης περιβάλλοντος καθώς προχωρά και επιλύει αντίστοιχες διαδικασίες αποφάσεων Markov με χρήση μεθόδων δυναμικού προγραμματισμού.

Η ¨*ιεραρχική ενισχυτική μάθηση*¨ (hierarchical reinforcement learning) μπορεί να χρησιμοποιηθεί για επίλυση προβλημάτων (σε πολλά αφαιρετικά επίπεδα) με πολύπλοκες συμπεριφορές (Forestier and Varaiya, 1975; Parr and Russell, 1998; Dietterich, 2000; Sutton et al., 2000; Andre and Russell, 2002).

2 ΒΑΣΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

2.1 Μηχανική Μάθηση

Μηχανική Μάθηση (Machine Learning) καλείται η δημιουργία μοντέλων ή προτύπων από σύνολα δεδομένων με χρήση υπολογιστικών συστημάτων. Εκτός από ορισμούς της Μάθησης που έχουν διατυπωθεί (Simon, 1983; Minsky, 1985; Michalski 1986) τις τελευταίες δεκαετίες, έχουν δοθεί διάφοροι σχετικοί ορισμοί για τη Μηχανική Μάθηση (Carbonell, 1987; Mitchell, 1997; Witten & Frank, 2000).

Ένα **Γνωστικό ή Γνωσιακό Σύστημα** (cognitive system) είναι ένα φυσικό ή τεχνικό σύστημα επεξεργασίας πληροφοριών με ιδιότητες μάθησης, συλλογισμού , αντίληψης, λήψη αποφάσεων , επικοινωνίας, δράσης κ.α., μπορεί να χρησιμοποιήσει δυο βασικές ιδιότητες:

i. απόκτηση γνώσης κατά την διάρκεια αλληλεπίδρασης με το περιβάλλον του.
ii. βελτίωση της εκτέλεσης ενεργειών με επαναλήψεις (δηλαδή βελτίωση της απόδοσης του) .

Ο άνθρωπος παρατηρώντας το περιβάλλον του και προσπαθώντας να δημιουργήσει απλοποιημένες (αφαιρετικές) παραστάσεις του, κατασκευάζει διάφορα μοντέλα, χρησιμοποιώντας διαδικασίες που βασίζονται σε μεθόδους επαγωγικής μάθησης (inductive learning) και επαγωγή (induction). Επιπρόσθετα, ο άνθρωπος μπορεί να δημιουργήσει διάφορες νέες δομές που καλούνται πρότυπο (patterns).

Η διαδικασία δημιουργίας τέτοιων προτύπων και μοντέλων από διάφορα σύνολα δεδομένα καλείται **μηχανική μάθηση** (machine learning).

2.2 Βασικές Αρχές Αλγορίθμων Μηχανικής Μάθησης

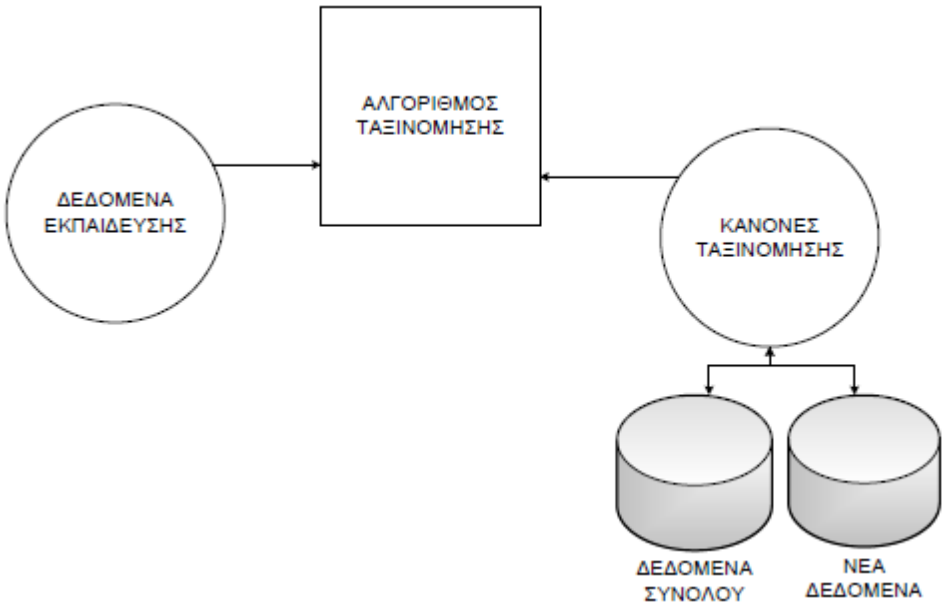
2.2.1 Εισαγωγικές Παρατηρήσεις

Υπάρχουν διαφορετικές κατηγορίες αλγορίθμων που σκοπό έχουν να δημιουργήσουν ένα μοντέλο ταξινόμησης και να κατηγοριοποιήσουν τα στιγμιότυπα σε ένα σύνολο έλεγχου.

Ο κάθε αλγόριθμος βασίζεται σε ξεχωριστά ποσοτικά μέτρα, έτσι ώστε να αναλύει διαφορετικά το σύνολο δεδομένων που δέχεται και να κατασκευάζονται διαφορετικά μοντέλα

ταξινόμησης. Κάποια βασικά μοντέλα είναι τα δέντρα αποφάσεων όπως οι πιο γνωστοί αλγόριθμοι είναι ID3 και C4.5, οι πιθανοκρατιών κατηγοριοποιητές όπως ο Naive Bayes, NaiveBayesNominal, και NaiveBayesSimple, οι διακριτοί κατηγοριοποιητές όπως οι multinomial και linear logistic regression κ.α.

Τα βασικά βήματα εκτέλεσης είναι δυο, αρχικά το στάδιο εκμάθησης και το στάδιο ταξινόμησης. Στο πρώτο στάδιο ο αλγόριθμος ταξινομητή κατασκευάζει τον ταξινομητή (classifier) και αναλύει το σύνολο δεδομένων εκπαίδευσης όπου αναπαριστάται συνήθως ως διάνυσμα χαρακτηριστικών (feature vector) της μορφής , X =< x1,x2,….xn>, με xi τα χαρακτηριστικά του στιγμιοτύπου που ανήκουν σε κλάση. Στο δεύτερο στάδιο εκτιμάται η ακρίβεια του αλγόριθμου που κατασκευάστηκε, με τις μεθόδους αποτίμησης ακρίβειας οι οποίες αναφέρονται σε επόμενο κεφάλαιο.



Σχήμα 2.1 : Λειτουργικότητα Αλγόριθμου Ταξινόμησης (Στο πρώτο στάδιο ο αλγόριθμος ταξινόμησης κατασκευάζει τον ταξινομητή και αναλύει το σύνολο δεδομένων εκπαίδευσης. Ο αλγόριθμος ταξινόμησης μπορεί να αναπαρασταθεί και ως σύνολα κανόνων, τους κανόνες ταξινόμησης για τα δεδομένα ή τα δεδομένα συνόλου.)

Χρησιμοποιώντας την έννοια της μάθησης, στόχος μας είναι το σύστημα να μάθει από μια συνάρτηση στόχο (target function) επαγωγικά και να εκφράσει τα δεδομένα ως ένα μοντέλο. Υπάρχουν δυο είδη μάθησης για τα δεδομένα μάθησης με επίβλεψη και μάθηση χωρίς επίβλεψη. Σε αυτό το εδάφιο θα ασχοληθούμε εκτενεστέρα με τη μάθηση με επιβλεπόμενα στοιχεία.

Στη μάθηση με επίβλεψη παρατηρούνται δυο είδη προβλημάτων (learning tasks). Το πρώτο είναι η *ταξινόμηση* (classification), στο οποίο δημιουργούνται μοντέλα πρόβλεψης διακριτών κλάσεων. Στο δεύτερο , το οποίο είναι η *παλινδρόμηση/παρεμβολή* (regression), δημιουργούνται μοντέλα πρόβλεψης αριθμητικών τιμών. Στη συνέχεια παρουσιάζεται μια σύντομη περιγραφή των πιο βασικών αλγορίθμων στη μηχανική μάθηση, όπως μάθηση εννοιών (concept learning),δένδρα ταξινόμησης απόφασης (decision trees), μάθηση κατά περίπτωση (instance-based learning), μάθηση κατά Bayes (Bayesian learning), παρεμβολή ή παλινδρόμηση (regression), νευρωνικά δίκτυα (neural networks), και μηχανές διανυσμάτων υποστήριξης (support vector machines).

2.2.2 Δένδρα Ταξινόμησης/Απόφασης

Τα δέντρα ταξινόμησης είναι μια από τις πιο δημοφιλείς ταξινόμησης στους αλγορίθμους μάθησης. Η μέθοδος αυτή παίρνει ως είσοδο ένα διάνυσμα τιμών σε κάποιες ιδιότητες και επιστρέφει μια έξοδο. Αυτή η έξοδος μπορεί να είναι διακριτή, οπότε ορίζεται ένα πρόβλημα ταξινόμησης, ενώ αν η έξοδος αυτή είναι συνεχής έχουμε ένα πρόβλημα παλινδρόμησης. Κυρίως σ αυτή την εργασία θα ασχοληθούμε με τα δέντρα ταξινόμησης ως προς προβλήματα ταξινόμησης (classification).

Τα δέντρα ταξινόμησης χρησιμοποιούνται για να προβλέψουν, με κάποιο βαθμό ακρίβειας, την τιμή της μεταβλητής που μοντελοποιούν με βάση τις τιμές των θεωρούμενων ανεξάρτητων χαρακτηριστικών.

Ένα δένδρο απόφασης αναπαριστά μια διαδικασία λήψης απόφασης, όπου για κάθε πιθανό σημείο ή κατάσταση έχουμε ένα κόμβο, ενώ για κάθε επιλογή που μπορεί να γίνει σε ένα σημείο απόφασης αναπαριστάται ένα «κόμβο-παιδί». Κάθε κόμβος ορίζει μια συνθήκη ελέγχου της τιμής κάποιου χαρακτηριστικού των περιπτώσεων. Κάθε κλαδί που φεύγει από ένα κόμβο αντιστοιχεί σε μια διαφορετική διακριτή τιμή του χαρακτηριστικού που σχετίζεται με τον κόμβο. Στα

κλαδιά καταλήγουν οι τελικοί κόμβοι που ανήκουν σε ένα μόνο σύνολο, όπου είναι οι τελικές αποφάσεις ή ενέργειες.

Για να διασπάσουμε ένα κόμβο (parent) με N εγγραφές σε κ παιδιά u_i, και ο αριθμός των εγγραφών είναι $N(u_i)$ με $N(u_i)$ με $\sum N(u_i) = N$.

Για να διαλέξουμε τη διάσπαση με το μεγαλύτερο κέρδος, υπολογίζουμε το μέγιστο Δ από την επόμενη σχέση

$$\Delta = \mathrm{I}(parent) - \sum_{i=1}^{k} \frac{N(u_i)}{N} i(u_i). \qquad (2.1)$$

Για να κατασκευάσουμε ένα δένδρο αποφάσεως παρουσιάζεται η ακόλουθη αλγοριθμική διαδικασία σε ψευδοκωδικη μορφή:

Βήμα 1: Δημιουργούμε ένα κόμβο που περιέχει όλες τις εγγραφές

Βήμα 2: Διασπάμε τον κόμβο με βάση μια συνθήκη, έτσι ώστε να διαχωριστούν οι εγγραφές σε κάποιο από τα γνωρίσματα.

Βήμα 3: Γίνεται αναδρομική κλίση του βήματος 2 σε κάθε κόμβο μέχρι να φτάσουμε στο τελικό κόμβο, έτσι ώστε να είναι μονοσύνολο (τελική απόφαση). Όταν κάποιος κόμβος δεν έχει παραδείγματα τότε αντιστοιχίζεται σε μια κατηγορία. Αν σε κάποιο κόμβο, υπάρχουν θετικά και αρνητικά παραδείγματα , αλλά έχουν εξαντληθεί όλα τα χαρακτηριστικά , τότε ο κόμβος χαρακτηρίζεται διφορούμενος. Τότε ο κόμβος αυτός μπορεί να αντιστοιχηθεί στην πλειοψηφία κατηγορία παραδειγμάτων ή στις κατηγορίες με τις αντίστοιχες συχνότητες εμφάνισης του. Αυτό μπορεί να συμβεί είτε από την ύπαρξη θορύβου στα δεδομένα ή παράβλεψη σημαντικών χαρακτηριστικών.

Βήμα 4: Αφού κατασκευαστεί το δένδρο αποφάσεων , μπορούν να γίνουν κάποιες βελτιστοποιήσεις με τη μέθοδο κλαδέματος (tree pruning).

Ο αριθμός των πιθανών δέντρων απόφασης είναι εκθετικός. Πολλοί αλγόριθμοι για να κατασκευάσουν ένα δένδρο απόφασης προσπαθούν να κτίσουν λαμβάνοντας μια σειρά από τοπικά βέλτιστες αποφάσεις. Οι πιο γνωστοί αλγόριθμοι είναι οι Hunt's Algorithm, CART, ID3, C4.5, SLIQ, SPRINT, BFTree, J48, J48Graft, LADTree, REPTree, SimpleCart κ.α. Εναλλακτικά τα δένδρα μπορούν

να αναπαρασταθούν και ως σύνολα κανόνων ιf then που ονομάζονται κανόνες ταξινόμησης (classification rules).

Πέρα από την κατασκευή του δένδρου , πρέπει να καθορίσουμε τις συνθήκες ελέγχου για τα γνωρίσματα . Εξαρτώνται από τον τύπο των γνωρισμάτων σε συνεχείς (continuous), διακριτές (norminal) και διατεταγμένες (ordinal) και από το είδος διαχωρισμού στο οποίο θα γίνει δυαδικός διαχωρισμός (2-way split) ή πολλαπλός διαχωρισμός (multi-way split). Με βάση το δυαδικό διαχωρισμό το σύνολο τιμών διαχωρίζεται σε δυο υποσύνολα, ώστε να βρει το βέλτιστο διαχωρισμό, ενώ με τον πολλαπλό χρησιμοποιούνται τόσες διασπάσεις όσες και διαφορετικές τιμές που δίνονται.

Μέτρα μη- Καθαρότητας

Στη κατασκευή ενός δέντρου τρία είναι τα μέτρα μη- καθαρότητας: η εντροπία , το ευρετήριο Gini και το λάθος ταξινόμησης.

Η **Εντροπία** (Entropy), μια θεμελιώδης έννοια που σχετίζεται με το δεύτερο νόμο της θερμοδυναμικής , ορίζει το μετρό της αταξίας ενός συστήματος και ορίζεται από τη σχέση

$$E(S) = -\sum_{i=1}^{c} p_i \log_2 (p_i),$$ όπου S είναι ένα σύνολο δεδομένων , (2.2)

pI είναι το ποσοστό των παραδειγμάτων του S που ανήκουν στην κατηγορία I και c διαφορετικές κατηγορίες.

Στη θεωρία πληροφοριών η **εντροπία της πληροφορίας** (entropy information) (REF) είναι ένας μηχανισμός διαχωρισμού που υπολογίζει ουσιαστικά την ανομοιογένεια που υπάρχει στο S αναφορικά με την υπό εξέταση εξαρτημένων μεταβλητών και ορίζεται από την σχέση:

$$E(S) = - p_+ . \log_2 (p_+) - p_- . \log_2 (p_-),$$ (2.3)

όπου S είναι το σύνολο των δεδομένων εκπαίδευσης στο στάδιο (κόμβο) του διαχωρισμού, p+ είναι το κλάσμα των θετικών παραδειγμάτων του S και p- είναι το κλάσμα των αρνητικών παραδειγμάτων του S.

Ένας εναλλακτικός τύπος για την εντροπία , που μέτρα την ομοιογένεια ενός κόμβου, ορίζεται ως

$$Entropy(t) = -\sum_{j=1}^{c} p(j/t) \log_2 p(j/t)$$

(2.4)

όπου p(j / t) είναι η σχετική συχνότητα της κλάσης j στο κόμβο t, c κλάσης.

Το **κέρδος πληροφορίας** (information gain) αναφέρεται στη μείωση εντροπίας ενός συνόλου εκπαίδευσης S με παράμετρο διαχωρισμού A. Σημειώνεται ότι όταν η εντροπία πληροφορίας (ή πληροφοριακή εντροπία) μειώνεται τότε η πυκνότητα πληροφορίας αυξάνεται, δηλαδή η περιγραφή γίνεται περισσότερο συμπαγής.

Η εντροπία βοηθά ώστε να προσδιορισθεί η κατάλληλη μεταβλητή που οδηγεί σε περισσότερο συμπαγές δέντρο.

Το κέρδος πληροφορίας G (S,A) ορίζεται ως

$$G(S, A) = E(S) - \sum_{u \in V(A)} |S_u| \cdot E(S_u) / |S|, \qquad (2.5)$$

όπου E(S) είναι η εντροπία πληροφορίας του υπό εξέταση κόμβου, A είναι μια ανεξάρτητη μεταβλητή, με τιμές V(A) με τις οποίες γίνεται ο επόμενος διαχωρισμός, u είναι μια από τις (δυνατές) τιμές A και S_u είναι το πλήθος των εγγραφών (με A=u) και $E(S_U)$ είναι η εντροπία πληροφορίας του υπό εξέταση κόμβου ως προς τη τιμή A=u.

Το ευρετήριο Gini γα κάθε κόμβο t , δίνετε από τον τύπο

$$GINI(t) = 1 - \sum_{j=1}^{c} [p(j / t)]^2 , \qquad (2.6)$$

όπου p (j / t) σχετική συχνότητα της κλάσης j στο κόμβο t (ποσοστό εγγραφών της κλάσης j στο κόμβο t) και c αριθμός κλάσεων. Το αναμενόμενο κέρδος σε πληροφορία μετά τον έλεγχο στο χαρακτηριστικό A ισούται με την πληροφορία που μας έλειπε πριν τον έλεγχο μείον τη αναμενόμενη πληροφορία που θα μας λείπει μετά τον έλεγχο.

Η ελάχιστη τιμή που μπορεί να πάρει είναι (0.0) , όταν όλες οι εγγραφές ανήκουν σε μια κλάση, ενώ η μέγιστη τιμή (1-1/c), όταν όλες οι εγγραφές είναι ομοιόμορφα κατανεμημένες στις κλάσεις.

Στους αλγορίθμους CART,SLIQ, και SPRINT χρησιμοποιείται ο παρακάτω τύπος

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i) ,$$

(2.7)

όπου n_i είναι ο αριθμός εγγραφών του παιδιού i και n είναι ο αριθμός εγγραφών του κόμβου p. Αυτός ο τύπος χρησιμοποιείται για την ποιότητα διαχωρισμού εγγραφών, όταν ένας κόμβος p διασπάται σε κ κόμβους.

Πλεονεκτήματα / Μειονεκτήματα Αλγόριθμου

Στα πλεονεκτήματα του αλγόριθμου είναι η μη παραμετρική προσέγγιση, δηλαδή δε στηρίζεται σε υπόθεση εκ των πρότερων γνώσεων σχετικά με τον τύπο της κατανομής πιθανότητας που ικανοποιεί η κλάση ή τα αλλά γνωρίσματα. Η κατασκευή βέλτιστου δένδρου απόφασης είναι ένα ¨NP-complete¨ πρόβλημα . Ένα ακόμα πλεονέκτημα όταν ένα δέντρο κατασκευαστεί, η ταξινόμηση νέων εγγράφων είναι πολύ γρήγορη της τάξεως O(h) και είναι εύκολα στην κατανόηση τους κυρίως στα μικρότερα δέντρα. Ακόμα δείχνει να έχει καλή συμπεριφορά στο θόρυβο δεδομένων και όταν υπάρχουν πλεονάζοντα γνωρίσματα δεν καταστρέφεται η κατασκευή.

Στα μειονεκτήματα του αλγόριθμου συγκαταλέγεται το ότι ο αλγόριθμος δε μπορεί να χειριστεί περίπλοκες σχέσεις μεταξύ γνωρισμάτων , χρησιμοποιεί απλά όρια απόφασης και αντιμετωπίζει προβλήματα όταν λείπουν πολλά δεδομένα.

2.2.3 Αλγόριθμοι Δένδρων Αποφάσεων

Στο εδάφιο αυτό παρουσιάζονται ορισμένοι αλγόριθμοι Δένδρων Αποφάσεων σε ψευδικωδική μορφή.

Αλγόριθμος C4.5

Στον αλγόριθμο C4.5 (Quinlan,1993) επιλέγεται ένας αρχικός κόμβος , ο οποίος θα σπάσει το αρχικό σύνολο εκπαίδευσης με βάση τη συνθήκη διάσπασης , έτσι θα βρεθεί ο κατάλληλος κόμβος ώστε να υπολογίζει όλα τα χαρακτηριστικά του συνόλου εκπαίδευσης και να επιδεχθεί το χαρακτηριστικό με τη καλύτερη τιμή. Αυτό θα επαναληφτεί αναδρομικά.

Algorithm C4.5

Input: an attribute-value dataset D

Tree={}

If D is "pure" OR other stopping criteria met then terminate

end if

for all attribute $a \in D$ **do**

Compute information theoretic criteria if we split on α

end for

$a_{best} = Best\ attribute\ according\ to\ above\ computed\ criteria$

Tree=Create a decision node that tests a_{best} in the root

$D_v = Induced\ sub-datasets\ from\ D\ based\ on\ a_{best}$

for all D_v **do**

$Tree_v = C4.5(D_v)$

Attach $Tree_v$ to the corresponding branch of Tree

end for

return Tree

2.2.4 Μάθηση Κανόνων Ταξινόμησης

Οι κανόνες ταξινόμησης θεωρούνται ως η πλέον δημοφιλής εναλλακτική ταξινόμηση στα δένδρα απόφασης, Οι κανόνες είναι από τις πιο εκφραστικές και κατανοητές αναπαραστάσεις για τον άνθρωπο.

Οι κυριότερες κατηγορίες κανόνων που συναντάμε είναι: (i) η προϋπόθεση κανόνα (antecedent) και (ii) το συμπέρασμα κανόνα (consequent). Στον κανόνα με προϋπόθεση υπάρχει ένα σύνολο ελέγχων ,το οποίο είναι όμοιο με τους ελέγχους στους κόμβους ενός δένδρου απόφασης. Οι έλεγχοι συνήθως χρησιμοποιούν τις λογικές συζεύξεις με συχνότερη να εμφανίζεται τη AND. Στον κανόνα με συμπέρασμα γίνεται η εκχώρηση ταξινόμησης, συνόλου ταξινομήσεως ή κατανομής πιθανότητας. Όταν συναντούμε ανεξάρτητους κανόνες χρησιμοποιούνται οι λογικές διαζεύξεις, κατά κύριο λόγο η OR, ώστε να ατυχούμε τη δημιουργία κανόνα. Το πρόβλημα που

συναντάται είναι ότι κάποιες φόρες οι υποδείξεις των κανόνων είναι διαφορετικές για το ίδιο παράδειγμα.

Για να μετατρέψουμε ένα δένδρο σε ένα σύνολο κανόνων θα πρέπει αρχικά να θέσουμε έναν κανόνα για κάθε φύλλο. Η προϋπόθεση περιέχει μια συνθήκη για κάθε κόμβο που συναντάται από τη ριζά ως το φύλλο , ενώ ως συμπέρασμα ορίζεται η τάξη εκχώρησης. Οι παραγόμενοι κανόνες είναι σαφείς και ορίζονται μονοσήμαντα , και αλλάζοντας τη σειρά που θα εκτελεστούν δεν θα αλλοιωθεί το αποτέλεσμα. Όταν υπάρξουν πολλοί κανόνες θα χρειαστεί «κλάδεμα» για την απομάκρυνση των περιττών ελέγχων και κανόνων.

Τα μειονεκτήματα στην ερμηνεία κανόνων είναι όταν έχουμε δυο ή περισσότερους κανόνες που δίνουν αντικρουόμενες υποδείξεις, έχουμε αδυναμία στο να εξάγουμε ένα συμπέρασμα. Υπάρχει μια ειδική περίπτωση στη δυαδική κλάση , όπου αν έχουμε ένα γεγονός το ίδιο δεν ανήκει στη μια κλάση τότε κατά ανάγκη θα εκχωρηθεί στην άλλη. Σε αυτήν την περίπτωση χρησιμοποιείται ένα τέχνασμα, δημιουργώντας κανόνα μόνο μια κλάσης, και κάνοντας χρήση της άλλης κλάσης ως προεπιλεγμένης σε κάθε άλλη περίπτωση. Ο κανόνας σε αυτή την περίπτωση μπορεί να γραφτεί σε διάζευξη διαδοχικών συζεύξεων, δηλαδή σε διαζευκτικής κανονικοποιημένη μορφή.

Οι πιο γνωστοί αλγόριθμοι αυτής της κατηγορίας είναι ο ZeroR και ο PART (Witten I.H., Frank E., Hall M.A.,2011). Συγκεκριμένα ο PART χρησιμοποιεί επιμέρους δένδρα απόφασης για να δημιουργήσει τον κατάλογο αποφάσεων που εμφανίζεται στη έξοδο , αλλά μόνο αυτή η τελική λίστα είναι αυτή που χρησιμοποιείται για να γίνει η ταξινόμηση. Έτσι δεν υπάρχει καμία ανάγκη να εξεταστούν τα επιμέρους δένδρα που παράγονται κατά τη διαδικασία της μάθησης ,απλά χρησιμοποιήστε τη λίστα των κανόνων που παρουσιάζεται στο εργαλείο λογισμικού WEKA (Witten et al., 1994). Το PART μπορεί να χρησιμοποιηθεί για επιλεγμένα χαρακτηριστικά σε συνδυασμό με ένα περιτύλιγμα αξιολόγησης υποσυνόλων στο πακέτο επιλογής χαρακτηριστικών (η εναλλακτικά η αξιολόγηση υποσύνολου ταξινομητή).

2.2.5 Μάθηση βασισμένη σε Στιγμιότυπα

Σε αντίθεση με τις άλλες μεθόδους μηχανικής μάθησης , οι όποιες κωδικοποιούν τα παραδείγματα εκπαίδευσης σε μια συμπαγή περιγραφή , στη μάθηση βασισμένη σε στιγμιότυπα ,

τα δεδομένα εκπαίδευσης μπορούν να διατηρηθούν αυτούσια. Μερικές φορές χρησιμοποιούμε τις μεθόδους παραμετρικής μάθησης όπου η ομαδοποίηση χωρίς επίβλεψη με χρήση μειγμάτων Gauss θεωρεί ότι τα δεδομένα εξηγούνται από το άθροισμα ενός σταθερού αριθμού κατανομών Gauss. Το πρόβλημα εμφανίζεται όταν έχουμε λίγα δεδομένα και πρέπει να διατηρείται η ίδια πολυπλοκότητα με ένα μεγαλύτερο σύνολο δεδομένων.

Οι μέθοδοι μη παραμετρικής μάθησης (non parametric learning) επιτρέπουν να μεγαλώνει η πολυπλοκότητα της υπόθεσης ανάλογα με τα δεδομένα. Αν έχουμε μεγάλο πλήθος δεδομένων, τότε έχουμε και αναλογική αυξανομένη πολυπλοκότητα. Δυο βασικές οικογένειες μεθόδων μη παραμετρικής μάθησης υπάρχουν βασισμένες σε στιγμιότυπα ή μάθησης βασισμένης στην μνήμη, οι όποιες δομούν τις υποθέσεις τους από τα ίδια τα στιγμιότυπα εκπαίδευσης.

Στην ταξινόμηση αλγορίθμων μάθηση βασισμένης σε στιγμιότυπα ανήκουν οι ακόλουθοι αλγόριθμοι: Πλησιέστερου γείτονα (Shekhar S.,Xiong H.2008), Ευκλείδειας Απόστασης (Euclidean Distance Algorithm) (Breu H.,Gil J.;Kirkpatrick D.;Werman M,1995), Μοντέλα Πυρήνων (kernel model) (Cortes C., Vapnik V.,1995), Μάθηση κατά Bayes (Russell , Norvig 2011), Παρεμβολή και Παλινδρόμηση (Berk R,2004), όπως γραμμική παλινδρόμηση. Λογική παλινδρόμηση, μέθοδος ελαχίστων τετραγώνων (least square methods).

Μοντέλα Πυρήνων

Στα μοντέλα πυρήνων (kernel model) θεωρούμε ότι κάθε στιγμιότυπο εκπαίδευσης δημιουργεί από μόνο του μια μικρή συνάρτηση πυκνότητας, που ονομάζετε συνάρτηση πυρήνα (Crammer, Koby et al. ,2001) και δίνεται από τον τύπο

$$P(x) = \frac{1}{N}\sum_{i=1}^{N} K(x, x_i) \quad (2.8)$$

Το x_i είναι στιγμιότυπο εκπαίδευσης και θα δημιουργήσει έναν πυρήνα $K(x,x_i)$.

Η επιβλεπόμενη μάθηση με πυρήνες γίνεται με τη λήψη ενός σταθμισμένου συνδυασμού όλων των προβλέψεων από το στιγμιότυπα εκπαίδευσης. Η τιμή του πυρήνα K(x,xi) δίνεται από το βάρος του i-οστού στιγμιότυπου για ένα σημείο ερωτήματος χ που δίνεται από τη τιμή του πυρήνα.

Για μια διακριτή πρόβλεψη μπορούμε να πάρουμε μια σταθμισμένη ψήφο, ενώ για μια συνεχή πρόβλεψη μπορούμε να πάρουμε το σταθμισμένο μέσο όρο ή μια σταθμισμένη γραμμική παλινδρόμηση. Οι προβλέψεις τέτοιου είδους απαιτούν εξέταση όλων των στιγμιότυπων εκπαίδευσης. Μπορούμε να συνδυάσουμε επίσης τη μέθοδο αυτή με μεθόδους δεικτοδότησης των πλησιέστερων γειτόνων έτσι ώστε να κάνουμε σταθμισμένες προβλέψεις μόνο από γειτονικά στιγμιότυπα.

2.2.6 Μάθηση κατά Bayes

Στη μάθηση κατά Bayes υπολογίζεται η πιθανότητα κάθε υπόθεσης, με βάση τα δεδομένα και κάνει προβλέψεις με χρήση όλων των υποθέσεων συμφώνα με τις πιθανότητες τους , αντί με την χρήση μόνο της βέλτιστης υπόθεσης.

Πιο συγκεκριμένα, κάθε παράδειγμα εκπαίδευσης μπορεί σταδιακά να μειώσει ή να αύξηση την πιθανότητα να είναι σωστή μια υπόθεση. Όμως αυτό δεν είναι εφικτό γιατί είναι απαραίτητο να γνωρίζουμε τις πιθανότητες πολλών τιμών. Αυτή η δυσκολία εφαρμογής έχει δώσει μεγάλη πρακτική αξία, στον απλό ταξινομητή Bayes, στον οποίο γίνεται η παραδοχή ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους.

Το πιο συνηθισμένο μοντέλο Bayes που χρησιμοποιείται είναι το αποκαλούμενο ¨απλοϊκό¨ μοντέλο Bayes. Το επίθετο απλοϊκό δόθηκε επειδή θεωρεί τα χαρακτηριστικά είναι υπό συνθήκη ανεξάρτητα μεταξύ τους, με δεδομένη κατηγορία.

Αν θεωρήσουμε ένα δυτιμες μεταβλητές μετά από την εκπαίδευση του μοντέλου κατ΄ αυτόν τον τρόπο, μπορεί να χρησιμοποιηθεί για την ταξινόμηση νέων παραδειγμάτων για τα οποία η μεταβλητή κατηγορίας C δεν έχει παρατηρηθεί. Με παρατηρηθείσες τιμές χαρακτηριστικών x1,x2,…..,xn, η πιθανότητα κάθε κατηγορίας δίνεται από τον τύπο

$$P(C \mid x_1, x_2, \ldots, x_n) = aP(C)\prod_j P(x_i / C) \qquad (2.9)$$

όπου C είναι η ριζά (μεταβλητή κατηγορίας) και είναι x_i είναι τα φύλλα (μεταβλητές χαρακτηριστικών).

Η μέθοδος μαθαίνει ικανοποιητικά καλά αλλά όχι τόσο καλά όσο η μάθηση δένδρου αποφάσεων, αυτό πιθανώς συμβαίνει επειδή η αληθής υπόθεση δεν αναπαριστάνεται ακριβώς με την χρήση αυτή της ταξινόμησης. Η μάθηση απλοϊκού μοντέλου Bayes μπορεί να κλιμακωθεί καλά σε μεγάλα προβλήματα , δηλαδή με n Boolean χαρακτηριστικά υπάρχουν (2n+1) παράμετροι και δεν απαιτείται αναζήτηση για να βρούμε την υπόθεση απλοϊκού μοντέλου Bayes με μέγιστη πιθανοφάνεια. Σημειώνεται ότι τα δεδομένα δεν επηρεάζονται από δεδομένα με θόρυβο.

Το σφάλμα βρίσκεται από τη διαφορά μεταξύ της πραγματικής τιμής και της προβλεπόμενης τιμής, όπου λέγεται άθροισμα τετραγωνικών σφαλμάτων. Αυτή η ποσότητα ελαχιστοποιείται από την συνήθη διαδικασία γραμμικής παλινδρόμησης. Έτσι συμπεραίνουμε ότι η ελαχιστοποίηση του αθροίσματος των τετραγώνων σφαλμάτων δίνει το μοντέλο ευθείας γραμμής μεγίστης πιθανοφάνειας, εφόσον τα δεδομένα δημιουργούνται με θόρυβο Gauss σταθερής διακύμανσης. (Russell , Norvig 2011)

2.2.7 Παρεμβολή και Παλινδρόμηση

Άλλη μια τεχνική μηχανικής μάθησης με επίβλεψη είναι η παλινδρόμηση, όπου είναι η διαδικασία προσδιορισμού της σχέσης μιας εξαρτημένης μεταβλητής y με μια ή περισσότερες άλλες ανεξάρτητες μεταβλητές x1,x2,…..,xn. Χρησιμοποιείται με σκοπό την εκχώρηση δεδομένων στις μεταβλητές πρόβλεψης, αν οι μεταβλητές είναι συνεχείς. Επίσης η ανάλυση της παλινδρόμησης έχει καθοριστική σημασία γιατί μας δείχνει στατιστικά την εκτίμηση των συσχετίσεων, δηλαδή ο βαθμός εμπιστοσύνης είναι κοντά στην εκτίμηση που έχει γίνει.

Η παρεμβολή παρουσιάζεται σε γραμμικά και μη γραμμικά μοντέλα, στη Γραμμική Παλινδρόμηση (linear regression) αντίστοιχα και στη Λογιστική Παλινδρόμηση (logistic regression). Όταν χρησιμοποιείται για παρεμβολή σημείων σε ενδιάμεσα τμήματα μπορεί να γίνει και η κατηγοριοποίηση. Η συνάρτηση παλινδρόμησης προβλέπει τη συνάρτηση χ με μεταβλητές x1,x2,…..,xn στην κλάση με τιμή y.

Το πιο γνωστό μοντέλο είναι το γραμμικό, όπου η αναμενόμενη τιμή της εξόδου μοντελοποιείται με μια γραμμική συνάρτηση ή άθροισμα με βάρη (weighted sum) των παραμέτρων εισόδου. Αυτό δίνεται από τον τύπο

$$y_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + \ldots + b_n x_{nj},\ j = 1, 2, \ldots, m,$$

(2.10)

όπου m είναι ο αριθμός των παραδειγμάτων εκπαίδευσης και b_i για i=1,…..,n για υπολογισμό των συντελεστών.

Μια από τις πιο διαδεδομένες μεθόδους λύσης είναι η μέθοδο των ¨ελαχίστων τετραγώνων¨ (least squares) που ελαχιστοποιεί το σφάλμα μεταξύ της εκτιμώμενης συνάρτησης και των πραγματικών δεδομένων. Με αυτή την μέθοδο επίλυσης προσπαθούμε να προσδιορίσουμε τη μορφή της άγνωστης σχέσης , στην οποία ταιριάζουν καλυτέρα τα πειραματικά μας δεδομένα, ελέγχοντας μια σειρά γνωστών σχέσεων.

Για τα μη γραμμικά μοντέλα παρεμβολής χρησιμοποιείται η Λογιστική Παλινδρόμηση, όπου παρατηρείται ότι τα σφάλματα δεν υπακούν στην κανονική κατανομή και η μεταβλητή απόκρισης είναι διακριτή. Σε αυτή την περίπτωση η τιμή εξόδου δεν είναι σταθμισμένο αθροισμάτων παραμέτρων εισόδου αλλά συνδέεται με αυτά με πιο πολύπλοκο τρόπο.

Σκοπός της λογιστικής παλινδρόμησης είναι η πρόβλεψη της απουσίας ή της παρουσίας ενός χαρακτηριστικού. Υπάρχουν κάποιες περιπτώσεις όπου μη γραμμικά προβλήματα μπορούν να μετατραπούν σε γραμμικά με κάποιο τρόπο κατάλληλο μετασχηματισμό, για να μπορέσουν να επιλυθούν με την μέθοδο ελαχίστων τετραγώνων.

2.3 Τεχνητά Νευρωνικά Δίκτυα

Μια εναλλακτική τεχνική μηχανικής μάθησης είναι τα τεχνητά νευρωνικά δίκτυα (ΤΝΔ)(artificial neural networks), αυτά είναι συστήματα επεξεργασίας δεδομένων που εξαρτώνται από ένα πλήθος τεχνητών νευρώνων οργανωμένων σε δομές παρόμοιες με αυτές του ανθρώπινου εγκεφάλου.

Το ΤΝΔ είναι ιδιαίτερα δημοφιλή σε προβλήματα που δεν μπορούν να γίνουν προβλέψεις, όπως προβλήματα σε πολλές ανθρώπινες δραστηριότητες που σχετίζονται με την ταξινόμηση (classification), αναγνώριση (recognition), αποτίμηση (assessment) και πρόβλεψη (prediction).

Τα τεχνητά νευρωνικά δίκτυα (artificial neural networks) είναι συστήματα επεξεργασίας δεδομένων που εξαρτώνται από ένα πλήθος τεχνητών νευρώνων οργανωμένων σε δομές παρόμοιες με αυτές του ανθρώπινου εγκεφάλου. Έτσι οργανώνονται σε επίπεδα , όπου το πρώτο είναι το επίπεδο εισόδου (input layer), όπου χρησιμοποιείται για να εισάγουμε τα δεδομένα μας.

Το ενδιάμεσο επίπεδο ονομάζεται κρυφό επίπεδο και μπορεί να απαρτίζεται από ένα ή και παραπάνω κρυφά επίπεδα. Τέλος υπάρχει και το επίπεδο εξόδου (output layer). Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδο τους με το αντίστοιχο συνοπτικό βάρος και υπολογίζουν το ολικό άθροισμα. Έτσι το άθροισμα τροφοδοτεί τη συνάρτηση ενεργοποίησης, την οποία υλοποιεί κάθε κόμβος. Κάθε φορά λαμβάνεται η τιμή της συνάρτησης και η έξοδος του νευρώνα για τις τρέχουσες τιμές. Η συνάρτηση αυτή δίνεται από τον τύπο

$$y_k = \varphi(\sum_{i=0}^{N} x_{ki} w_{ki}), \tag{2.11}$$

Όπου xki είναι η i-οστη εισοδος του κ νευρώνα, w_{ki} το i-οστό συνοπτικό βάρος του κ νευρώνα , φ η συνάρτηση ενεργοποίησης του νευρωνικού δικτύου και y_k η έξοδος του κ νευρώνα. Στον κ οστό νευρώνα έχουμε το συνοπτικό βάρος w_{ki} όπου καλείται πόλωση ή κατώφλι. Αν το συνολικό άθροισμα από τις υπόλοιπες εισόδους του νευρώνα είναι μεγαλύτερο από την αρχική του τιμή που θα πάρει και θα είναι 1, τότε ο νευρώνας ενεργοποιείται. Αν είναι μικρότερο από την τιμή της εισόδου , τότε ο νευρώνας παραμένει ανενεργός.

Οι νευρώνες μπορεί να είναι **πλήρως συνδεδεμένοι**, δηλαδή είναι συνδεδεμένοι με τους υπόλοιπους νευρώνες αλλιώς είναι **μερικώς συνδεδεμένοι** (partially connected). Όταν δεν υπάρχουν συνδέσεις μεταξύ νευρώνων προηγούμενου επιπέδου τα δίκτυα χαρακτηρίζονται **πρόσθια τροφοδοτούμενα**, διαφορετικά αν έχουμε το αντίθετο ή καθώς και στην περίπτωση συνδέσεων μεταξύ των νευρώνων ίδιου επιπέδου, χαρακτηρίζονται με **ανατροφοδότηση**(feedback).

Τα τεχνικά νευρωνικά δίκτυα (ΤΝΔ) εκτελούν δύο βασικές λειτουργιές. Αρχικά κατά την διαδικασία της εκπαίδευσης, χρησιμοποιείται η λειτουργία της μάθησης όπου τροποποιεί την τιμή των βαρών του δικτύου , ώστε αν δώσουμε ένα συγκεκριμένο διάνυσμα εισόδου να πάρουμε ένα συγκεκριμένο διάνυσμα εξόδου. Άλλη μια λειτουργία είναι η *ανάκληση* (recall) όπου υπολογίζει ένα διάνυσμα εξόδου για ένα συγκεκριμένο διάνυσμα εισόδου και τιμές βαρών.

Τα ΤΝΔ μπορούν να εμφανίσουν φαινόμενα υποταιριάσματος ή *ατελούς μάθησης (underfitting) ή υπερταιριάσμος (overfitting)*. Αν το δίκτυο αυτό δεν είναι αρκετά περίπλοκο μπορεί να αποτύχει να μοντελοποιήσει τα δεδομένα εκπαίδευσης οπότε μπορεί να οδηγηθεί στο φαινόμενο του υποταιριάσματος. Αντίθετα αν έχουμε ένα πολύπλοκο νευρωνικό δίκτυο μπορεί να

μοντελοποίηση υπερβολικά τα δεδομένα εκπαίδευσης καθώς και τον θόρυβο που μπορεί να υπάρχει μέσα σε αυτά. Στην περίπτωση αυτή παρουσιάζεται το φαινόμενο υπερταιριάσματος όπου δίνει σωστή πρόβλεψη για τα δεδομένα εκπαίδευσης αλλά παράγει λάθος προβλέψεις για τις επόμενα δεδομένα εισόδου.

Τα ΤΝΔ είναι ιδιαίτερα δημοφιλή σε προβλήματα που δεν μπορούν να γίνουν προβλέψεις, όπως προβλήματα σε πολλές ανθρώπινες δραστηριότητες που σχετίζονται με την ταξινόμηση (classification) , αναγνώριση (recognition), αποτίμηση (assessment) και πρόβλεψη (prediction).

Στα νευρωνικά δίκτυα παρουσιάζονται τέσσερις ιδιότητες:

i. εκμάθηση μέσω παραδειγμάτων,
ii. ικανότητα τους για την αναγνώριση προτύπων,
iii. ανοχή σε σφάλματα,
iv. δυνατότητα θεώρησης ως κατανεμημένη μνήμη και ως μνήμη συσχέτιση.

Ένα σημαντικό πλεονέκτημα του είναι η ανοχή που παρουσιάζουν σε δεδομένα εκπαίδευσης με θόρυβο, δηλαδή σε δεδομένα όπου έχουν λανθασμένες τιμές ή λανθασμένες καταχωρήσεις. Αντιθέτως όμως δε μπορούν να εξηγήσουν ποιοτικά τη γνώση που μοντελοποιούν.

2.4 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine-SVM) στηρίζονται στη θεωρία στατιστικής μάθησης και στα νευρωνικά δίκτυα τύπου Perceptron (Cortes C.,1995). Τα τελευταία χρόνια θεωρείται η πλέον διαδεδομένη μέθοδος για γραμμικές ή μη μεθόδους παρεμβολής και ταξινόμησης. Οι Μηχανές διανυσμάτων υποστήριξης προσπαθούν να βρουν μια υπερεπιφανεια (hypersurface) ώστε να μπορέσει να διαχωρίσει στο χώρο τα θετικά και τα αρνητικά παραδείγματα. Η επιλογή της επιφάνειας γίνετε με τέτοιον τρόπο ώστε να απέχει όσο το δυνατόν περισσότερο από τα κοντινότερα θετικά και αρνητικά παραδείγματα (maximum margin hypersurface). Έτσι μπορεί να ταξινομήσει περιπτώσεις που είναι παρόμοιες και όχι πανομοιότυπες με κάποιο παράδειγμα εκπαίδευσης με αποτέλεσμα μια αριθμητική τιμή στο διάστημα [-1, +1].

Ένα βασικό πλεονέκτημα που έχουν οι μηχανές διανυσμάτων έναντι στα νευρωνικά δίκτυα τύπου Perceptron είναι ότι μπορούν να παράγουν πιο σύνθετες υπερεπιφανειες, ενσωματώνοντας

μετασχηματισμούς και συνδυασμούς των αρχικών μεταβλητών ανάλογα με το πρόβλημα και ξεπερνώντας προβλήματα τοπικών ελαχίστων και διασποράς των λύσεων στο χώρο αναζήτησης. Γι αυτό χρησιμοποιούν ένα πεπερασμένο υποσύνολο που καλείται διάνυσμα υποστήριξης (support vectors) και συναρτήσεις πυρήνα (kernel functions) στο σύνολο εκπαίδευσης, ώστε να μετασχηματίσουν τον αρχικό χώρο υποθέσεων για να βρουν τη βέλτιστη μη γραμμική υπερεπιφανεια που ελαχιστοποιεί το σφάλμα ταξινόμησης.

2.5 Αλγόριθμοι Μάθησης Ομαδικών Ταξινομητών

Το πρόβλημα των ανομοιογενών τάξεων να μπορεί να επιλυθεί χρησιμοποιώντας τους αλγόριθμους μάθησης συνόλων Boosting (AdaBoost και τους τροποποιημένους αλγορίθμους AdamBoost.M1 και AdamBoost.M2) (Schapire, 1990; Freund and Schapire, 1997; Schapire and Siner, 1999) και Bagging (και τους τροποποιημένους αλγορίθμους κολλημένους μικρούς ψήφους με σημαντικό δείγμα) (pasting small votes with importance sampling algorithm) (Breiman, 1999).

Ο αλγόριθμος Bagging προέρχεται από ένα συνδυασμό δυο παλαιοτέρων τεχνικών του Bootstap και Aggregation, και οι βασικές λειτουργίες που κάνει είναι δειγματοληψία με επανένταξη (sampling with replacement),να κατασκευάζει ταξινομητή για κάθε δείγμα δεδομένων και κάθε δείγμα από αυτά να έχει πιθανότητα επιλογής [1-$\left(1-\frac{1}{n}\right)^{n}$]. Στον αλγόριθμο Boosting δεν δίνεται το ίδιο βάρος σε όλους τους ταξινομητές, αλλά προτεραιότητα έχει η ακρίβεια του. Η κεντρική ιδέα του αλγορίθμου είναι ότι έχουμε την δυνατότητα να εξετάσουμε τη πιθανότητα επιλογής των ταξινομήσεων στο σύνολο εκπαίδευσης.

Στο εδάφιο αυτό παρουσιάζονται συνοπτικά αλγόριθμοι μάθησης συνόλων ταξινομητών με ψευδοκωδικη μορφή. Συγκεκριμένα παρουσιάζονται οι αλγόριθμοι AdaBoost, Bagging και Ivotes σε ψευδοκωδικη μορφή:

Αλγόριθμος AdaBoost

Είσοδος: αριθμός επαναλήψεων T; συντελεστής ασθενούς μάθησης I (weak learner); σύνολο δοκιμών (training set) $S = \{x_i, y_i\}, i = 1, 2, \ldots, N;\ y_i \ [-1,\ +1]$.

Έξοδος: ταξινομητής ενίσχυσης (boosted classifier) $H(x) = \text{sign}[\sum_{t=1}^{T} a_t * h_t(x)\ ]$, όπου h_t και a_t είναι αντίστοιχα ειδικοί ταξινομητές (induced classifiers) και τα καθορισμένα βάρη τους.

Υπολογιστική Διαδικασία:

$D_1(i) \leftarrow 1/N$, for $i = 1, 2, \ldots, N$
For t = 1 to T do
$h_t \leftarrow I(S, D_t)$
$\varepsilon_t \leftarrow D_t(i) \ . \ [h_t(x_i) \neq y_i]$
if $(\varepsilon_t > 0.5)$ then
$T \leftarrow (t-1)$
return
endif
$a_t = \ln[(1-\varepsilon_r) \ / \ \varepsilon_r] \ / \ 2$
$z = -\ a_n h_t(x_i) \ . y_i$
$D_{t+1}(i) = D_t(i) \ . \ e \uparrow z$, for $i = 1, 2, \ldots, N$
Normalize D_{i+1} to be a proper distribution
endfor

Αλγόριθμος LogitBoost

Ο αλγόριθμος LogitBoost βρίσκεται στη κατηγορία Boosting στους αλγορίθμους ταξινόμησης ο οποίος δημιουργεί ένα εναλλακτικό δένδρο απόφασης (REF).

Η βασική ιδέα είναι να συνδυαστεί ένα σύνολο αδυναμιών κατηγοριοποιήσεων προκειμένου να σχηματίσουν να ένα δυνατό κατηγοριοποιητή ο οποίος θα έχει πολύ καλύτερα αποτελέσματα από την τυχαία πρόβλεψη.

Algorithm LogitBoost

Input: Instance distribution D;

Base learning algorithm L;

Number of learning rounds T.

Process:

$D_1 = D.$ %*Initialize distribution*

for $t = 1,\ldots,T$:

$h_t = L(D_t);$ %*Train a weak learner from distribution* D_t

$e_t = \Pr_{x\sim D_t,y}\, I[h_t(x) \neq y];$ %*Measure the error of* h_t

$D_{t+1} = AdjustDistribution(D_t, e_t)$

end

Output: $H(x) = CombineOutputs(\{h_t(x)\})$

Αλγόριθμος Bagging

Είσοδος : αριθμός επαναλήψεων T; συντελεστής ασθενούς μάθησης I(weak learner); μέγεθος αυτοδυναμίας (bootstrap) n; σύνολο δοκιμών (training set) S.

Έξοδος: ταξινομητής 'σακκουλιάσματος' (bagged classifier)

$H(x) = \text{sign}\left[\sum_{t=1}^{T} h_t(x)\right]$, όπου $h_t \in [-1, +1]$, είναι ειδικοί ταξινομητές (induced classifiers).

Υπολογιστική Διαδικασία:

for $t = 1$ to T do
$S_t \leftarrow$ Random Sample Replacement (n, S)
$h_t \leftarrow I(S_t)$
endfor

Αλγόριθμος Ivotes

Είσοδος : αριθμός επαναλήψεων T; συντελεστής ασθενούς μάθησης I (weak learner); μέγεθος αυτοδυναμίας (bootstrap) n; σύνολο δοκιμών (training set) S.

Έξοδος: ταξινομητής 'σακκουλιάσματος' (bagged classifier)

$H(x) = \text{sign}[\sum_{t=1}^{T} h_t(x)]$, όπου $h_t \in [-1, +1]$, είναι ειδικοί ταξινομητές (induced classifiers).

Υπολογιστική Διαδικασία:

$e_{NEW} \leftarrow 0.5$
repeat
$e_{OLD} \leftarrow e_{NEW}$
$St \leftarrow 0.0$
while size$(S_t) < n$ do{importance sampling}
$x \leftarrow$ RandomInstance (S)
if x misclassified by out – of – bag classifier then
$S_t \leftarrow S_t\{x\}$
else
$S_t \leftarrow S_t\{x\}$ with probability $e_{OLD} / (1 - e_{OLD})$
endif
endwhile
$h_t \leftarrow I(S_t)$
eNEW $\leftarrow$ error of out – of – bag classifier
until $e_{NEW} > e_{OLD}$

3 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ ΔΙΑΧΕΙΡΙΣΗΣ ΑΝΟΜΟΙΟΓΕΝΩΝ ΔΕΔΟΜΕΝΩΝ

3.1 Εισαγωγικές Παρατηρήσεις

Τα ανομοιογενή δεδομένα (imbalanced data) είναι πολύ σημαντικά στην ¨εξόρυξη γνώσης¨ (data mining) και στα δεδομένα ταξινόμησης. Το πρόβλημα της εκμάθησης για ανομοιογενή σύνολο δεδομένων συμβαίνει όταν ο αριθμός των δειγμάτων σε μια κλάση είναι σημαντικά μεγαλύτερος από μια άλλη κλάση. Το πρόβλημα των ανομοιογενών δεδομένων συνδέεται συχνά με το ασύμμετρο κόστος των ακατάλληλα ταξινομημένων στοιχείων διαφόρων τάξεων (Japkowicz and Stephen, 2002; Fawcett and Provost, 1997).

Για παράδειγμα σύνολα δεδομένων που εμπεριέχουν ανομοιογενή δεδομένα περιλαμβάνονται σε δεδομένα ασθενών καρκίνων και ίασης τους στην ιατρική επιστήμη, σε περίπτωση μη λειτουργίας επικοινωνιακών εξοπλισμών , στην εξαπάτηση στις αναλήψεις πιστωτικών καρτών, καθώς και σε πληθώρα δεδομένων από δορυφορικές εικόνες για αναζήτηση πετρελαιοκηλίδων και άλλων σχετικών πληροφοριών.

Η Ανομοιογενής *Μάθηση* ή *΄Μη-ισορροπημένη Μάθηση΄* (Imbalance Learning) αναφέρεται στη ικανότητα των ανομοιογενών δεδομένων να δεσμεύονται από την απόδοση των περισσοτέρων αλγορίθμων μάθησης, που προϋποθέτουν ομοιογενείς τάξεις κατανομών ή ίσα κόστη εσφαλμένης ταξινόμησης. Οι αλγόριθμοι αυτοί όταν παρουσιαστούν σε πολύπλοκα σύνολα ανομοιογενών δεδομένων , αποτυγχάνουν να παραστήσουν τα χαρακτηριστικά κατανομών και παρέχουν μη ικανοποιητικές ακρίβειες για τις τάξεις δεδομένων, σε πραγματικά προβλήματα το πρόβλημα ανομοιογενούς μάθησης αντιπροσωπεύει ένα επαναλαμβανόμενο σημαντικό πρόβλημα με πολυποίκιλες επιπτώσεις που πρέπει να διερευνηθούν.

Το πρόβλημα ανομοιογενούς μάθησης αποτελεί αντικείμενο έντονης έρευνας σε διεθνή περιοδικά του κλάδου, ειδικές εκδόσεις ερευνητικών περιοδικών και διεθνή συνέδρια επιστημονικών οργανισμών όπως Institute of Electrical and Electronics Engineers (IEEE), Association for Computing Machinery (ACM), Association of the Advancement of Artificial Intelligence (AAAI), ACM Special Interest Group on Knowledge Discovery and Data Mining Exploration (ACM SIGKDD Exploration) κ.α. (Japkowicz, 2000; Chawla et al., 2004).

Ένα σύνολο δεδομένων λέγεται ¨*ανομοιογενές*¨, όταν η κλάση των ενδιαφερόντων (class of interest)(ή κλάση της μειοψηφίας –minority class) είναι μικρότερη ή πιο σπάνια από την κανονική συμπεριφορά (κυρίαρχη κλάση- majority class).

Η επίδραση του προβλήματος στο σύνολο των ανομοιογενών δεδομένων στη ταξινόμηση βρίσκεται στο ότι οι ταξινομητές τείνουν να αποδείξουν ότι ο ανομοιογενής βαθμός της ακριβείας με κυρίαρχη κλάση (majority class) με πολύ μεγάλη ακρίβεια (περίπου >90%) και στην κλάση της μειονότητας (minority class) με δεδομένα ταξινόμησης που κυμαίνονται από 0-10%.

Η αποτίμηση μετρικών μεθόδων για ανομοιογενή δεδομένα συγκρίνει και αξιολογεί την απόδοση διαφορετικών αλγορίθμων ανομοιογενούς μάθησης. Σημειώνεται ότι κάθε σύνολο δεδομένων που διαθέτει μια άνιση κατανομή μεταξύ κλάσεων μπορεί να θεωρηθεί ως σύνολο ανομοιογενών δεδομένων (He and Shen, 2007; Kubat et al., 1998).

Τα ανομοιογενή δεδομένα διακρίνονται σε εσωτερικά (intrinsic), δηλαδή η ανομοιογένεια είναι άμεσο αποτέλεσμα της φύσης του διαστήματος δεδομένων, και εξωτερικά (extrinsic), δηλαδή η ανομοιογένεια είναι έμμεσο αποτέλεσμα της φύσης του διαστήματος δεδομένων. Η πολυπλοκότητα δεδομένων αναφέρεται γενικά σε έλλειψη αντιπροσωπευτικών δεδομένων, επικαλύψεων, μικρούς διαχωρισμούς κ.α.

Οι αλγόριθμοι ταξινόμησης (Classification algorithms) δεν παρουσιάζουν ικανοποιητική απόδοση , όταν τα δεδομένα είναι στραμμένα ως προς μια κλάση.

Η άμεση επίδραση του προβλήματος ανομοιογενούς μάθησης σε αλγορίθμους μάθησης σχετίζεται με αλγόριθμους μάθησης δένδρων αποφάσεων (Japkowwicz and Stephen, 2002; Weiss and Provost, 2003; Chawla, 2003). Τα δένδρα αποφάσεων χρησιμοποιούν αναδρομικούς αλγορίθμους αναζήτησης που χρησιμοποιούν σχήματα επιλογής χαρακτηριστικών για την επιλογή των καλύτερων χαρακτηριστικών ως κριτήριο διαχωρισμού σε κάθε κόμβο του δένδρου. Ένα φύλλο του δένδρου μπορεί να δημιουργηθεί για κάθε δυνατές τιμές που αντιστοιχούν στο χαρακτηριστικό διαχωρισμού (Quinlan, 1986; Mitchell, 1997).

Το πρόβλημα των ανομοιογενών δεδομένων στη περίπτωση περιοχών πραγματικού κόσμου αντιπροσωπεύει ένα σημαντικό επαναληπτικό πρόβλημα με πολυποίκιλες επιπτώσεις που απαιτούν προσεκτικές διερευνήσεις. Για την περίπτωση αυτή απαιτείται ένας ταξινομητής που παρέχει υψηλή ακρίβεια για την μειονοτική τάξη να διακυβεύεται η ακρίβεια της πλειοψηφικής τάξης.

Αναφέρεται ότι υπάρχει ισχυρή σύνδεση μεταξύ μάθησης με ευαισθησία κόστους και μάθησης με ανομοιογενή δεδομένα με συνέπεια οι θεωρητικές βάσεις και αλγόριθμοι των μεθόδων μάθησης με ευαισθησία κόστους να μπορούν να εφαρμόσουν σε προβλήματα μάθησης με ανομοιογενή δεδομένα. Σε ορισμένες περιπτώσεις η μάθηση με ευαισθησία κόστους αναφέρεται ότι είναι υπέρτερη από τις μεθόδους δειγματοληψίας (Chawla et al. 2004). Οι τεχνικές μάθησης με ευαισθησία κόστους θεωρούνται εναλλακτική προσέγγιση στις μεθόδους δειγματοληψίας για περιοχές μάθησης με ανομοιογενή δεδομένα.

Ένα σύνολο δεδομένων θεωρείται ότι είναι ένα ανομοιογενές (μη-ισορροπημένοο) αν οι κατηγορίες ταξινόμησης δεν αντιπροσωπεύονται με ισότιμο τρόπο. Οι τεχνικές μηχανικής μάθησης εφαρμόζονται σε πολύπλοκα προβλήματα ¨πραγματικού -κόσμου¨ που χαρακτηρίζονται από ανομοιογενή δεδομένα. Η αξιολόγηση της απόδοσης ενός ταξινομητή μπορεί να γίνει με προβλεπόμενη ακρίβεια που μπορεί να μην είναι κατάλληλη όταν τα δεδομένα είναι ανομοιογενή και όταν το κόστος διαφορετικών σφαλμάτων ποικίλει (Chawla et al. 2004).

3.2 Αλγοριθμικές Μέθοδοι Μάθησης σε Προβλήματα Ανομοιογενών Τάξεων

3.2.1 Εισαγωγικές Παρατηρήσεις

Τις τελευταίες δεκαετίες η έρευνα σε Μηχανική Μάθηση και Εξόρυξη Δεδομένων έχει αυξηθεί σημαντικά με την ανάπτυξη προηγμένων αλγορίθμων και τεχνικών. Παράλληλη ανάπτυξη είχε και η ανάπτυξη γλωσσών για χρήση σε μηχανική μάθηση και εξόρυξη δεδομένων, που παρέχει εναλλακτική αλληλεπίδραση σε αλγορίθμους και συστήματα τα οποία μπορούν να αυξήσουν σημαντικά τη χρησιμότητα τέτοιων συστημάτων.

Στη μηχανική μάθηση το σύνολο ταξινομητών αυξάνει την ακρίβεια των απλών ταξινομητών συνδέοντας μερικούς από αυτούς, αλλά καμία από αυτές τις τεχνικές μάθησης μπορεί να λύσει μόνη της το ανομοιογενές πρόβλημα και το σύνολο των αλγορίθμων μάθησης πρέπει να σχεδιάσει με ειδικό τρόπο.

Το πρόβλημα των ανομοιογενών κλάσεων είναι από τα πλέον σημαντικά στο πεδίο εξόρυξης δεδομένων και παρουσιάζεται σε ορισμένες εφαρμογές που εμφανίζουν το πρόβλημα

αυτό, όπως διάγνωση σφαλμάτων (Yang and Tang, 2009; Zhu and Song, 2010), ιατρικές διαγνώσεις (Mazurowski, 2008), ανίχνευση ανωμαλιών (Mazurowski, 2008), αναγνώριση προσώπων (Liu and Chen, 2005), δημιουργία ειδικών φακέλων ηλεκτρονικού ταχυδρομείου (e-mail folding) (Bermejo et al., 2011), ανίχνευση κηλίδων πετρελαίου(Kubat et al., 1998) κ.α.

Οι τεχνικές επίλυσης του προβλήματος ανομοιογενών τάξεων μπορούν να ταξινομηθούν στις ακόλουθες κατηγορίες:

i. τεχνικές επίπεδου αλγορίθμου (internal CI technique) με δημιουργία ή τροποποιήσεις υπαρχόντων αλγορίθμων (Quinlan, 1991; Wu and Chang, 2005)
ii. τεχνικές επίπεδου δεδομένων (external CI technique) με βήμα προδιαδικασιας, όπου η κατανομή δεδομένων επανα-ομογενοποιείται για να μειωθεί η επίδραση της κατανομής τάξεων στη διαδικασία μάθησης (Batista et al., 2004; Chawla et al., 2004)
iii. μέθοδοι ευαισθησίας κόστους, που παρουσιάζουν τις προηγούμενες δυο τεχνικές, για να συγχωνεύσουν διαφορετικά εσφαλμένα κόστη για κάθε τάξη στη διαδικασία μάθησης (Chawla et al., 2008; Freitas et al., 2007).

3.2.2 Αλγοριθμικές Τεχνικές για Ανομοιογενή Δεδομένα : Εισαγωγικές Παρατηρήσεις

Πολλά δύσκολα προβλήματα πραγματικού –κόσμου μηχανικής μάθησης σχετίζονται με ανομοιογενή δεδομένα μάθησης (imbalanced learning data), όπου τουλάχιστον μια τάξη δεν αντιπροσωπεύει κατάλληλα σχετικά με άλλες τάξεις π.χ. ανίχνευση διαφόρων απατών, έλεγχος και ιατρική δεδομένων διάγνωσης, κατηγοριοποίηση κείμενων κ.α. Το πρόβλημα των ανομοιογενών δεδομένων συνδέεται συχνά με το ασύμμετρο κόστος των ακατάλληλα ταξινομημένων στοιχείων διαφόρων τάξεων (Japkowicz and Stephen, 2002; Fawcett and Provost, 1997).

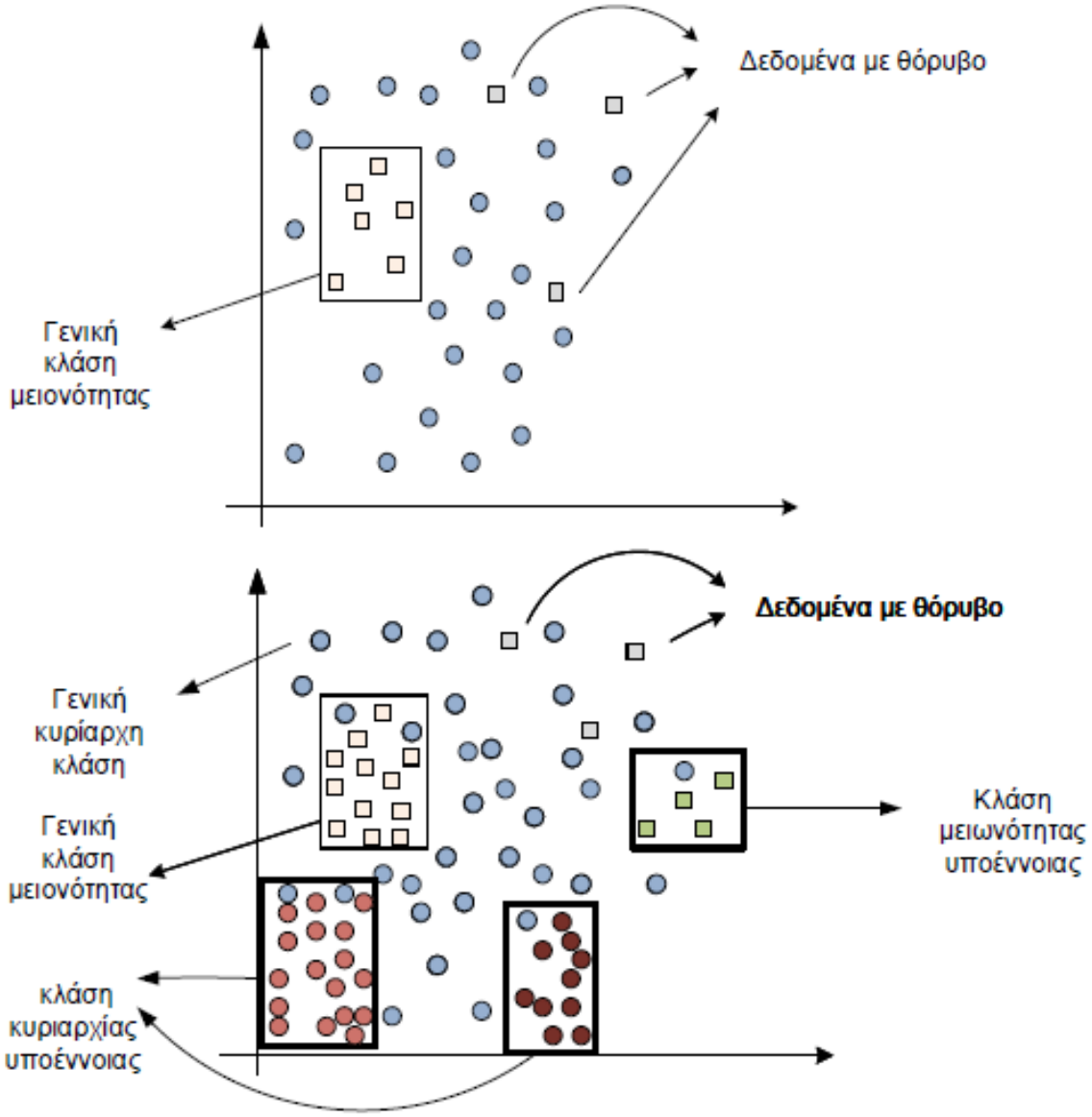
Οι μέθοδοι που χρησιμοποιούνται για την επίλυση του προβλήματος των ανομοιογενών δεδομένων περιλαμβάνουν τις ακόλουθες κατηγορίες:

i. Προ-επεξεργασία για τα δεδομένα
ii. Μετα-επεξεργασία για το μοντέλο μάθησης.

Η κατανόηση μεθόδων ανακάλυψης γνώσης και η ανάλυση πρωτογενών δεδομένων για υποστήριξη διαδικασιών για λήψη αποφάσεων αποτελούν σημαντικούς παράγοντες σχετικά με τη συνεχή επέκταση διαθεσιμότητας δεδομένων σε εφαρμογές πολλών μεγάλης τάξης πολύπλοκων δικτυακών συστημάτων, όπως διαδίκτυο, ασφάλεια, συστήματα παρακολούθησης και οικονομίας.

Το πρόβλημα μάθησης από ανομοιογενή δεδομένα αποτελεί ένα ενδιαφέρον ερευνητικό πεδίο για την ακαδημαϊκή κοινότητα και τη βιομηχανία, και αναφέρεται με την απόδοση των αλγορίθμων μάθησης παρουσία υπο-αντιπροσωπευτικών δεδομένων και αυστηρών τάξεων ασύμμετρων κατανομών (severe class distribution skews) . Εξαιτίας των πολύπλοκων χαρακτηριστικών των ανομοιογενών συνόλων δεδομένων, η μάθηση από τέτοια δεδομένα απαιτεί αρχές , αλγορίθμους και εργαλεία για τη μετατροπή τεράστιων ποσών πρωτογενών δεδομένων σε κατάλληλες παραστάσεις πληροφοριών και γνώσεων.

Το πρόβλημα της ανομοιογενούς μάθησης εμφανίζει την ικανότητα των ανομοιογενών (μη-ισορροπημένων) δεδομένων να συμβιβάζει την απόδοση των περισσότερων προτύπων αλγορίθμων μάθησης , που υποθέτουν ομοιογενείς τάξεις κατανομών ή ίσο κόστος εσφαλμένων ταξινομήσεων. Στην περίπτωση πολύπλοκων συνόλων ανομοιογενών δεδομένων, τέτοιοι αλγόριθμοι αποτυγχάνουν να παραστήσουν κατάλληλα τα κατανεμημένα χαρακτηριστικά των δεδομένων και παρέχουν ακατάλληλες ακρίβειες στις τάξεις δεδομένων.



Εικόνα 3.1: Αναπαράσταση δεδομένων μέσα σε μια κλάση και μέσα σε διάφορες κλάσεις. (He and Garcia,

Το πρόβλημα των ανομοιογενών δεδομένων στη περίπτωση περιοχών πραγματικού κόσμου αντιπροσωπεύει ένα σημαντικό επαναληπτικό πρόβλημα με πολυποίκιλες επιπτώσεις που απαιτούν αυξανόμενες εξερευνήσεις. Για την περίπτωση αυτή απαιτείται ένας ταξινομητής που παρέχει υψηλή ακρίβεια για τη μειονοτική τάξη χωρίς να διακυβεύεται η ακρίβεια της πλειοψηφικής τάξης.

Αναφέρεται ότι υπάρχει ισχυρή σύνδεση μεταξύ μάθησης με ευαισθησία κόστους και μάθησης με ανομοιογενή δεδομένα, με συνέπεια οι θεωρητικές βάσεις και αλγόριθμοι των μεθόδων μάθησης με ευαισθησία κόστους να μπορούν να εφαρμοστούν σε προβλήματα μάθησης με ανομοιογενή δεδομένα. Σε ορισμένες περιπτώσεις η μάθηση με ευαισθησία κόστους αναφέρεται ότι είναι υπέρτερη από τις μεθόδους δειγματοληψίας. Οι τεχνικές μάθησης με ευαισθησία κόστους θεωρούνται εναλλακτική προσέγγιση στις μεθόδους δειγματοληψίας για περιοχές μάθησης με ανομοιογενή δεδομένα.

Η ¨καμπύλη ROC¨ (Receiver Operating Characteristic) θεωρείται χρήσιμη επειδή προσφέρει μια ορατή παράσταση των σχετικών διαφορών μεταξύ πλεονεκτημάτων (true positives) και κόστους (false positives) των ταξινομήσεων σε σχέση με τις κατανομές δεδομένων (Fawcett, 2006; Bradley, 1997).

Η ανομοιογένεια (μη-ισορροπία) μπορεί να είναι ένα σύνολο κατανεμημένων τάξεων ή διαφορετικό κόστος σφάλματα ή παραδείγματα (Japkowicz, 2000; Chawla et al. 2003; Fem et al., 2004). Σαν προτεινόμενο μέτρο της απόδοσης μπορεί να χρησιμοποιηθεί η μέθοδος καμπύλων ROC (Fem et al., 2004).

Οι καμπύλες ROC μπορούν να χρησιμοποιηθούν για συνολική απόδοση ταξινομητών που αναφέρονται σε ένα φάσμα ανταλλαγών μεταξύ αληθών θετικών τιμών σφαλμάτων και μη-αληθών θετικών τιμών σφαλμάτων (Swets, 1988).

Μεθοδος Κατωφλίου (Threshold Method)

Ένα ¨κατώφλι¨ (threshold) λέγεται ότι είναι συνολικά βέλτιστο, αν ο αριθμός των μη ταξινομημένων στοιχείων είναι ελάχιστος. Η γραφική του παράσταση είναι ένα διτροπικό ιστόγραμμα (bimodal), δηλαδή ένα ιστόγραμμα με δυο προφανείς σχετικές λειτουργίες ή κορυφές δεδομένων. Για παράδειγμα, αν είχαμε μια απεικόνιση , μπορούσαμε να ξεχωρίσουμε ένα αντικείμενο με το περιβάλλον που βρισκόταν πίσω του.

Παραθέτοντας ένα παράδειγμα από την επεξεργασία εικόνας, εξετάζεται ποια προβλήματα παρουσιάζονται μέσω της μεθόδου αυτής.

Το πρόβλημα με τη μέθοδο κατωφλίου είναι ότι θεωρούμε σημαντική μόνο την ένταση της εικόνας και δεν δίνουμε σημασία στις σχέσεις μεταξύ των εικονοστοιχείων (pixels). Δεν γνωρίζουμε

αν τα εικονοστοιχεία που προσδιορίζονται από την διαδικασία κατωφλίου είναι συνεχόμενα. Μπορούμε εύκολα να περιβάλουμε εξωγενή εικονοστοιχεία που δεν αποτελούν μέρος της επιθυμητής περιοχής και μπορούμε να χάσουμε εξίσου εύκολα απομονωμένα εικονοστοιχεία εντός της περιοχής (ειδικά κοντά στα όρια της περιοχής). Αυτά τα αποτελέσματα δε είναι ικανοποιητικά καθώς ο θόρυβος αυξάνεται, απλά επειδή είναι πιο πιθανό ότι μια ένταση εικονοστοιχείου (pixels) δεν αντιπροσωπεύει την κανονική ένταση στην περιοχή.

Όταν χρησιμοποιείται η μέθοδος κατωφλίου, που συνήθως πρέπει να ασχοληθεί με το αντικείμενο, μερικές φορές χάνει πάρα πολύ από την περιοχή και παίρνει πολλά εικονοστοιχεία από το φόντο (για παράδειγμα, σκιές των αντικειμένων της εικόνας).

Ένα άλλο πρόβλημα με το συνολικό κατώφλι (global thresholding) είναι ότι οι αλλαγές στο φωτισμό σε όλο το περιβάλλον μπορεί να προκαλέσει κάποια μέρη να είναι πιο φωτεινά (το φως) και σε ορισμένες περιοχές πιο σκούρο (στη σκιά), με τρόπους που δεν έχουν καμιά σχέση με τα αντικείμενα της εικόνας. Μπορούμε να το αντιμετωπίσουμε εν μέρει, το πρόβλημα του άνισου φωτισμού με προσδιορισμό κατωφλίων σε τοπικό επίπεδο. Δηλαδή, αντί να έχει ένα ενιαίο συνολικό κατώφλι, επιτρέπεται στο κατώφλι να εξομαλύνει εκεί να υπάρχουν διαφορές από μόνο του στο συνολικό κατώφλι σε όλη την εικόνα.

Η μέθοδος QUEST (Watson and Pelli, 1983) είναι μια αποτελεσματική μέθοδος για τη μέτρηση κατωφλίων που βασίζεται σε τρία στάδια : (i) Προσδιορισμός των προηγούμενων γνώσεων και υποθέσεων, συμπεριλαμβανομένης της αρχικής συνάρτησης πυκνότητας πιθανότητας (p.d.f.) του κατωφλίου (δηλαδή σχετική πιθανότητα των διαφορετικών κατωφλίου του πληθυσμού). (ii) Μια μέθοδος για την επιλογή του ερέθισμα της έντασης οποιαδήποτε δοκιμής. (iii) Μια μέθοδος για την επιλογή του τελικού επιθυμητού κατωφλίου. (King-Smith et al, 1994)

Λόγω της έλλειψης αρκετών παραδοσιακών μεθόδων αυτόματου κατωφλίου (automatic threshold methods), προκειμένου να κατατμηθεί η διαβάθμιση της εικόνας καλύτερα δημιουργήθηκε μια βελτιωμένη μέθοδος. Η βελτιωμένη μέθοδος διπλής κατωφλίου (double-threshold method) συνδυάζεται με τη μέθοδος της μέγιστης διακύμανσης κλάσεων, τη μέθοδο εκτίμησης της περιοχής και μέθοδο διπλού κατωφλίου. Η μέθοδος αυτής μπορεί να επιλέξει αυτόματα δυο διαφορετικά κατώτατα όρια σε τμήματα εικόνων βαθμωτά. Η προσομοίωση σε υπολογιστή εκτελείται επι τις παραδοσιακές μεθόδους και αυτού του αλγορίθμου και αποδεικνύει ότι η μέθοδος αυτή μπορεί να πάρει ικανοποιητικό αποτέλεσμα. (Shen et al.,2004).

3.2.3 Τεχνικές Αλγορίθμου Επιπέδου: Μάθηση με Ευαισθησία Κόστους

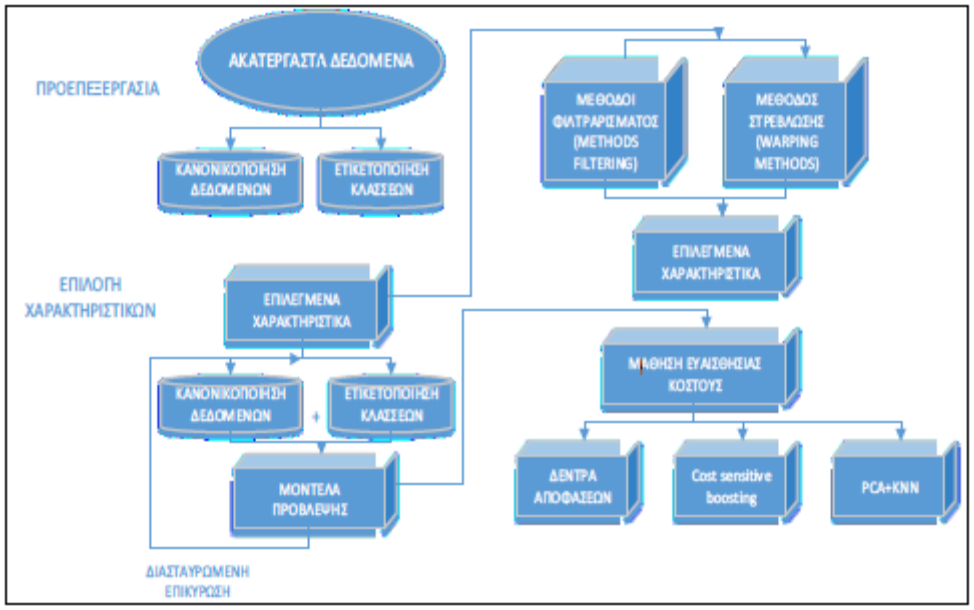
Η ***Μάθηση με Ευαισθησία Κόστους*** (Cost Sensitive Learning) είναι ένας τύπος μάθησης που αναφέρεται σε διάφορους τύπους κόστους με σκοπό την ελαχιστοποίηση του κόστους αυτού. Η βασική διαφορά μεταξύ μάθησης με ευαισθησία κόστους και μάθησης άνευ ευαισθησία κόστους είναι ότι η πρώτη διαχειρίζεται τις διάφορες τεχνικές μη-ταξινομήσεων διαφορετικά από την δεύτερη, με το να θεωρεί τέτοιες μη- ταξινομήσεις ότι δεν υπάρχουν, με στόχο να επιτύχει υψηλή ακρίβεια παραδειγμάτων ταξινόμησης σε ένα σύνολο γνωστών κλάσεων.

Η τάξη των ανομοιογενών (μη ισορροπημένων) συνόλων δεδομένων (imbalanced datasets) εμφανίζεται σε πολλές καθημερινές εφαρμογές, όπου η τάξη κατανομών είναι ανομοιογενείς σε υψηλό βαθμό και οι τεχνικές μάθησης με ευαισθησία κόστους αποτελούν μια κοινή προσέγγιση για την επίλυση τέτοιων προβλημάτων(Ling and Sheng, 2008).

3.2.4 Δομή Συστήματος Μάθησης με Ευαισθησία Κόστους

Η Μάθηση με ευαισθησία κόστους μπορεί να χωριστεί σε δυο κατηγορίες:

Η πρώτη κατηγορία περιλαμβάνει σχεδιασμό ταξινομητών που έχουν την ίδια ευαισθησία κόστους (άμεσες μέθοδοι), π.χ. ICET. (Turney, 1995), δένδρα αποφάσεων με ευαισθησία κόστους (Drummond-Holte, 2000; Ling et al. 2004). Η δεύτερη κατηγορία περιλαμβάνει το σχεδιασμό ενός ¨περιτυλίγματος¨ (wrapper), που μετατρέπει κάθε ταξινομητή άνευ-κόστους (cost insensitive) σε αντίστοιχους ταξινομητές με ευαισθησία κόστους. Η κατηγορία αυτή καλείται επίσης **μέθοδος μετά-μάθησης με ευαισθησία κόστους** (cost sensitive meta-learning method) και διαχωρίζεται σε ¨μέθοδο κατωφλίου¨ (thresholding) και ¨μέθοδο δειγματοληψίας¨ (sampling).



Εικόνα 3.2: Διαδικασία προεπεξεργασίας δεδομένων και εισαγωγή μάθησης ευαισθησίας κόστους στα μοντέλα πρόβλεψης.

Η κατηγορία μάθησης με ευαισθησία κόστους περιλαμβάνει τις ακόλουθες μεθόδους (με μη-ταξινόμηση κόστους):

- Άμεσοι μέθοδοι (Drummond and Holte, 2000)
- Μέθοδοι Μετα-εκμάθησης
- Μέθοδοι Μετα-κόστους (Domingos, 1999)
- Ταξινομητής με ευαισθησία κόστους [Cost Sensitive Classifier (CSC)] (Witten and Frank, 2005)
- ¨Απλοϊκή¨ Bayes ευαισθησία κόστους (Chai et al. 2004)
- Εμπειρική Οριακή μέθοδος (Sheng and Ling, 2006)
- Μέθοδος κόστους (costing) (Zadrozny et al. 2003)
- Μέθοδος επιβάρυνσης (weighting) (Ting, 1998)

3.2.5 Μέτρηση Ευαισθησίας Κόστους

Οι μετρήσεις ευαισθησίας κόστους συνήθως υποθέτουν ότι το κόστος της πραγματοποίησης σφαλμάτων είναι γνωστό (Domingo, 1999; Turney, 2000; Elkan, 2001). Ο πίνακας κόστους ορίζει το κόστος που προκύπτει σε εσφαλμένα θετικά (false positive) και σφαλμένα αρνητικά. Για παράδειγμα, η μεταβλητή x μπορεί να συνδέεται με το κόστος C(i, j, x) που ορίζει το κόστος για την πρόβλεψη μια τάξης i για x όταν η αληθής τάξη είναι j.

Το αποτέλεσμα είναι να ληφθεί μια απόφαση για την ελαχιστοποίηση του αναμενόμενου κόστους. Η βέλτιστη πρόβλεψη για το x ορίζεται ως

$$\sum_f P(j|x) * C(i,j,x) \qquad (3.1)$$

Όπου C είναι το κόστος και P η αντίστοιχη πιθανότητα της τάξης i για x

3.3 Τεχνικές Επιπέδου Δεδομένων

Το πρόβλημα ανομοιογενούς μάθησης σχετίζεται με την απόδοση αλγορίθμων μάθησης που χρησιμοποιούν ανεπαρκή δεδομένα και ασύμμετρες τάξεις κατανομών. Η μάθηση από τέτοια δεδομένα λόγω των πολύπλοκων χαρακτηριστικών των συνόλων ανομοιογενών δεδομένων , απαιτούν νέους αλγόριθμους και κατάλληλα εργαλεία για να μετασχηματίσουν με αποδοτικό τρόπο τεράστια ποσά πρωτογενών δεδομένων σε επιθυμητή παράσταση πληροφοριών και γνώσεων.

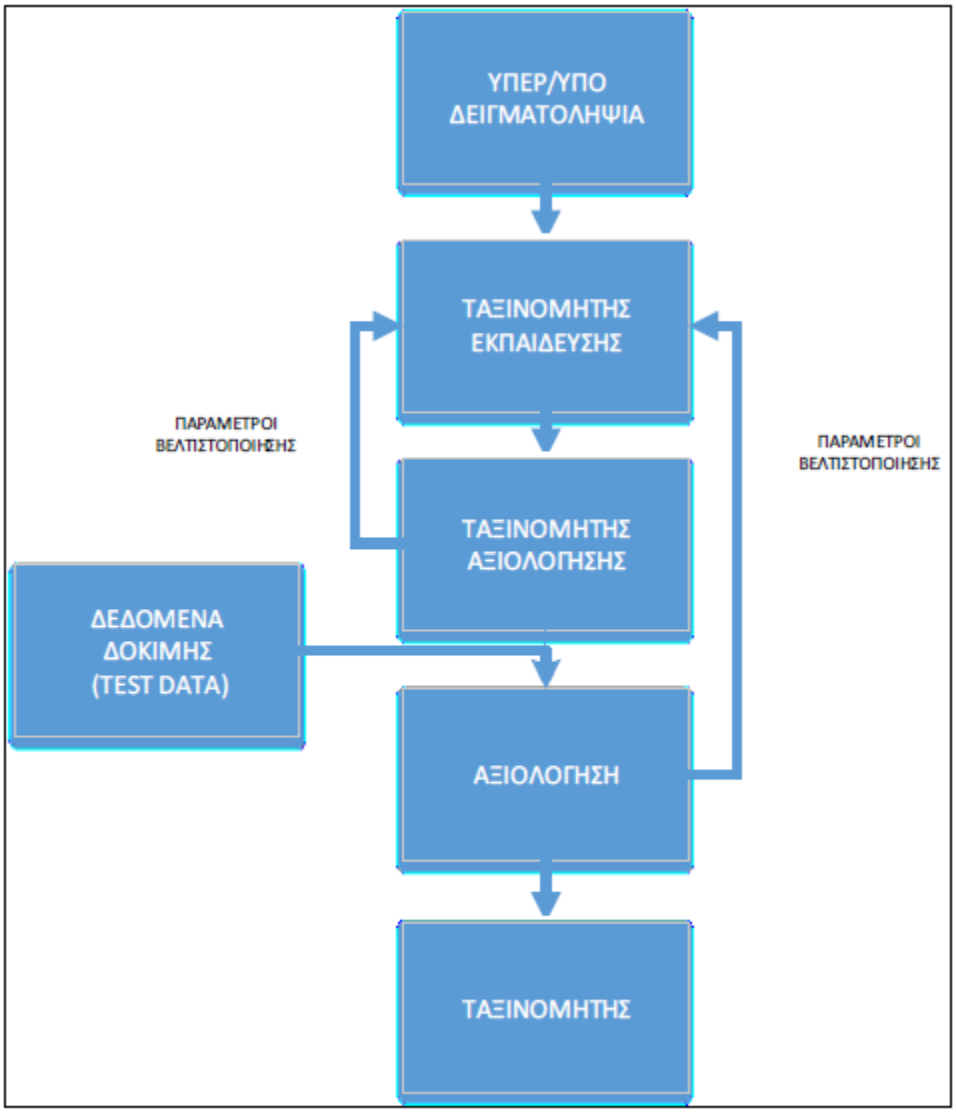
3.3.1 Μέθοδοι Προεπεξεργασίας Δεδομένων

Οι μέθοδοι προ-επεξεργασίας δεδομένων μπορούν να χρησιμοποιηθούν με τους αλγόριθμους μάθησης συνόλων (Batista et al., 2004; Fernandez et al., 2008). Η επίδραση της αλλαγής κατανομής τάξεων για χρήση σε ανομοιογενή σύνολα δεδομένων μπορεί να γίνει με τεχνικές **επανα-δειγματοληψίας (resampling)**. Οι τεχνικές αυτές που χρησιμοποιούνται με αλγόριθμους μάθησης συνόλων περιλαμβάνουν τις ακόλουθες κατηγορίες:

i. τυχαία υπο-δειγματοληψία (random undersampling)
ii. τυχαία υπερ-δειγματοληψία (random oversampling)
iii. τεχνική υπερ-δειγματοληψίας με συνθετική μειονότητα (Synthetic Minority Oversampling Technique [SMOTE]) (Chawla et al., 2002)
iv. τροποποιημένη τεχνική υπερ-δειγματοληψίας με συνθετική μειονότητα (Modified Synthetic Minority Oversampling Technique [MSMOTE]), (Hu et al., 2009)
v. επιλεκτική προ-επεξεργασία ανομοιογενών δεδομένων (Selektive Preprocessing of Imbalanced Data [SPIDER]) (Stefanowski and Wilk, 2008).
vi.

3.3.2 Τυχαία Επιλογή Συνόλων Δεδομένων Υποδειγματοληψίας και Υπερδειγματοληψίας

Οι μηχανικές διαδικασίες της τυχαίας υπερδειγματοληψιας ακολουθούν την περιγραφή της προσθέτοντας ένα σύνολο που προκύπτει από τη τάξη μειοψηφίας, δηλαδή για ένα σύνολο τυχαία επιλεγμένων παραδειγμάτων μειοψηφίας αυξάνεται το αρχικό σύνολο αναπαράγοντας τα επιλεγμένα παραδείγματα και προσθέτοντας στο αρχικό σύνολο (Mease et al., 2007).



Εικόνα 3.2: Διαδικασίες Δεδομένων Δοκιμής Υποδειγματοληψίας και Υπερδειγματοληψίας

3.3.3 Ενιαίο Σύνολο Δεδομένων Υποδειγματοληψίας

Δυο χαρακτηριστικά παραδείγματα ενημερωμένης υποδειγματοληψίας (informed undersampling) αποτελούν ο ¨***αλγόριθμους EasyEnsemble*** ¨ και ο ¨αλγόριθμους BalanceCascade ¨ (Liu et al., 2006). Οι αλγόριθμοι αυτοί μπορούν να ξεπεράσουν το μειονέκτημα της απώλειας πληροφοριών που παρουσιάζεται στην αρχική μέθοδο τυχαίας υποδειγματολειψίας. Ο

αλγόριθμους EasyEnsemble, ένας αλγόριθμος μάθησης χωρίς επίβλεψη (unsupervised learning algorithm), δημιουργεί ένα σύστημα συνολικής μάθησης (ensemble learning system), με ανεξάρτητη δειγματοληψία διαφόρων υποσυνόλων από την τάξη πλειοψηφίας και με δημιουργία πολλαπλών ταξινομητών που βασίζεται στο συνδυασμό κάθε υποσύνολου με τα δεδομένα της τάξης μειοψηφίας (Liu et al., 2006).

3.3.4 Ενιαία Σύνολα Δεδομένων Υποδειγματοληψιας

Η Συνθετική Τεχνική Υπερδειγματοληψιας Μειονότητας (Synthetic Minority Oversampling Technique [SMOTE]) δοκιμάζεται σε ποικιλία συνόλων δεδομένων με μεταβαλλόμενους βαθμούς ανομοιογένειας και με μεταβαλλόμενα ποσά δεδομένων στα δοκιμαστικά σύνολα παρέχοντας διαφορετικά πεδία δοκιμών.

Για τη δημιουργία ενός συνθετικού δείγματος επιλέγεται τυχαία ένας από K-πλησιέστερους γείτονες και πολλαπλασιάζεται η διαφορά διανυσμάτων με έναν τυχαίο αριθμό μεταξύ [0,1] και τελικά το διάνυσμα αυτό προστίθεται στο xi , δηλαδή προκύπτει το επαναληπτικό σχήμα:

$$x_{i+1} = x_i + (x_i * - x_i).\ \ d, i = 0,1,2,.....\ \ x_i \in S_{min} \qquad (3.2)$$

όπου xi είναι το στιγμιότυπο μειοψηφίας που θεωρείται, xi * είναι ένας από K-πλησιέστερους γείτονες για το xi και d είναι ένας τυχαίος πραγματικός αριθμος, με d ∈ [o,1] (Chawla et al., 2002).

Διάφοροι προσαρμοσμένες μέθοδοι δειγματοληψίας (adaptive sampling methods), όπως ο αλγόριθμος Border-line-SMOTE (Han et al., 2005) και τον αλγόριθμο ADA-SYN (Adaptive Synthetic Sampling) (He et al., 2008), έχουν επίσης παρουσιαστεί για τη βελτίωση του αλγορίθμου SMOTE.

3.3.5 Τεχνική Συνθετικής Μειονοτικής Υπερδειγματοληψιας (SMOTE)

Η υπερδειγματοληψία με αντιγραφή μπορεί να οδηγήσει σε παρόμοια αλλά περισσότερο ειδικές περιοχές σε μελλοντικούς χώρους όπως οι περιοχές αποφάσεων για την μειονοτική τάξη. Αυτό μπορεί να οδηγήσει σε υπερπλήρωση των πολλαπλών αντιγράφων των παραδειγμάτων της

μειονοτικής τάξης. Για να αντιμετωπιστεί η υπερπλήρωση και να διευρυνθεί η περιοχή αποφάσεων των παραδειγμάτων της μειονοτικής τάξης έχουν αναπτυχτεί νέες τεχνικές για τη δημιουργία συνθετικών παραδειγμάτων που λειτουργούν σε ¨μελλοντικό¨ χώρο παρά σε χώρο δεδομένων (Chawla et al., 2002).

Τα συνθετικα παραδειγματα επιτρεπουν στον ταξινομητη να δημιουργησει μεγαλυτερες και ολιγότερες ειδικές περιοχές , παρά μικρότερες και περισσότερο ειδικές περιοχές, όπως γίνεται με υπερδειγματοληψία με αντιγραφή. Άμεσο αποτέλεσμα είναι ότι τα δένδρα αποφάσεων γενικεύουν καλύτερα.

Η τεχνική SMOTE δοκιμάζεται σε ποικιλία συνόλων δεδομένων με μεταβαλλόμενους βαθμούς ανομοιογένειας και με μεταβαλλόμενα ποσά δεδομένων στα δοκιμαστικά σύνολα παρέχοντας διαφορετικά πεδία δοκιμών. Η τεχνική αυτή χρησιμοποιεί ως ταξινομητές C.4 και Ripper (Cohen, 1995) και αναφέρεται ότι υπερτερεί σε απόδοση άλλες μεθόδους που συμπεριλαμβάνουν δειγματοληπτικές στρατηγικές (Ripper’s Loss Ratio) και απλοϊκές Bayes μεταβάλλοντας κατάλληλα τις τάξεις.

3.3.6 Τεχνικές Δειγματοληψίας BorderLIne-SMOTE

Η τεχνική ***Συνθετικής Μειονοτικής Υπερδειγματοληψίας*** (Synthetic Minority Oversampling Technique [SMOTE]) είναι μια ¨νοήμων¨ μέθοδος υπερδειγματοληψίας που προσθέτει νέα τεχνητά μειονοτικά παραδείγματα με προεκβολή μεταξύ προϋπαρχόντων τεκμηρίων μάλλον παρά από αντιγραμμένων αυθεντικών παραδειγμάτων (Chawla et al., 2002). Η τεχνική υπολογίζει τους κ-πλησιέστερους γείτονες της μειονοτικής τάξης για κάθε μειονοτικό παράδειγμα.

Η ***τεχνική συνοριακής γραμμής*** (borderline SMOTE) (BSM) (Han et al., 2005) είναι τροποποίηση της τεχνικής SMOTE, που επιλέγει μειονοτικά παραδείγματα που θεωρούνται ότι είναι στο σύνορο (border) της περιοχής SMOTE μόνο για να υπερδειγματίζουν στιγμιότυπα (instancses) αυτές, από το να τις υπερδειγματίσει όλες ή ένα τυχαίο υποσύνολο.

Δυο νέες μέθοδοι Υπερδειγματοληψίας , η Borderline-SMOTE1 και η Borderline-SMOTE2 στα οποία μόνο τα παραδείγματα μειονότητας κοντά στη διαχωριστική γραμμή είναι υπερδειγματολειπτικά (Oversampled). Για την κλάση μειονότητας τα πειράματα δείχνουν ότι οι προσεγγίσεις του Han, Wang και Mao (Borderline-SMOTE: A New Over-Sampling Method in

Imbalanced Data Sets Learning) πέτυχαν καλύτερο TP-rate και F-value από ότι τις μεθόδους SMOTE και τυχαία υπερδειγματοληψία.

3.3.7 Τεχνικές Δειγματοληψίας ADASYN

Στον ***αλγόριθμο Adasyn*** χρησιμοποιείται μια κατανομή βαρών (weighted distribution) για διαφορετικά παραδείγματα κλάσης της μειονότητας (minority class) συμφώνα με το επίπεδο της δυσκολίας στην εκμάθηση , όπου τα περισσότερα σύνθετα δεδομένα (synthetic data) παράγουν παραδείγματα των μειονοτικών κλάσεων που είναι πιο δύσκολο να μάθει σε σύγκριση με τα παραδείγματα των μειονοτήτων (minority examples) που είναι πιο εύκολο να μάθουν.

Η αλγοριθμική προσέγγιση Adasyn βελτιώνει τη μάθηση σε σχέση με τις κατανομές των δεδομένων με δυο τρόπους : (i) μείωση της μεροληψίας (bias) που εισήγαγε η ανομοιογενής κλάση και (ii) η προσαρμοστική αλλαγή του ορίου απόφασης ταξινόμησης προς τα δύσκολα παραδείγματα (adaptively shifting the classification decision boundary toward the difficult examples). Αναλύσεις προσομοιώσεις για διάφορα σύνολα μηχανικής μάθησης δεδομένα δείχνουν την αποτελεσματικότητα αυτής της μεθόδου σε πέντε μετρήσεις αξιολόγησης.

Αλγόριθμος ADASYN

Η επιτυχία των πρόσφατων συνθετικών προσεγγίσεων συμπεριλαμβανόμενου του SMOTE, SMOTEBoost , και του DataBoost-IM , οδήγησε σε μια προσαρμοστική μέθοδο διευκολύνοντας την μάθηση για ανομοιογενή δεδομένα (He et al., 2008). Στόχος είναι διπλός : i) τη μείωση της μεροληψίας (bias) και ii) η προσαρμοστική μάθηση. Ο αλγόριθμος ταξινόμησης adasyn για το πρόβλημα δυο κλάσεων περιγράφεται από την ακόλουθη αλγοριθμική διαδικασία (σε ψευδοκωδική μορφή):

Αλγόριθμος ADASYN (D_{tr}, m, n, y_i, s_i)

Είσοδος: Τα στοιχεία εκπαίδευσης D_{tr} με m δείγματα $\{x_i, y_i\}$, i = 1, ..., m, όπου x_i είναι ένα παράδειγμα σε n διάσταση στο χώρο των χαρακτηριστικών X και $y_i \in Y = \{1, -1\}$ είναι η ετικέτα της ταυτότητας της κλάσης σχετίζεται με x_i. (Θέσε το m_s και m_l σαν τον αριθμό των παραδειγμάτων των κλάσεων των μειονοτήτων και των παραδειγμάτων των κλάσεων της πλειονότητας, αντίστοιχα. Ως εκ τούτου, $m_s \leq m_l$ και $m_s + m_l = m$.)

Υπολογιστική Διαδικασία/ Εκτέλεση

(1) Υπολόγισε τον βαθμό της ανομοιογενής κλάσης:

$d = m_s / m_l$, όπου $d \in (0, 1]$. (4.3)

(2) Αν $d < d_{th}$ τότε (d_{th} είναι το παρόν κατώφλι για το μέγιστο ανεκτό βαθμό της ανομοιογενής κλάσης:

(a) Υπολόγισε τον αριθμό των παραδειγμάτων σύνθετων δεδομένων (synthetic data) που χρειάζονται για να παραχθούν για την κλάση των μειονοτήτων:

$$G = (m_l - m_s) \times \beta, \quad (4.4)$$

όπου $\beta \in [0, 1]$ είναι μια παράμετρος που χρησιμοποιείται για να καθοριστεί το επιθυμητό επίπεδο ισορροπίας μετά τη δημιουργία των συνθετικών στοιχείων. Η τιμή β=1 σημαίνει ένα πλήρως ισορροπημένο σύνολο δεδομένων που έχει δημιουργηθεί μετά τη γενίκευση της διαδικασίας.

(b) Για κάθε παράδειγμα, $x_i \in$ minorityclass, βρες τον K κοντινότερο γείτονα με βάση την Ευκλείδεια απόσταση στην n διάσταση, και υπολογίζετε η αναλογία r_i ορίζεται ως:

$$r_i = \Delta i / K, \; i = 1, \ldots, m_s \quad (4.5)$$

όπου Δ_i είναι ο αριθμός των παραδειγμάτων στον K κοντινότερο γείτονα του x_i που ανήκει στην κύρια κλάση, επομένως $r_i \in [0, 1]$;

(c) Κανονικοποιήσαμε του r_i σύμφωνα με

$$\hat{r}_i = r_i / \sum_{i=1}^{m_s} r_i, \quad (4.6)$$

ούτως ώστε το $\hat{r}_i$ είναι μια κατανομή πυκνότητας ($\sum_i \hat{r}_i = 1$).

(d) Υπολογισμός του αριθμού των παραδειγμάτων των συνθετικών δεδομένων που χρειάζονται για την δημιουργία για κάθε παράδειγμα μειονότητας x_i :

$g_i = \hat{r}_i \times G$, (4.7)

όπου G είναι ο συνολικός αριθμός των παραδειγμάτων των συνθετικών δεδομένων που χρειάζεται να δημιουργούνται για την κατηγορία μειονότητας όπως ορίζεται στην εξίσωση 4.3.

(e) Για κάθε παράδειγμα μειονοτικής κλάσης δεδομένων x_i, δημιουργεί το g_i παραδείγματα συνθετικών στοιχείων σύμφωνα με τα ακόλουθα βήματα:

Κάνε από 1 έως g_i :

(i) Επέλεξε τυχαία ένα από τα παραδείγματα μειονοτικών δεδομένων, x_{zi}, από το K κοντινότερο γείτονα για δεδομένα x_i.

(ii) Δημιουργώντας παράδειγμα συνθετικών δεδομένων :

$s_i = x_i + (x_{zi} - x_i) \times \lambda$, (4.8)

όπου $(x_{zi} - x_i)$ είναι το διαφορετικό διάνυσμα σε n διαστάσεις, και λ είναι ένας τυχαίος πραγματικός αριθμός $\lambda \in [0, 1]$.

Τέλος υπολογιστικής διαδικασίας/ βρόχου.

Εικόνα 3.3: Αλγόριθμος Adasyn

3.3.8 Δειγματοληψία Με Τεχνικές Εκκαθάρισης Δεδομένων (Data Cleaning)

Διάφορες τεχνικές εκκαθάρισης δεδομένων (Data Cleaning Techniques) μπορούν να εφαρμοστούν αποδοτικά για την απομάκρυνση της υπερκάλυψης (overlapping) που παρουσιάζεται από τις μεθόδους δειγματοληψίας. Η μέθοδος δειγματοληψίας. Η μέθοδος ¨***συνδέσμων Tomek***¨ ορίζεται ως ένα ζεύγος πλησιέστερων γειτόνων αντίθετων τάξεων με ελαχιστοποιημένη απόσταση (Tomek, 1976).

Υποθέστε ότι δίνεται ένα ζεύγος στιγμιότυπων (xi , xj)και η απόσταση τους d (xi , xj) με xi ∈Smin και xj ∈Smax , τότε το ζεύγος (xi , xj) καλείται ένας σύνδεσμος Tomek αν δεν υπάρχει στιγμιότυπο xk , τέτοιο ώστε να ισχύει

$$d(x_i , x_k) < d(x_i , x_j) \text{ ή } d(x_j , x_k) < d(x_i , x_j). \quad (3.3)$$

Αν δύο στιγμιότυπα σχηματίζουν ένα σύνδεσμο Tomek , τοτε είτε το ένα από τα στιγμιότυπα αυτά αποτελεί ¨θόρυβο¨ ή και τα δυο είναι κοντά σε ένα σύνορο.

Οι σύνδεσμοι Tomek μπορούν να χρησιμοποιηθούν για να εκκαθαρισθούν ανεπιθύμητες επικαλύψεις μεταξύ τάξεων μετά από συνθετική δειγματοληψία όπου όλοι οι σύνδεσμοι Tomek απομακρύνονται μέχρις ότου όλα τα ζεύγη πλησιέστερων γειτόνων με ελαχιστοποιημένη απόσταση είναι της ιδίας τάξης. Με την απομάκρυνση παραδειγμάτων επικάλυψης μπορεί να επιτευχθούν συμπλέγματα με καλά ορισμένους κανόνες ταξινόμησης με βελτιωμένη απόδοση ταξινόμησης (Batista et al., 2004; Laurikkala, 2001; Kubat and Matwin, 1997).

3.3.9 Εναλλακτικές Μέθοδοι

Με την απομάκρυνση της υπερκάλυψης μπορεί κανείς να καθορίσει κανόνες ταξινόμησης για να βελτιώσει την απόδοση της ταξινόμησης. Κάποιες ενδεικτικές εργασίες σε αυτή την περιοχή περιλαμβάνεται η μέθοδος της επιλογής μια πλευράς (one-slide selection) (OSS) (Kubat& Matwin,1997), η μέθοδος ολοκλήρωσης του συμπυκνωμένου κανόνα του κοντινότερου γείτονα και οι σύνδεσμοι Tomek (condensed nearest neighbor rule and Tomek links) (CNN+Tomek Links) (Batista et al.,2004), οι κανόνες καθαρισμού γειτονίας (NCL) (Laurikkala, 2001) βασισμένο στον κανόνα επεξεργασίας του πληρέστερου γείτονα (ENN), η οποία καταργεί παραδείγματα που διαφέρουν από δυο από τους τρεις κοντινότερους γείτονες και οι ολοκληρώσεις του SMOTE με ENN (SMOTE+ENN) και ο SMOTE με τις συνδέσεις Tomek (SMOTE+Tomek) (Batista et al., 2004).

3.3.10 Δείγματα Βασισμένα σε Συστάδες/Συμπλέγματα (clusters)

Οι αλγόριθμοι δειγματοληψίας που βασίζονται σε συμπλέγματα (Cluster based Sampling Methods) παρέχουν ένα επιπρόσθετο στοιχειό ευελιξίας που δεν υπάρχει στους περισσότερους απλούς και συνθετικής δειγματοληψίας αλγόριθμους, έτσι ώστε να μπορούν να χρησιμοποιηθούν σε πολύ ειδικά προβλήματα, ο αλγόριθμος **υπερδειγματοληψίας που βασίζεται σε συμπλέγματα** (Cluster Based Oversampling [CBO] algorithm) μπορεί να χρησιμοποιηθεί αποδοτικά για το ανομοιογενές πρόβλημα εντός-τάξεων σε συνεργασία με το ανομοιογενές πρόβλημα μεταξύ-τάξεων (Jo and Japkowicz, 2004) και να αποτελέσει μια αποδοτική στρατηγική για τα ανομοιογενή σύνολα δεδομένων.

Ειδικές Ενσωματώσεις Δειγμάτων (Sampling and Boosting)

Η ολοκλήρωση στρατηγικών δειγματοληψίας με τεχνικές συνολικής μάθησης έχουν χρησιμοποιηθεί με επιτυχία για τη βελτίωση απόδοσης. Ο **αλγόριθμος SMOTE- Boost** βασίζεται στην ολοκλήρωση του αλγορίθμου SMOTE με τον αλγόριθμο Adaboost.M2, και παρουσιάζει τη συνθετική δειγματοληψία σε κάθε ενδυναμωτική επανάληψη (boosting iteration). Με τον τρόπο αυτό κάθε διάδοχος συνολικός ταξινομητής εστιάζεται περισσότερο στην τάξη μειοψηφίας. Επειδή κάθε συνολικός ταξινομητής κατασκευάζεται σε μια διαφορετική δειγματοληψία δεδομένων, ο τελευταίος ταξινομητής αναμένεται να έχει μια ευρύτερη και καλά ορισμένη περιοχή αποφάσεων για την τάξη μειοψηφίας (Chawla et al., 2003).

Ο **αλγόριθμος DataBoost-IM** συνδυάζει τεχνικές δημιουργίας δεδομένων με τον αλγόριθμο AdaBoost.M1 για να πετύχει μια προβλέψιμη υψηλή ακρίβεια για την τάξη μειοψηφιας χωρις να χασει ακριβεια στην ταξη πλειοψηφιας. Ο αλγόριθμος Databost-IM δημιουργεί συνθετικά δείγματα σύμφωνα με την αναλογία των ¨δύσκολων προς μάθηση¨ δειγμάτων μεταξύ τάξεων (Guo and Viktor, (2004a-2004b)).

Φίλτρα Επαναδειγματοληψίας

Το φίλτρο επαδειγματοληψίας (resample) είναι ένα φίλτρο στα δεδομένα με επίβλεψη, όπου παράγει ένα τυχαίο υπόδειγμα (subsample) από ένα σύνολο δεδομένων χρησιμοποιώντας είτε δειγματοληψία με ή χωρίς αντικατάσταση. Το αρχικό σύνολο δεδομένων πρέπει να χωράει ολόκληρο στη μνήμη. Ο αριθμός των στιγμιοτύπων στο σύνολο δεδομένων που δημιουργούνται πρέπει να είναι καθορισμένος. Το σύνολο δεδομένων πρέπει να έχει ονομαστικά χαρακτηριστικά κλάσης, αν δεν έχει θα πρέπει να χρησιμοποιηθεί η έκδοση των μη επιβλεπόμενων δεδομένων (unsupervised). Το φίλτρο μπορεί να γίνει για να διατηρηθεί η κατανομή των χαρακτηριστικών κλάσης στο υπόδειγμα, ή να πολώσει την κατανομή κλάσης προς την ομοιόμορφη κατανομή.

3.4 Μεθοδολογίες Τεχνικών Συνόλων για Ανομοιογενή Σύνολο Δεδομένων

3.4.1 Ανασκόπηση Αλγορίθμων Μηχανικής Μάθησης και αλγόριθμοι για Μάθηση Συνόλων ταξινόμησης

Μια μικρή ανομοιογένεια στην κλάση κατανομής, υπό ορισμένες συνθήκες, κρίνεται ότι δεν είναι σοβαρή, αλλά όταν κάποιες κλάσεις είναι σε μεγάλο βαθμό υπό εκπροσωπημένες , ένα μεγάλο ποσοστό των αντίστοιχων αλγορίθμων μηχανικής μάθησης είναι πιθανό να δημιουργήσουν διάφορες δυσκολίες. Κάποιες περιπτώσεις ανήκουν στην κατηγορία των μικρών κλάσεων που χάνονται ανάμεσα στις πιο συχνές περιπτώσεις κατά τη διάρκεια της μάθησης. Κατά συνέπεια, οι ταξινομητές όπως τα δένδρα αποφάσεων και οι κανόνες μάθησης δεν είναι σε θέση να ταξινομήσουν σωστά νέες περιπτώσεις από την κλάση μειονοτήτων.

Οι ταξινομητές δένδρων απόφασης συνήθως χρησιμοποιούν τις λεγόμενες τεχνικές «μετά-κλαδέματος» (**post-pruning**), που αξιολογούν την απόδοση των δένδρων απόφασης που κλαδεύονται χρησιμοποιώντας ένα σύνολο επικύρωσης (validation set –Esposito et al., 1997). Οποιοσδήποτε κόμβος μπορεί να αφαιρεθεί και να ανατεθεί στην πιο κοινή κλάση των στιγμιοτύπων εκπαίδευσης που είναι ταξινομημένοι οι υπό εξέταση κομβόι. Έτσι αν μια κλάση είναι σπάνια, ο αλγόριθμος δένδρου απόφασης συχνά ¨κλαδεύει¨ το δένδρο κάτω από κάθε κόμβο που είναι μονός του (φύλλο), που ταξινομεί όλα τα στιγμιότυπα ως μελή μια κοινής κλάσης και που οδηγεί σε όχι τόσο μεγάλη ακρίβεια στα στιγμιότυπα της κλάσης μειονοτήτων. Στην ερευνητική μελέτη αυτή χρησιμοποιήθηκαν τα γνωστά δένδρα αποφάσεων (Quilan, 1993).

Μια ενδιαφέρουσα αναφορά στο πεδίο των αλγορίθμων της μηχανικής μάθησης στα ανομοιογενή δεδομένα έχει προταθεί από Weiss and Provost (2003). Το ερευνητικό έργο αυτό εστιάζεται κυρίως σε διαφορές και ομοιότητες ανάμεσα στα προβλήματα των σπανίων κλάσεων και σπανίων υποθέσεων. Στη συνέχεια γίνεται μια αναφορά των ζητημάτων όπου μελετήθηκαν και το εύρος των λύσεων που προτάθηκαν στον τομέα της εξόρυξης ανομοιογενών δεδομένων.

Οι ερευνητικές προσπάθειες που απασχολήθηκαν για τα ανομοιογενή δεδομένα στη επιβλεπόμενη μάθηση, μπορούν να ταξινομηθούν στις ακόλουθες τέσσερις κατηγορίες:

i. υποδειγματοληψία στην κυρίαρχη κλάση, έτσι ώστε να ταιριάζει με το μέγεθος της άλλης κατηγορίας.

ii. υπερδειγματοληψία στην κλάση της μειονότητας έτσι ώστε να ταιριάζει με το μέγεθος της άλλης κατηγορίας
iii. η εσωτερική πόλωση στην διακριτική διαδικασία, έτσι ώστε να ταιριάξει με το μέγεθος της άλλης κατηγορίας και
iv. ειδικά εμπειρογνωμικά συστήματα, που αποτελούν ισορροπημένους ταξινομητές.

Μια απλή μέθοδος επιπέδου δεδομένων (data level), που ,μπορεί να χρησιμοποιηθεί στα ανομοιογενή δεδομένα, είναι να επανατοποθετήσουμε βάρη (reweigh) εκπαιδεύοντας στιγμιότυπα ανάλογα με την κάθε κλάση (Domingos,1998). Η ιδέα είναι να αλλαχτούν οι κατανομές κλάσης στο σύνολο δεδομένων ορίζονται στην πιο κοστοβόρα κλάση. Ας υποθέσουμε ότι τα στιγμιότυπα της θετικής κλάσης είναι πέντε φορές περισσότερες από τα στιγμιότυπα της αρνητικής κλάσης. Εάν ο αριθμός των αρνητικών στιγμιοτύπων είναι τεχνητά αυξημένος από έναν παράγοντα πέντε, τότε το σύστημα εκμάθησης σκοπεύει να μειώσει το αριθμό των σφαλμάτων ταξινόμησης, έτσι θα καταλήξουμε που θα παρεκκλίνει στην αποφυγή του σφάλματος στην αρνητική κλάση, δεδομένου ότι κάθε τέτοιο σφάλμα υπάρχει σχετικά επιβάρυνση (τέσσερις φορές περισσότερο).

Στην επόμενη παράγραφο περιγράφονται μέθοδοι μηχανικής μάθησης και εξηγούνται οι λόγοι για τη χαμηλή τους απόδοση στα ανομοιογενή σύνολα δεδομένων. Επιπλέον, παρουσιάζεται μια επισκόπηση στο χειρισμό ανομοιογενών συνόλων δεδομένων, με βασικές λεπτομέρειες της προσέγγισης που χρησιμοποιείται.

Με τα αποτελέσματα της ανομοιογένειας στα σύνολα δεδομένων έχει ασχοληθεί ο Japkowicz (2000). Ο συγγραφέας αξιολογεί κυρίως δυο στρατηγικές δεδομένων επιπέδου: υποδειγματοληψίας και επαναδειγματοληψία (under-sampling and resampling) και εξετάζονται δυο επαναδειγματοληπτικές μέθοδοι. Η τυχαία επαναδειγματοληψία αποτελείται από επεναδειγματοληψία της μικρότερης κλάσης τυχαία μέχρι να αποτελέσει τόσα δείγματα όσα και της κυρίαρχης κλάσης και στην εστιασμένη επαναδειγματοληψία αποτελείται για την επαναδειγματοληψία μόνο εκείνων των στιγμιοτύπων όπου έλαβαν χώρα στα όρια μεταξύ μειονοτικών και κυριάρχων κλάσεων. Η τυχαία υπερδειγματοληψία, στην οποία συμπεριλάμβαναν υποδειγματολειπτικά τα δείγματα κυρίαρχης κλάσης τυχαία μέχρι ο αριθμός τους να αντιστοιχηθεί στον αριθμό δειγμάτων της μειονότητας, εστιάζεται στην υποδειγματοληψία συμπεριλαμβανομένης της υποδειγματοληψίας της κυρίαρχης κλάσης δειγμάτων που βρίσκεται πιο

μακριά. Οι προσεγγίσεις δειγματοληψίας είναι αποτελεσματικές και έχει παρατηρηθεί ότι η χρήση των εξελιγμένων τεχνικών δειγματοληψίας δεν προσφέρει κανέναν σαφές πλεονέκτημα στο πεδίο αυτό (Japkowicz, 2000).

Ο Kubat και ο Matwin (1997) υποδειγματίζουν επιλεκτικά στην κυρίαρχη κλάση, κρατώντας τον αρχικό πληθυσμό της κλάσης της μειονότητας με ικανοποιητικά αποτελέσματα. Ο Batista και οι άλλοι χρησιμοποιούν μια πιο εξελιγμένη τεχνική υποδειγματοληψίας προκειμένου να ελαχιστοποιηθεί η ποσότητα των δυνητικά χρήσιμων δεδομένων. Τα στιγμιότυπα της κυρίαρχης κλάσης είναι ταξινομημένα ως στιγμιότυπα ¨ασφαλή¨, ¨οριακά¨ και ¨θορύβου¨. Οι οριακές υποθέσεις και του θορύβου ανιχνεύτηκαν χρησιμοποιώντας το ¨σύνδεσμο Tomek¨ και μπορούν να απομακρυνθούν από το σύνολο δεδομένων μας. Οι ασφαλείς κλάσεις κυριαρχίας και όλα τα στιγμιότυπα κλάσης μειονότητας χρησιμοποιούνται για την εκπαίδευση του συστήματος μάθησης. Ένας σύνδεσμος Tomek μπορεί να οριστεί δίνοντας δυο στιγμιότυπα x και y που ανήκουν σε διαφορετικές κλάσεις, με d(x,y) να είναι η απόσταση των x και y. Ένας ζεύγος (x,y) είναι ο ¨σύνδεσμος Tomek¨ , αν δεν υπάρχει μια περίπτωση z , τέτοια ώστε d(x, z) < d(x, y) ή d(y, z) < d(y, x).

Οι τεχνικές υποδειγματοληψίας και υπερδειγματοληψίας εμφανίζουν αρκετά μειονεκτήματα. Συγκεκριμένα , η υποδειγματοληψία μπορεί να απομακρύνει χρήσιμα στοιχειά και η υπερδειγματοληψία μπορεί να αυξήσει την πιθανότητα εμφάνισης του υπερταιριάσματος (overfitting), δεδομένου ότι περισσότεροι από τις υπερδειγματοληψίες μεθόδους κάνουν ακριβή αντίγραφα των στιγμιοτύπων της κλάσης μειονότητας. Με τον τρόπο αυτό, για παράδειγμα ένας συμβολικός ταξινομητής μπορεί να κατασκευάσει κανόνες που είναι φαινομενικά ακριβείς, αλλά στην πραγματικότητα, επικαλύπτουν ένα αναπαραγόμενο στιγμιότυπο.

Μια εναλλακτική προσέγγιση του επιπέδου δεδομένων έχει παρουσιαστεί από τους Ling και Li (1998). Συνδύασαν την υπερδειγματοληπτική κλάση μειονότητας με την υποδειγματοληπτική κλάση κυριαρχίας. Ωστόσο, ο υπερδειγματοληπτικός και υποδειγματοληπτικός συνδυασμός δεν παρέχει καμιά σημαντική βελτίωση. Ο Chawla (2002) προτείνει μια υπερδειγματοληπτική προσέγγιση στην οποία η κλάση μειονότητας είναι υπερδειγματοληπτική με τη δημιουργία ¨συνθετικών¨ στιγμιοτύπων πάρα με υπερδειγματοληψία με αντικατάσταση με καλυτέρα αποτελέσματα.

Οι ερευνητές Ng και Dash (2006) χρησιμοποιούν μια παρόμοια τεχνική για να εκπαιδεύσουν τους ταξινομητές σε μεγάλα και ανομοιογενή σύνολα δεδομένων. Στο ερευνητικό τους έργο, δειγματοληπτούν συνεχώς και από την κλάση κυριαρχίας και από την κλάση μειονότητας (χωρίς αντικατάσταση) , έως ότου δεν υπάρξει σαφή βελτίωση σε σχέση με την προηγούμενη ταξινόμηση, όπου χτίσθηκε ένα μικρότερο σύνολο δεδομένων. Όταν η προσφορά παραδειγμάτων μειονότητας έχει εξαντληθεί εξακολουθούν να δοκιμάζουν από την κλάση κυριαρχίας χωρίς αντικατάσταση, αλλά χρησιμοποιούν τυχαία υπερδειγματοληπτικά για να διατηρήσουν τις κατανομές κλάσεων ισορροπημένες.

Μια υβριδική διαδικασία δειγματοληψίας που χρησιμοποιεί ένα συνδυασμό των δυο τεχνικών δειγματοληψίας για να δημιουργήσει ένα ισορροπημένο σύνολο δεδομένων έχει παρουσιαστεί πρόσφατα (Seiffert et al., 2009). Αυτή η υβριδική τεχνική περιλαμβάνει εν μέρει δείγματα τα δεδομένα εκπαίδευσης χρησιμοποιώντας μια δειγματοληπτική τεχνική και στη συνέχεια την ολοκλήρωση της τεχνικής με άλλη τεχνική. Για παράδειγμα, θα μπορούσε κάποιος να υπερδειγματαληπτεί στην κλάση μειοψηφίας να μειώσει την ανομοιογένεια και στη συνέχεια να υποδειγματίζει την κυρίαρχη κλάση για να πετύχει ισορροπία στο σύνολο δεδομένων.

Ο Garcia (2009) προτείνει μια υποδειγματοληπτική διαδικασία καθοδηγούμενη από εξελικτικούς αλγορίθμους για να εκτελέσει μια επιλογή εκπαίδευσης συνόλων δεδομένου για την ενίσχυση των δένδρων αποφάσεων που λαμβάνονται από τον αλγόριθμο C4.5 και τα σύνολα κανόνων λαμβάνονται από τον κανόνα PART (Garcia et al., 2009). Η πρόταση αυτή συγκρουόμενη με άλλες υποδειγματοληπτικές τεχνικές και τα αποτελέσματα δείχνουν ότι η νέα προσέγγιση είναι πολύ ανταγωνιστική όσο αφορά την ακρίβεια , όταν συγκριθεί με υπερδειγματοληψία και υπερτερεί της τυπικής υποδειγματοληψίας.

Ο Khoshgoftaar (2010) πρότεινε μια ¨γενετική ¨ προσέγγιση βασισμένη σε αλγορίθμους. Η εξελικτική δειγματοληψία λειτουργεί ως μια τεχνική κυρίαρχη υποδειγματοληψία, όπου τα στιγμιότυπα από την κυρίαρχη κλάση αφαιρούνται επιλεκτικά. Αυτό διατηρεί τη σχετική ακεραιότητα της κυρίαρχης κλάσης, διατηρώντας παράλληλα την αρχική μειονοτική ομάδα κλάσεων. Η κύρια κριτική της υποδειγματοληψίας είναι ότι η πληροφορία μπορεί να χαθεί, όταν τα παραδείγματα αφαιρούνται από τα δεδομένα εκπαίδευσης, ενώ η υπερδειγματοληψία αυξάνει το μέγεθος του συνόλου δεδομένων προσθέτοντας παράλληλα είτε όχι (στην περίπτωση της τυχαίας

υπερδειγματοληψίας) ή συνθετικών πληροφοριών (σε περίπτωση των πιο ¨έξυπνων¨ υπεδειγματοληπτικών τεχνικών).

Το πρόβλημα ταξινόμησης ανομοιογενών δεδομένων είναι στενά συνδεδεμένο με το πρόβλημα της ταξινόμησης ευαισθησίας κόστους (cost-sensitive classification) (Chawla et al., 2008). Μια προσέγγιση αλγοριθμικού επιπέδου για την ενσωμάτωση του κόστους στη διαδικασία λήψης είναι ο καθορισμός σταθερών και άνισων μη ταξινομημένων κοστών ανάμεσα στις κλάσεις (Chen et al., 2009). Το μοντέλο κόστους μπορεί να πάρει μορφή του πίνακα κόστους, όπου το κόστος της ταξινόμησης ενός δείγματος απεικονίζεται για την αληθή κλάση j στην κλάση i αντιστοιχεί στον πίνακα στην είσοδο του πίνακα λij. Αυτός ο πινάκας συνήθως εκφράζεται σε όρους του μέσου μη ταξινομημένου κόστους για το πρόβλημα. Τα διαγώνια στοιχεία είναι συνήθως μηδενικά, όπου σημαίνει σωστή ταξινόμηση δεν έχει κόστος. Ο υποθετικός κίνδυνος για την κατασκευή μια απόφασης αi ορίζεται

$$R(\alpha i \mid x) = \sum \lambda ij \;\; P\,(\nu j \mid x) \qquad (3.4)$$

Η εξίσωση (3.4) δηλώνει ότι ο κίνδυνος της επιλογής κλάσης i ορίζεται από τον δημιουργημένο κόστος ταξινόμησης και την αβεβαιότητα της γνώσης μας σχετικά με την πραγματική κλάση x που εκφράζονται από τις εκ των υστέρων πιθανότητες. Ο στόχος της ευαισθησίας του κόστους ταξινόμησης είναι να ελαχιστοποιήσει το κόστος της εσφαλμένης ταξινόμησης, η οποία μπορεί να υλοποιηθεί με την επιλογή της κλάσης (ν_j) με την ελάχιστη προϋπόθεση κινδύνου (Chen et al., 2009).

Μια πολλαπλή προσέγγιση αναδειγματοληψίας (Estabrooks και Japkowicz, 2004) έχει χρησιμοποιηθεί για να συνδυάσει τα αποτελέσματα πολλών ταξινομητών, κάθε μια προκαλείται μετά από υπέρ-δειγματοληψία ή υπό-δειγματοληψία των δεδομένων με διαφορετικές τιμές υπέρ/ υπό-δειγματοληψίας. Η προσέγγιση αυτή αναγνωρίζει το γεγονός ότι εξακολουθεί να μην είναι σαφές ποια μέθοδος δειγματοληψίας αποδίδει καλύτερα, ποιος δειγματοληπτικός ρυθμός θα πρέπει να χρησιμοποιείται και αν ο αριθμός δειγματοληψίας είναι η σωστή επιλογή. Η συνδυαστική ¨μέθοδος MetaCost¨ (Domingos, 1998) είναι μια άλλη μέθοδος για την κατασκευή ενός ταξινομητή

κόστους ευαισθησίας (cost-sensitive). Η διαδικασία ξεκινά να εκπαιδεύει ένα μοντέλο με ευαισθησία κόστους, εφαρμόζοντας μια οικονομικά ευαίσθητη διαδικασία, η οποία χρησιμοποιεί ένα βασικό αλγόριθμο μάθησης. Στην συνέχεια η μέθοδος MetaCost εκτιμά τις πιθανότητες της κλάσης με τη χρήση της ¨μεθόδου σακουλιάσματος¨ (Bagging) και στη συνέχεια επαναετικετοποιεί (relabels) τα εκπαιδευμένα στιγμιότυπα με τις ελάχιστες αναμενόμενες κλάσης του κόστους και τέλος επανεκπαιδεύεται ένα μοντέλο χρησιμοποιώντας το τροποποιημένο σύνολο εκπαίδευσης.

Οι Chan και Stolfo (2001) τρέχουν μια σειρά προκαταρκτικών πειραμάτων, ώστε να προσδιορίσουν μια καλή κατανομή κλάσης και στη συνέχεια το δείγμα με αυτό τον τρόπο παράγει πολλαπλά σύνολα εκπαίδευσης με την επιθυμητή κλάση κατηγορίας. Κάθε σύνολο εκπαίδευσης περιλαμβάνει όλα τα παραδείγματα της κλάσης μειονότητας και ένα υποσύνολο των παραδειγμάτων περιλαμβάνει παραδείγματα της κυρίαρχης κλάσης. Ωστόσο κάθε παράδειγμα της κυρίαρχης κλάσης εμφανίζεται σε τουλάχιστον ένα σύνολο εκπαίδευσης έτσι ώστε κανένα δεδομένο να μην πάει χαμένο. Ο αλγόριθμος εκμάθησης εφαρμόζεται σε κάθε σύνολο εκπαίδευσης και η μετά-εκπαίδευση χρησιμοποιείται για το σχηματισμό ενός σύνθετου αλγορίθμου εκπαίδευσης από τους ταξινομητές που προκύπτουν ως αποτέλεσμα.

Ο Molinara (2007) παρουσιασε μια ενδιαφερουσα συνδιαστικη προσεγγιση. Για να κατασκευαστει ένα συνολο ταξινομητών, δοκιμασε δυο στρατηγικες διαχωρισμου, οι οποίες βασίζονται στην ομαδοποίηση και στην τυχαία επιλογή των δειγμάτων της κατηγορίας πλειοψηφίας. Όπως οι βασικοί ταξινομητές έχουν υιοθετήσει μια διαδικασία ενίσχυσης (boosting), ενώ δυναμική επιλογή, σημαίνει και λήψη κυρίαρχης νίκης που έχει χρησιμοποιηθεί ως κριτήριο την συσσωμάτωση ατομικών αποφάσεων. Αυτή η προσέγγιση έχει εφαρμοστεί για την ανίχνευση μικρό-ταξινομήσεων. Αυτή η προσέγγιση έχει εφαρμοστεί για την ανίχνευση μικρό-ταξινομήσεων (micro- classifications) σε ένα σύνολο δεδομένων των ψηφιακών μαστογραφιών διαθέσιμες στο κοινό. Τα αποτελέσματα , που αναφέρθηκαν σε όρους της ακρίβειας σε θετικά και αρνητικά δείγματα , έδειξαν ότι οι πρώτες αυξήθηκαν , ενώ οι δευτέρες μειωθήκαν. Καλύτερη απόδοση επιτυγχάνεται όταν το σύνολο αποτελείται από έναν αριθμό βασικών ταξινομητών ίσο με την αναλογία μεταξύ του αριθμού της πλειοψηφίας και της μειοψηφίας κλάσης δειγμάτων.

Ο Sung Chiang-Lin και άλλοι (2009) προτείνουν το σύνολο αλγορίθμων ανομοιογενών με μετά-δεδομένη ταξινόμηση (Meta Imbalanced Classification Ensemble-MICE), προκειμένου να μετριάσει η επίδραση ισορροπημένων δεδομένων. Στο MICE , η πλειοψηφική ομάδα κατανέμεται

με βάση τα μετασχηματισμένα χαρακτηριστικά από το ¨εσωτερικό γινόμενο¨ για να διατηρηθεί η γεωμετρική σχέση μεταξύ των δυο ομάδων.

Μια παρόμοια συνδυαστική προσέγγιση έχει προταθεί πρόσφατα (Soda, 2009). Είναι βασισμένο σε ένα σύνολο ταξινομητών εκπαιδευμένα σχετικά με τα ομοιογενή υποσύνολα του αρχικού συνόλου εκπαίδευσης ανομοιογένειας που δουλεύουν σε συνδυασμό με τον εκπαιδευόμενο ταξινομητή στο αρχικό σύνολο ανομοιογενών δεδομένων. Ο Cleofas και άλλοι (2009) πρότειναν δύο σύνολα μοντέλων (χρησιμοποιώντας ένα σπονδυλωτό νευρωνικό δίκτυο και τον κανόνα του πλησιέστερου γείτονα), που έχουν εκπαιδευτεί σε σύνολα δεδομένων υπό-δειγματοληψίας με γενετικούς αλγορίθμους.

Οι αλγόριθμοι μάθησης συνόλων περιλαμβάνουν τον αλγόριθμο ¨*Σακουλιάσματος*¨ (Bagging algorithm) (Breitman, 1996-1999) και τον αλγόριθμο ¨*Ενίσχυσης*¨ (Boosting algorithm) (Schapire, 1990). Στην οικογένεια του αλγορίθμου Ενίσχυσης ανήκουν οι αλγόριθμοι AdaBoost (Freund and Schapire, 1997; Wu et al., 2007) και τροποποιημένοι AdaBoost.M1 και AdaBoost.M2 (Freund and Schapire, 1997; Schapire and Singer, 2001).

Οι οικογένεια αλγορίθμων ενίσχυσης (Boosting algorithms) περιλαμβάνει τις ακόλουθες κατηγορίες αλγορίθμων:

A*) Οι αλγόριθμοι ενίσχυσης με ευαισθησία κόστους* (Cost sensitive Boosting algorithms), που περιλαμβάνουν τις ακόλουθες αλγοριθμικές διαδικασίες:

A1) AdaCost (Fan et al., 1999)
A2) Cost Sensitive Boosting (CSB)
A3) AdaC1 (Sun et al., 2007)
A4) AdaC2 και AdaC3

B) Οι *αλγόριθμοι συνόλων*, που βασίζονται σε ενίσχυση (Boosting based Ensembles), περιέχουν τις αλγοριθμικές διαδικασίες:

B1) SMOTEBoost (Chawla et al., 2003) και MSMOTEBoost (Hu et al., 2009)
B2) RUSBoost (Seifert et al., 2010)

B3) DataBoost-IM (Guo and Viktor, 2004)

C) Οι *αλγόριθμοι συνόλων*, που βασίζονται σε ¨σακούλιασμα¨ (Bagging based Ensembles), περιέχουν τις αλγοριθμικές διαδικασίες:

C1) OverBagging (Wang and Yao, 2009)
C2) UnderBagging (Barandela et al., 2003)
C3) UnderOverBagging (Wang and Yao, 2009)
C4) IIVotes (Blaszczynski et al., 2010)

D) *Υβριδικά (Μεικτογενή) Σύνολα* (Hybrid Ensembles), που συνδυάζουν αλγορίθμους ενίσχυσης και σακουλιάσματος με τεχνικές προ-επεξεργασίας, περιέχουν

D1) EasyEnsemble (Liu et al., 2009)
D2) BalanceCascade (Liu et al., 2009).

3.4.2 Αλγοριθμικές Μέθοδοι Μάθησης Συνόλων

Τα σύνολα (ensembles) έχουν σχεδιαστεί για την αύξηση της ακρίβειας ενός απλού ταξινομητή με δοκιμή (training) διάφορων ταξινομητών και συνδυάζοντας τις αποφάσεις τους για να δώσουν στην έξοδο μια απλή ετικέτα τάξης (single class label) (Polikar, 2006; Rokach, 2010).

Ο όρος ¨*μέθοδοι συνόλων*¨ (ensemble methods) χρησιμοποιείται στην μηχανική μάθηση και αναφέρεται συνήθως σε συλλογές ταξινομητών που είναι μικρές παραλλαγές του ίδιου ταξινομητή, ενώ ο όρος ¨**συστήματα πολλαπλών ταξινομητών**¨ (multiple classifier systems) αποτελεί μια ευρύτερη κατηγορία η οποία περιλαμβάνει συνδυασμούς που χρησιμοποιούν υβριδικούς συνδυασμούς διαφορετικών μοντέλων (Ho et al., 1994; Ho, 2002). Ο σχηματισμός συνόλων με δημιουργία διαφοροποιημένων ταξινομητών (διατηρώντας τη συνύπαρξη τους με το σύνολο δοκιμών [training set]) αποτελεί βασικό παράγοντα για να γίνουν ακριβείς. Σημειώνεται ότι η διαφοροποίηση (diversity) στα σύνολα έχει ένα θεωρητικό υπόβαθρο σε προβλήματα

παλινδρόμησης (regression) (Ueda and Nakano, 1996; Krogh and Vedelsby, 1995), ενώ η έννοια της διαφοροποίησης στην ταξινόμηση αναφέρεται ότι δεν είναι ακόμη σαφώς ορισμένη (Brown et al., 2005).

Οι αλγόριθμοι μάθησης συνόλων (ensemble learning algorithms) περιλαμβάνουν τους αλγορίθμους AdaBoost (Schapire, 1990; Freund and Schapire, 1997), Bagging (Breiman, 1996) και αρκετούς σχετικούς τροποποιημένους αλγορίθμους (Kuncheva, 2004). Οι τροποποιημένοι αλγόριθμοι μάθησης συνόλων συνήθως περιλαμβάνουν προσεγγίσεις επιπέδου δεδομένων για να προ-επεξεργαστούν τα δεδομένα πριν από τη μάθηση κάθε ταξινομητή (Chawla et al. 2003; Seiffert et al., 2010; Blaszczynski et al., 2010; Liu et al., 2009). Εναλλακτικές προσεγγίσεις για τους αλγορίθμους μάθησης συνόλων περιλαμβάνουν τους ακολούθους αλγορίθμους:

SMOTEBoost (Chawla et al., 2003)

RUSBoost (Seiffert et al., 2010)

SMOTE-Bagging (Wang and Yao, 2009)

EasyEnsemble (Liu et al., 2009)

IIVotes (Blaszczynski et al., 2010)

Οι αλγόριθμοι αυτοί , αν και μπορούν να χρησιμοποιηθούν για την επίλυση του προβλήματος ανομοιογενών τάξεων, έχουν χρησιμοποιηθεί σε περιορισμένο αριθμό προβλημάτων, δεν έχουν συγκριθεί οι αποδόσεις τους και γενικά υπάρχει έλλειψη ενός ενοποιημένου πλαισίου για την κατηγοριοποίηση τους (Seiffert et al., 2010; Liu et al., 2009).

Στην ταξινόμηση ένα ανομοιογενές σύνολο δεδομένων ορίζεται όταν ο αριθμός στιγμιότυπων που αντιπροσωπεύει μια τάξη είναι μικρότερος από τους αντίστοιχους αριθμούς άλλων τάξεων. Η τάξη με το μικρότερο αριθμό στιγμιότυπων είναι συνήθως η πλέον ενδιαφέρουσα από την άποψη μάθησης (Chawla et al., 2004) και το πρόβλημα εμφανίζεται σε σημαντικές εφαρμογές , όπως ιατρικές διαγνώσεις (Mazurowski et al., 2008; Freitas et al., 2007; Kilic et al., 2007; Celebi et al., 2007; Peng, 2008), ανίχνευση ρύπανσης (Lu and Wang, 2008), διαχείριση επικινδυνότητας (risk management) (Huang et al., 2006), ανίχνευση απάτης (Cieslak, 2006) κ.α.

3.4.3 Αλγόριθμος Εύκολων Συνόλων (Easy Ensemble) και Αλγόριθμος Ισορροπημένων Διαδοχικών Συνόλων (Balance Cascade)

Οι μέθοδοι συνόλων (ensemble) χρήση πολλαπλών μοντέλων (ensemble methods using multiple models) χρησιμοποιώντας στατιστικές μεθόδους και μεθόδους μηχανικής μάθησης οδηγούν στην απόκτηση καλύτερης απόδοσης πρόβλεψης από προβλέψεις που θα μπορούσαν να προκύψουν από συνθετικά μοντέλα. Σε αντίθεση με ένα στατιστικό σύνολο στη στατιστική μηχανική, η μηχανική μάθηση συνόλου αναφέρεται μόνο σε ένα συγκεκριμένο πεπερασμένο σύνολο εναλλακτικών μοντέλων, αλλά συνήθως επιτρέπει την πιο ευέλικτη δομή να υπάρχει μεταξύ αυτών των εναλλακτικών λύσεων.

Οι *αλγόριθμοι μάθησης με επίβλεψη* συνήθως περιγράφονται ως εκπλήρωση της αποστολής αναζήτησης μέσα από ένα υποθετικό χώρο για να βρουν μια κατάλληλη υπόθεση που θα κάνουν καλή πρόβλεψη με ένα συγκεκριμένο πρόβλημα. Ακόμη και αν ο χώρος της υπόθεσης περιέχει υποθέσεις που είναι ιδιαίτερα κατάλληλες για ένα συγκεκριμένο πρόβλημα, μπορεί να είναι πολύ δύσκολο να προκύψει ένα καλό αποτέλεσμα. Ορισμένα σύνολα μπορούν να συνδυάζουν πολλαπλές υποθέσεις για να σχηματίσουν μια καλύτερη υπόθεση. Με άλλα λόγια, ένα σύνολο είναι μια τεχνική που συνδυάζει πολλούς αδύναμους εκπαιδευόμενους (weak learners) σε μια προσπάθεια να παράγει ένα ισχυρό εκπαιδευόμενο. Ο ορός σύνολο προορίζεται συνήθως για τις μεθόδους που δημιουργούν πολλαπλές υποθέσεις με τον ίδιο βασικό εκπαιδευόμενο. Ο ευρύτερος όρος των πολλαπλών συστημάτων ταξινομητή καλύπτει επίσης υβριδοποίηση των υποθέσεων που δεν προκαλείται από τον ίδιο βασικό εκπαιδευόμενο.

Αξιολογώντας την πρόβλεψη ενός συνόλου συνήθως απαιτεί περισσότερο από ότι την αξιολόγηση υπολογισμού την πρόβλεψη ενός ενιαίου μοντέλου, έτσι τα σύνολα μπορούν να θεωρηθούν ως ένας τρόπος για την αντιστάθμιση των αλγορίθμων κακή εκμάθηση εκτελώντας πολλών επιπλέον υπολογισμών. Οι ταχείς αλγόριθμοι, όπως δέντρα απόφασης που χρησιμοποιούνται συνήθως με σύνολα (π.χ. Random Forest), αν και οι αλγόριθμοι με μέτρια ταχύτητα εκτέλεσης μπορούν να επωφεληθούν από τις τεχνικές συνόλων.

Ο αλγόριθμος ισορροπημένων *διαδοχικών συνόλων* (Balance cascade algorithm) διευρύνει την κυρίαρχη κλάση με επιβλεπόμενο τρόπο, όπου η easy ensemble μαθαίνει διαφορετικές πτυχές

της αρχικής κυρίαρχης κλάσης με επιβλεπόμενο τρόπο. Στον αλγόριθμο αυτό η διαδικασία όπου προσπαθούμε να απομακρύνουμε παραδείγματα από την κυρίαρχη κλάση μέχρι κανέναν να μην είναι μη ταξινομημένο (miss-classified). Αν έχουμε αν ανομοιογενή σύνολο δεδομένων εκπαίδευσης, προσπαθούμε να ισορροπήσουμε την κυρίαρχη κλάση και την κλάση μειοψηφίας. Έτσι δημιουργείται το μοντέλο, και μαζί με την κλάση κυριαρχίας των αρχικών ανομοιογενών συνόλων δεδομένων εκπαίδευσης ταξινομούνται στην κλάση κυριαρχίας παρατήρησης που είχαν μείνει εκτός. Στην συνέχεια απομακρύνονται τα σωστά ταξινομημένα παραδείγματα και έτσι προκύπτει το νέο σύνολο εκπαίδευσης.

3.4.4 Αλγόριθμος AdaBoost

Ο ¨**αλγόριθμος AdaBoost**¨, συντομογραφία του Adaptive Boosting, είναι ένα μετά-αλγόριθμος μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί με άλλους αλγορίθμους μάθησης για να βελτιώσει την απόδοση τους. Ο αλγόριθμος αυτός είναι ευαίσθητος σε θορυβώδη δεδομένα και προσαρμόσιμος με την έννοια ότι διαδοχικοί ταξινομητές που κατασκευάζονται είναι προσανατολισμένοι προς στιγμιότυπα με εσφαλμένη ταξινόμηση από προηγούμενους ταξινομητές. Οι ταξινομητές που χρησιμοποιεί μπορεί να είναι ¨ασθενής¨ (δηλαδή να δείχνουν μια ουσιαστική τιμή σφάλματος), αλλά εφόσον η απόδοση τους δεν είναι τυχαία αναμένεται να βελτιώσουν το τελικό μοντέλο.

Ο αλγόριθμος AdaBoost δημιουργεί και καλεί ένα νέο ασθενή ταξινομητή σε κάθε μια σειρά τιμών t=1,2,…..,T. Για κάθε κλήση, μια κατανομή βαρών Dt ενημερώνεται έτσι ώστε να δείχνει τη σημασία των παραδειγμάτων στο σύνολο δεδομένων για την ταξινόμηση. Σε κάθε σειρά τα βάρη κάθε παραδείγματος εσφαλμένης ταξινόμησης αυξάνονται και τα βάρη κάθε σωστής ταξινόμησης μειώνονται, έτσι ώστε ο νέος ταξινομητής εστιάζεται σε παραδείγματα που έχουν αποφύγει σωστή ταξινόμηση (Fan et al.,1999; Freud and Schapire, 1997).

Ο αλγόριθμος AdaBoost που χρησιμοποιεί το ελάχιστο αναμενόμενο κριτήριο κόστους δίνεται από την ακόλουθη αλγοριθμική διαδικασία (Ting, 2000)

Υποθέστε ότι T είναι ένα σύνολο δοκιμών που περιέχει N- παραδείγματα (xn, yn) και έστω ότι L είναι μια βάση αλγορίθμων μάθησης και cost είναι ένας πινάκας κόστους. Ο αλγόριθμος AdaBoost δίνεται από την ακόλουθη διαδικασία (Ting, 2000):

Αλγόριθμος AdaBoost (T, L, Cost, K, H*)

Είσοδος: T (σύνολο δοκιμών που περιέχει N-παραδείγματα), N, L (βάση αλγορίθμων μάθησης), Cost (πίνακας κόστους)

Έξοδος: H*, $P(i/x)\log(F_k).h_k^i(x)$

Υπολογιστική Διαδικασία:

Αρχικοποίηση: θεωρείστε όλα τα βάρη στιγμιότυπων $w_1(n)=1$

Για $k = 1,\ldots,K$

(i) θεωρήστε ένα μοντέλο h_k εφαρμόζοντας L σε T υπό την κατανομή βαρών w_k. Σημειώστε ότι $h_k(x)$ δηλώνει την προβλεπόμενη τάξη και $h_k^i(x)[0, 1]$

δηλώνει το επίπεδο εμπιστοσύνης της πρόβλεψης για την τάξη i.

(ii) το T ταξινομείται χρησιμοποιώντας h_k. Το σφάλμα ε_k του μοντέλου ορίζεται ως

$$\varepsilon_k = \left(\sum_{\substack{xn \in T \\ hk(xn) \neq yn}} w_k(n) \right) / N.$$

(iii) το στιγμιότυπο βάρους $w_{(k+1)}$ δημιουργείται από το w_k με τον ακόλουθο τρόπο:

$$w_{(k+1)}(n) = \begin{cases} w_{(k)}(n) \cdot F_k, & \text{αν } h_k(x_n) y_n \\ w_{(k)}(n) / F_k, & \text{διαφορετικά} \end{cases}$$

$$\text{όπου } F_k = \left[(1-\varepsilon_k) / \varepsilon_k \right]^{1/2}.$$

$$H^*(x) = \arg\min_j \sum_i \sum_k \log F(k) \cdot h_k^i(x).\text{Cost}(i, j).$$

Εικόνα 3.4: Αλγόριθμος AdaBoost

3.4.5 Αλγόριθμος AdaCost

Ο ¨*αλγόριθμος* ***AdaCost***¨, μια τροποποίηση του αλγορίθμου AdamBoost (Feund and Schapire, 1997), είναι μια εσφαλμένης ταξινόμησης ενδυναμωμένη μέθοδος με ευαισθησία κόστους, που χρησιμοποιεί το κόστος των εσφαλμένων ταξινομήσεων για να ενημερώσει τη δοκιμαστική κατανομή σε διαδοχικές επαναλήψεις με αντικειμενικό σκοπό να ελαττώσει το άνω όριο του συσσωρευμένου κόστους με εσφαλμένο κόστος του δοκιμαστικού συνόλου (Fan et al., 1999).

Ο αλγόριθμος AdaCost (Fan et al., 1999) παρουσιάζει μια συνάρτηση καθορισμού κόστους που είναι ενσωματωμένη σε ένα κανόνα ενημερωμένων βαρών, εκχωρεί υψηλές αρχικές βάρη σε δαπανηρά στιγμιότυπα και ο κανόνας με ενημερωμένα βάρη θεωρεί το κόστος και αυξάνει τα βάρη των δαπανηρών εσφαλμένων ταξινομήσεων. Στην συνέχεια παρουσιάζεται ο αλγόριθμος AdaCost σε μορφή ψευδοκώδικα:

Αλγόριθμος AdaCost (I,T,S,D,a,b)

Είσοδος: I (α παρακινητής βάσης), T (αριθμός επαναλήψεων)

$$S = \{(x_1, c_1, y_1) \ldots (x_m, c_m, y_m)\},\ x_i X,\ c_i R^+,\ y_i \{-1,\ +1\}$$

Έξοδος: D, α ,β

Υπολογιστική Διαδικασία:

Βήμα 1: αρχικοποίησε $D_1(i)$ (T: τέτοιο ώστε $D_1(i) = c_i / \sum_j^m c_j$)

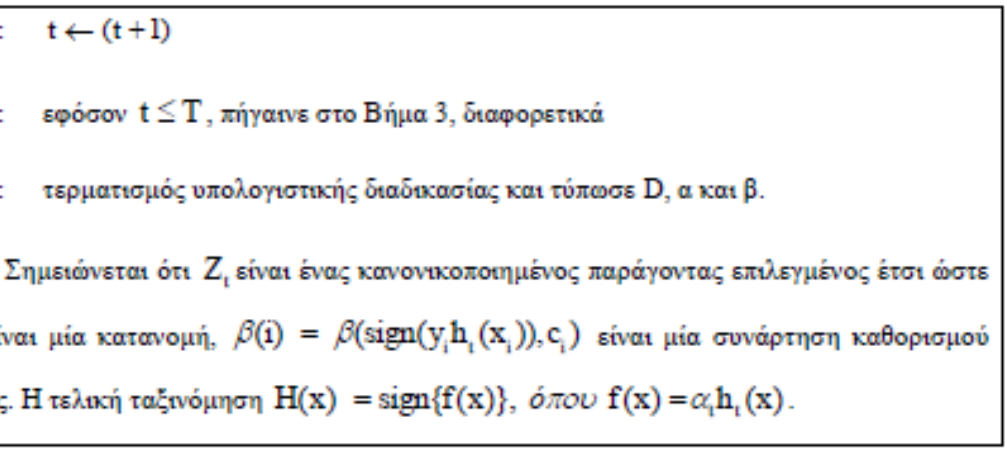
Βήμα 2: επανέλαβε τα ακόλουθα:

Βήμα 3: θεώρησε τον 'ασθενή παρακινητή' (weak inducer) χρησιμοποιώντας την κατανομή D_t

Βήμα 4: υπολόγισε τον ασθενή ταξινομητή $h : X \to R$

Βήμα 5: επέλεξε $\alpha_t \in R$ και $\beta(i) \in R^+$

Βήμα 6: ενημέρωσε $D_{t+1}(i) = D_t(i) \exp[-\alpha_t y_i h_t(x_i) \beta(i)] / Z_t$



Εικόνα 3.5: Αλγόριθμος AdaCost

3.4.6 Αλγόριθμος ενίσχυσης ευαισθησίας κόστους (Cost Sensitive boosting)

Η στρατηγική ρύθμισης βάρους (weighting strategy) των AdaBoost είναι να αυξήσει τα βάρη των μη ταξινομημένων δειγμάτων και να μειώσει τα βάρη των σωστά ταξινομημένων δειγμάτων μέχρι ότου οι κατανομές δείγματος με βάρος μεταξύ των μη ταξινομημένων δειγμάτων και των σωστά ταξινομημένων δειγμάτων είναι παρόμοιες σε κάθε γύρο. Η στρατηγική ρύθμισης βάρους, διακρίνει τα δείγματα στις εξόδους της ταξινόμηση τους ως αποτέλεσμα να ταξινομούνται σωστά ή όχι.

Η στρατηγική αυτή όμως, αντιμετωπίζει δείγματα από διάφορες κλάσεις το ίδιο, δηλαδή τα βάρη των μη ταξινομημένων δειγμάτων από διάφορες κλάσεις αυξήθηκαν από ένα πανομοιότυπο λόγο, και τα βάρη από την σωστή ταξινόμηση δειγμάτων από διαφορετικές κλάσεις μειωθήκαν από άλλο πανομοιότυπο λόγο. Ο μαθησιακός στόχος στην αντιμετώπιση του προβλήματος της ανομοιογενής κλάσης είναι να βελτιώσει την απόδοση αναγνώρισης σε μικρές κλάσεις.

Ο μαθησιακός στόχος αυτός αναμένει ότι η στρατηγική ρύθμισης βάρους ενός αλγορίθμου ενίσχυσης (boosting algorithm) θα διατηρήσει ένα σημαντικό μέγεθος δείγματος με βάρη της μικρής κλάσης. Μια επιθυμητή στρατηγική ενίσχυσης με την οποία είναι σε θέση να διακρίνει διαφορετικούς τύπους δειγμάτων, και να ενισχύσει περισσότερα βάρη σε αυτά τα δείγματα που συνδέονται με υψηλότερη σημασία ταυτοποίησης.

Σχετικά με το κόστος ευαισθησίας στα δέντρα απόφασης είναι βασισμένο στον αλγόριθμο C4.5 , που έχουμε περιγράψει σε παραπάνω κεφάλαιο. Όταν χρίζεις ένα δέντρο απόφασης, σε κάθε

βήμα, αντί να διαλέγεις ένα χαρακτηριστικό ελαχιστοποιεί την εντροπία (όπως ο αλγόριθμος C4.5) , το κόστος ευαισθησίας των δέντρων απόφασης διαλεγόμενα χαρακτηριστικό που μειώνει και ελαχιστοποιεί το συνολικό κόστος των μη ταξινομημένων χαρακτηριστικών (misclassification cost) και το κόστος έλεγχου (test cost), για το διαμοιρασμό. Παρόμοια με το αλγόριθμο C4.5 , το κόστος ευαισθησίας των δένδρων απόφασης επιλέγει ένα τοπικό βέλτιστο χαρακτηριστικό.

Κατασκευάζει ένα δέντρο απόφασης ευαίσθητο στο κόστος, οι αποφάσεις με τις όποιες τα χαρακτηριστικά διαχωρίζονται, μπορούν να προσδιοριστούν υπολογίζοντας το μη ταξινομημένο κόστος για κάθε πιθανό διαχωρισμό, διαλέγοντας αυτό με το μικρότερο κόστος. Ο Elkan (2001) έδειξε ότι αυτή η προσέγγιση οδηγεί σε ένα ταξινομημένο μοντέλο όπου ελαχιστοποιεί το κόστος των μη ταξινομημένων χαρακτηριστικών του συνόλου εκπαίδευσης , αλλά δεν παράγει ένα βέλτιστο μοντέλο όταν εφαρμόζονται δεδομένα, τα οποία δεν είναι ορατά (unseen data), λόγω του υπερταιριάσματος.

3.4.7 Αλγόριθμος MetaCost

Τις τελευταίες δεκαετίες έχουν αναπτυχτεί σημαντικά νέοι αλγόριθμοι ταξινόμησης (classifiers) για ανομοιογενή δεδομένα που χρησιμοποιούνται στη Μηχανική Μάθηση, Στατιστική και αλλά σχετικά πεδία και εφαρμογές. Οι περισσότεροι από τους αλγορίθμους αυτούς υποθέτουν ότι όλα τα σφάλματα έχουν το ίδιο κόστος και η κατασκευή κάθε τέτοιου ταξινομητή με ευαισθησία κόστους αποτελεί εργώδη και συχνά μη-τετριμμένη εργασία.

Ο ¨*αλγόριθμος MetaCost*¨ αποτελεί μια διαδικασία που χρησιμοποιεί ένα αυθαίρετο ταξινομητή με ευαισθησία κόστους κατασκευάζοντας μια προσαρμοσμένη διαδικασία ελαχιστοποίησης κόστους. Η *συνδυαστική μέθοδος MetaCost* (Domingos, 1998) είναι μια άλλη μέθοδος για την κατασκευή ενός ταξινομητή με κόστος ευαισθησίας (cost-sensitive). Η διαδικασία ξεκινά με την εκπαίδευση ενός μοντέλου με ευαισθησία κόστους, εφαρμόζοντας μια οικονομικά ευαίσθητη διαδικασία, η οποία χρησιμοποιεί έναν αλγόριθμο μάθησης βάσης. Στη συνέχεια , η μέθοδος MetaCost εκτιμά τις πιθανότητες της κλάσης με χρήση της μεθόδου σακουλιάσματος (Bagging), στη συνέχεια επαναετικετοποιεί (relabels) τα εκπαιδευμένα στιγμιότυπα με ελάχιστες αναμενόμενες κλάσεις του κόστους και τέλος επανεκπαιδεύεται ένα μοντέλο χρησιμοποιώντας το τροποποιημένο σύνολο εκπαίδευσης. Ο ταξινομητής αυτός χρησιμοποιείται ως black-box, χωρίς

γνώση της λειτουργικότητας του ή αλλαγών που επιτελούνται, και εφαρμόζεται για κάθε αριθμό τάξεων και για αυθαίρετους πίνακες κόστους. Ο αλγόριθμος MetaCost σε ορισμένες περιπτώσεις μπορεί να παράγει μεγάλες μειώσεις κόστους σε σύγκριση με τον ταξινομητή C4.5 (Domingos, 1996).

Η ταξινόμηση θεωρείται σημαντικός παράγων στο πεδίο εξόρυξης δεδομένων (Data Mining) και αποτελεί ενδιαφέρον ερευνητικό θέμα σε επιστημονικά πεδία , όπως μηχανική μάθηση, αναγνώριση σχεδίων, νευρωνικά δίκτυα, στατιστική και αλλά σχετικές περιοχές.

Σύγχρονες προσεγγίσεις για τους αλγορίθμους ταξινόμησης περιλαμβάνουν (i) επαγωγικούς κανόνες (Michalski, 1983; Domingos, 1996), (ii) επαγωγικά δένδρα αποφάσεων (Breiman et al., 1984; Quilan, 1993), (iii) μάθηση βασισμένη σε στιγμιότυπα (Dasarathy, 1991; Aha et al., 1991), (iv) γραμμικούς και νευρωνικούς ταξινομητές (Bishop, 1995), (v) μάθηση Bays (Domingos and Pazzani, 1997; Duda and Hart, 1973) κ.α.

Το κόστος μη-ταξινόμησης (misclassification) μπορεί να περιγράφει ως ένας αυθαίρετος πινάκας κόστους C(i,j) όπου C είναι το κόστος για την πρόβλεψη ότι ένα παράδειγμα ανήκει σε μια τάξη i όταν πράγματι ανήκει στην τάξη j. Προσπάθειες έχουν καταβληθεί για τη δημιουργία αλγορίθμων με ευαισθησία κόστους με τη μετατροπή ταξινομητών που βασίζονται σε σφάλματα σε αντίστοιχους ταξινομητές με ευαισθησία κόστους (Tuney, 1997; Breiman et al., 1984; Provost et al., 1998).

Για την κατασκευή ταξινομητών με κόστος ευαισθησία θεωρείται ένα παράδειγμα x με πιθανότητα P(j/x) για κάθε τάξη j. Η βέλτιστη πρόβλεψη Bayes για το x είναι η τάξη I που ελαχιστοποιεί την ¨υποθετική επικινδυνότητα¨ (conditional risk) (Duda and Hart, 1973):

$$R(i/x)=\sum P(j/x).\ C(i,j). \qquad (5.1)$$

Ο αλγόριθμος MetaCost βασίζεται στην επανα-ετικετοποίηση δοκιμαστικών παραδειγμάτων με υπολογισμένες τάξεις, ελαχιστοποιημένες ως προς το κόστος και στην εφαρμογή της μάθησης που βασίζεται σε σφάλματα σε νέα δοκιμαστικά σύνολα. Ο αλγόριθμος αυτός μπορεί να εφαρμοστεί σε μεγάλες βάσεις δεδομένων (Domingos, 1999).

Ο αλγόριθμος MetaCost αυξάνει το χρόνο μάθησης κατά ένα σταθερό παράγοντα (περίπου ίσο προς αριθμό των επαναληπτικών δειγμάτων [resamples]) σε σύγκριση με τον ταξινομητή που

βασίζεται σε σφάλμα. Ένας τρόπος αντιμετώπισης είναι να παραλληλοποιήσουμε τις πολλαπλές εκτελέσεις του ταξινομητή που βασίζεται σε σφάλματα καθώς επίσης να επιλέξουμε τα επαναληπτικά δείγματα που είναι μικρότερα από τα αρχικά δοκιμαστικά σύνολα.

Τα πλεονεκτήματα του αλγορίθμου MetaCost περιλαμβάνουν τα ακόλουθα:

- Ο MetaCost εφαρμόζεται σε κάθε αριθμό τάξεων και για αυθαίρετους πίνακες κόστους και παράγει μεγαλύτερες μειώσεις κόστους σε σύγκριση με ταξινομητές άνευ- κόστους (cost blind)
- Καθιερώνει ένα πρότυπο για σύγκριση αλγορίθμων, χρησιμοποιώντας C4.5 με υπο-δείγματα
- Η υποδειγματοληψία παράγει μια ευαισθησία σε αλλαγές στο κόστος εσφαλμένης ταξινόμησης (misclassification) και κατανομή τάξεων
- Η υπερδειγματοληψία δείχνει μικρή ευαισθησία και υπάρχει συχνά μικρή διαφορά στην απόδοση όταν το κόστος εσφαλμένης ταξινόμησης μεταβάλλεται. Μπορεί να γίνει επίσης ευαίσθητη κόστους αν χρησιμοποιηθούν κατάλληλες παράμετροι τερματισμού σε αναλογία με το ποσό της υπερδειγματοληψίας.
- Το επιπρόσθετο υπολογιστικό κόστος από τη χρησιμοποίηση υπερδειγματοληψίας είναι αδικαιολόγητο καθώς επιτυγχάνεται η απόδοση να είναι περίπου η ίδια με την περίπτωση της υποδειγματοληψίας.

Αλγόριθμος Metacost (S,L,C,m,n,p,q)

Είσοδος:εκμάθηση συνόλου (ensemble) S, εκμάθηση ταξινομητή L, πίνακας κόστους C, πλήθος της επαναδειγματολειψίας (resample) m, πλήθος των παραδειγμάτων σε κάθε επαναδειγματολειψία n, p είναι αληθής αν και μόνο αν η L παράγει μια κλάση από πιθανότητες, q είναι αλυθής αν και μόνο αν όλες οι επαναδειγματοληψίες χρησιμοποιούνται σε κάθε παράδειγμα.

Έξοδος: x

Υπολογιστική Διαδικασία:

Για i=1, m

Θέσε SI για επαναδειγματοληψία S με n παραδείγματα

θέσε Mi = μοντέλο κατασκευασμένο εφαρμόζοντας L στο S,

για κάθε παράδειγμα x στο S

για κάθε κλάση σύνολο j

$$P(j / x) = \frac{1}{\sum_i 1} \sum_i P(j / x, M_i)$$

,

Όπου αν p τότε P(j / x, Mi) παράγεται από Mi .

Τότε P(j / x, Mi) =1 για κάθε κλάση που προκύπτουν από Mi για x, και 0 για όλες τις άλλες.

Αν q τότε η κλάση i επι όλων Mi

τότε η κλάση i επι όλων Mi τέτοια ώστε $x \in S_i$

$$x = \arg\min_i \sum_j P(j / x) C(i, j)$$

θέσε την κλάση του ,

θέσε M= μοντέλο παραγωγής από την εφαρμογή του L στο S,

επιστροφή στο M.

Εικόνα 3.6: Αλγόριθμος Metacost

Σημειώνεται ότι ο αλγόριθμος MetaCost είναι ένας μετά-ταξινομητής που παράγει παρόμοια αποτελέσματα με εκείνα που δημιουργήθηκαν με το πέρασμα του βασικού ταξινομητή μάθησης (bagging) που με τη σειρά του δημιουργήθηκε από τον ταξινομητή CostSensitive που λειτούργει με ελάχιστο κόστος. Η εφαρμογή αυτή μπορεί να χρησιμοποιηθεί όλες τις επαναλήψεις bagging όταν τα δεδομένα μάθησης επαναταξινομούνται (Domingos, 1999).

3.4.8 Αλγόριθμος Rotation Forest

Ο ¨***αλγόριθμος Rotation Forest***¨ είναι μια νέα μέθοδος με την οποία κάποιος μπορεί να δημιουργήσει σύνολα ταξινομητή χρησιμοποιώντας ανεξάρτητα εκπαιδευμένα δέντρα απόφασης (Rodriguez, Kuncheva, 2006). Σύμφωνα με σχετικές έρευνες του Rodriguez και Kuncheva, διαπιστώθηκε ότι ο αλγόριθμος Rotantion Forest είναι πιο ακριβής από τους αλγορίθμους Bagging, Adaboost και Random Forest σε μια συλλογή δεδομένων. Τα καλύτερα αποτελέσματα βρεθήκαν , έχοντας τα χαρακτηριστικά γνωρίσματα που εξάγονται μέσω των PCA (principal component analysis) σε σύγκριση με εκείνα που προέρχονται από μη παραμετρική διακριτή ανάλυση (non-parametric discriminant analysis-NDA) ή τυχαίες προβλέψεις. Ο κύριος παράγοντας για την επιτυχία του αλγορίθμου Rotation forest είναι ότι ο πίνακας μετασχηματισμού που χρησιμοποιήθηκε για τον υπολογισμό των γραμμικών χαρακτηριστικών τα οποία εξάγονται, είναι αραιός.

Ο αλγόριθμος Rotation Forest συνδυάζει τις προβλέψεις πολλαπλών ταξινομητών αντί του ενός ταξινομητή και παρατηρείται μείωση της διασποράς και της πόλωσης. Επίσης εξαρτάται λιγότερο από τις ιδιαιτερότητες ενός συνόλου εκπαίδευσης και παρέχει μια πλέον κατανοητή έννοια τάξης από έναν ταξινομητή.

Αλγόριθμος Rotation Forest (X,Y,L,K,W)

Είσοδος: X είναι το αντικείμενο στα δεδομένα εκπαίδευσης που καθορίζονται σε έναν (Nxn) πίνακα, Y είναι οι ετικέτες των δεδομενων εκπαίδευσης σε ένα (Nx1) πίνακα, L είναι

ο αριθμός των ταξινομητών στο σύνολο, K είναι ο αριθμός των υποσυνόλων, (w1, …, wc) είναι το σύνολο των ετικετών της κλάσης

Έξοδος: D_i είναι ταξινομητές χρησιμοποιώντας $(XR_i^a Y)$ σαν σύνολο εκπαίδευσης

Υπολογιστική διαδικασία

Για $i = 1,...,L$

Προετοιμασία του πίνακα περιστροφής (rotation matrix) R_i^a :

Διαιρώ το F (το σύνολο των χαρακτηριστικών) σε K υποσύνολα: $F_{i,j}$ (για j=1,…,K)

Για j=1,…, K

Έστω $X_{i,j}$ είναι το σύνολο των δεδομένων X του χαρακτηριστικού σε $F_{i,j}$

Εξάλειψη από $X_{i,j}$ ένα τυχαίο υποσύνολο $F_{i,j}$ των κλάσεων

Επιλέξτε ένα bootstrap δείγμα από $X_{i,j}$ του μεγέθους 75% τον αριθμό των αντικειμένων σε $X_{i,j}$. Δείξτε το νέο σύνολο από $X'_{i,j}$

Εφάρμοσε PCA στο $X'_{i,j}$ για να ληφθούν οι συντελεστές σε ένα πίνακα $C_{i,j}$

Ταξινόμησε τον $C_{i,j}$, για j=1,…,K στον πίνακα περιστροφής R_i

$$R_i = \begin{bmatrix} a_{i,1}^{(1)}, a_{i,1}^{(2)}, ..., a_{i,1}^{(M_1)}, & [0] & ... & [0] \\ [0] & a_{i,2}^{(1)}, a_{i,2}^{(2)}, ..., a_{i,2}^{(M_2)} & ... & [0] \\ . & . & & .. \\ . & . & & \\ [0] & [0] & ... & a_{i,K}^{(1)}, a_{i,K}^{(2)}, ..., a_{i,K}^{(M_K)} \end{bmatrix}$$

Κατασκευάζεται R_i^a με αναδιάταξη των στηλών του R_i έτσι ώστε να ταιριάζει με τη σειρά των χαρακτηριστικών στο F.

-κατασκευάζεται ταξινομητής D_i χρησιμοποιώντας $(XR_i^a Y)$ σαν σύνολο εκπαίδευσης

Φάση ταξινόμησης

Για ένα δοσμένο x, έστω $d_{i,j}(xR_i^a)$ είναι η πιθανότητα που αποδίδεται από τον ταξινομητή D_i στην υπόθεση ότι το x προέρχεται από την κλάση w_j. Υπολογίστε την εμπιστοσύνη για κάθε κλάση, w_j, από το μέσο όρο της συνδυαστικής μεθόδου:

$$\mu_j(x) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(xR_i^a), j = 1,...,c$$

Εκχώρηση του x στην κλάση με την μεγαλύτερη εμπιστοσύνη.

Εικόνα 3.7: Αλγόριθμος Rotation Forest

Ο αλγόριθμος Rotation Forest παρουσιάσθηκε πρόσφατα από την ερευνητική ομάδα του Rodriguez (Rodriguez et al., 2006).

4. ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΕΙΣ ΜΕΘΟΔΩΝ ΜΗΧ.ΜΑΘΗΣΗΣ ΣΕ ΑΝΟΜΟΙΟΓΕΝΗ ΔΕΔΟΜΕΝΑ

4.1 Εισαγωγή

Το πρόβλημα της μάθησης από ανομοιογενή δεδομένα αναφέρεται στην απόδοση αλγορίθμων μάθησης με παρουσία δεδομένων που υπο-παρίστανται και έντονα ασύμμετρες τάξεις κατανομών. Τα σύνολα ανομοιογενών δεδομένων έχουν πολύπλοκα χαρακτηριστικά και η μάθηση από τέτοια δεδομένα προϋποθέτει νέες αρχές, νέους αλγορίθμους και νέα εργαλεία για να μετασχηματίσουν με αποδοτικό τρόπο τεράστια ποσά πρωτογενών δεδομένων σε κατάλληλες πληροφορίες και παραστάσεις γνώσεων. Τέτοια σύνολα ανομοιογενών δεδομένων προκύπτουν σε διάφορα μεγάλης κλίμακας πολύπλοκα δικτυακά συστήματα, όπως διαδίκτυο, ασφάλεια, επιτήρηση, οικονομία, όπου απαιτείται η βασική κατανόηση της ανακάλυψης γνώσης και εφαρμογή τεχνικών χρήσεων δεδομένων σε διαδικασίες λήψης αποφάσεων.

Η ανακάλυψη γνώσης (knowledgediscovery) και η μηχανική δεδομένων (data engineering) παίζουν σημαντικό ρόλο σε ευρύ φάσμα επιστημονικών πεδίων και εφαρμογών του από την επεξεργασία πληροφοριών επιχειρήσεων έως τα διοικητικά συστήματα λήψης αποφάσεων, από την ανάλυση δεδομένων (σεμικροκλίμακα) έως την ανακάλυψη γνώση (σε μακροκλίμακα).

Το πρόβλημα ανομοιογενών δεδομένων σχετίζεται με την ικανότητα των ανομοιογενών δεδομένων να μπορούν να συμβιβάζουν σε μεγάλο βαθμό την απόδοση των περισσοτέρων αλγορίθμων μάθησης.Με δεδομένο ότι οι περισσότεροι αλγόριθμοι προϋποθέτουν ομοιογενείς τάξεις κατανομών ή ισοδύναμο κόστος εσφαλμένων ταξινομήσεων, όταν χρησιμοποιούν πολύπλοκα σύνολαανομοιογενών δεδομένων δεν αντιπροσωπεύουν τα κατανεμημένα χαρακτηριστικά των δεδομένων και δίνουν ανακριβή αποτελέσματα για τις τάξεις δεδομένων. Σε περιπτώσεις πραγματικών προβλημάτων το πρόβλημα ανομοιογενών δεδομένων μπορεί να δημιουργήσει σημαντικά προβλήματα με ποικίλες επιπτώσεις που χρειάζονται ειδική αντιμετώπιση και διερεύνηση(HeandGarcia,2009).

Κάθε σύνολο δεδομένων που διαθέτει μια άνιση κατανομή στις τάξεις του μπορεί να θεωρηθεί ως ανομοιογενές, αλλά τα ανομοιογενή δεδομένα αντιστοιχούν σε σύνολα δεδομένων που έχουν σημαντική ανομοιογένεια και σε ορισμένες περιπτώσεις έντονη ανομοιογένεια.

Τα μετρικά συστήματα χρησιμοποιούνται για να αξιολογήσουν την ακρίβεια μεθόδων μηχανικής μάθησης με ανομοιογενή δεδομένα. Τα σημαντικότερα από αυτά είναι (i) οι καμπύλες Διαχείρισης Λειτουργικών Χαρακτηριστικών (ReceiverOperatingCharacteristic (ROC)), (ii) οι καμπύλες ανάκλησης ακρίβειας (CostCurves), καθώς επίσης οι αξιολογήσεις μετρικών συστημάτων που βασίζονται σε άτακτα ή με πολλές κλάσεις δεδομένα με πίνακες κόστους, χρησιμοποιώντας εργαλεία όπως F-measure, G-meanκ.τ.λ. Στα επόμενα κεφάλαια παρουσιάζονται συνοπτικά οι κατηγορίες αυτές.

4.2 Καμπύλες Διαχείρισης Λειτουργικών Χαρακτηριστικών (ROCCurves)

4.2.1 Εισαγωγικές παρατηρήσεις

Οι **καμπύλες ROC** χρησιμοποιούνται για την αξιολόγηση τεχνικών μηχανικής μάθησης (machinelearning) και την εξόρυξη δεδομένων (datamining) με μία από τις πρώτες εφαρμογές τη σύγκριση και την αξιολόγηση διαφόρων αλγορίθμων ταξινόμησης (Spackman, 1989; Fawcett, 2006). Οι καμπύλες ROCχρησιμοποιήθηκαν για πρώτη φορά κατά τη διάρκεια του δεύτερου παγκοσμίου πολέμου για την ανάλυση σημαντικών ραντάρ στη συνέχεια χρησιμοποιήθηκαν στη θεωρία ανίχνευσης σημάτων. Στις αρχές της δεκαετίας 1950-1960 χρησιμοποιήθηκαν στην ψυχοφυσική (psychophysics) για την ανίχνευσηασθενών σημάτων στην περίπτωση ανθρώπων και ζώων. Οι καμπύλες ROC χρησιμοποιούνται εκτεταμένα στην Ιατρική για την αξιολόγηση διαγνωστικών δοκιμών, καθώς επίσης σε ιατρική έρευνα, βιομετρική, επιδημιολογία, ραδιολογία, και σε κοινωνικές επιστήμες, όπου χρησιμοποιείται η *ανάλυση ROC*, μία τεχνική για την αξιολόγηση της ακρίβειας ειδικών μοντέλων(defaultprobabilitymodel).

Μια ‘καμπύλη ROC’στη θεωρία ανίχνευσης σημάτων (Signaldetectiontheory) αποτελεί ένα γραφικό διάγραμμα που απεικονίζει την απόδοση ενός συστήματος δυαδικών ταξινομητών με μεταβλητό όριο διάκρισης. Η ανάλυση ROCπαρέχει εργαλεία για την επιλογή πιθανών βέλτιστων μοντέλων και την απόρριψη μη ικανοποιητικών μοντέλων ανεξάρτητα από το περιεχόμενο κόστους ή κλάσης κατανομής. Η ανάλυση ROC σχετίζεται με άμεσο τρόπο με την ανάλυση κόστους της διαγνωστικής λήψης αποφάσεων.

Στην μετεωρολογία, η συλλογή για την πρόβλεψη καιρικών συνθηκών και κατασκευή στατιστικών μοντέλων βοηθά σημαντικά τις ακριβέστερες προβλέψεις από τα δεδομένα αυτά. Η ανάλυση καμπυλών ROC χρησιμοποιείται για διαγνωστικές μελέτες στην κλινική Χημεία,

Φαρμακολογία και Φυσιολογία, και θεωρείται ως πρότυπο για περιγραφή και σύγκριση της ακρίβειας διάφορων διαγνωστικών δοκιμών (diagnostictests). Παραδείγματα εφαρμογών αποτελούν οι προβλέψεις καιρού, που δημοσιεύονται σε εφημερίδες, ραδιοφώνου, τηλεόραση, διαδίκτυο κ.α. Ακριβείς καιρικές προγνώσεις είναι βασικές για μεταφορές, ναυσιπλοϊα, γεωργία και κατασκευές κ.α.

4.2.2 Μέτρα εκτίμησης σε Καμπύλες ROC

Οι *καμπύλες Διαχείρισης Λειτουργικών Χαρακτηριστικών* [ReceiverOperatingCharacteristics (ROC)] ή καμπύλες ROC, είναι γραφικές μέθοδοι αξιολόγησης των χαρακτηριστικών διαγνωστικών δοκιμών. Αποτελούν γραφικές παραστάσεις για την ανταλλαγή (trade-off) μεταξύ ευαισθησίας και ιδιαιτερότητας, ειδικότερα μεταξύ των εσφαλμένων αρνητικών [falseNegative (FN)] και εσφαλμένων θετικών [FalsePositive (FP)] τιμών για κάθε πιθανή διακοπή. Μία καμπύλη ROC δείχνει τη σχέση δύο ειδικών κατανομών υπό την ίδια τάξη μονοτονικών μετασχηματισμών. Τέτοια γραφήματα χρησιμεύουν για την ικανότητα/δυνατότητα προφυλακτικών δοκιμών (screeningtests) για να διαπιστωθεί αν άτομα είναι υγιή ή ασθενή. Μπορεί επίσης να χρησιμοποιηθεί σε άλλες μελέτες, όπως η διάκριση αντίδρασης ερεθιστικών ως ασθενών ερεθισμάτων ή μη-ερεθισμάτων (nonstimuli).

Η καμπύλη ROC παρέχει μία οπτική αναπαράσταση των σχετικών διαφορών μεταξύ των ωφελημάτων (αληθή θετικά) και του κόστους (εσφαλμένα θετικά) της ταξινόμησης για τις κατανομές δεδομένων. Στην περίπτωση ταξινομήσεων σκληρού-τύπου (hard-typeclassifiers) με διακριτές ετικέτες τάξεων, κάθε ταξινομητής παράγει ένα ζεύγος (TP_rate, FP_rate) που αντιστοιχεί σε ένα απλό σημείο της καμπύλης ROC (Fawcett, 2006- 2002; Domignos, 1999; FreudandSchapire, 1996). Οι καμπύλες ROC χρησιμοποιούνται για την αξιολόγηση της ακρίβειας των προβλέψεων. Σημιώνεται ότι οι προβλέψεις αποτελούν βασικό μέρος κάθε επιχείρησης και έρευνας επιστημονικών πεδίων (Stehman, 1997).

Ηαξιολόγηση των *καμπύλων ROC* χρησιμοποιεί την αναλογία δύο μετρικών αξιολόγησης (με δύο στήλες): την αληθή θετική τιμή (TPrate) και την εσφαλμένη θετική τιμή (FTrate).

Το διάγραμμα της καμπύλης ROC σχηματίζεται σχεδιάζοντας τις τιμές TPκαι FP, και κάθε σημείο του διαστήματος ROC αντιστοιχεί στην απόδοση ενός απλού ταξινομητή για μία δεδομένη

κατανομή.Η καμπύλη ROC είναι χρήσιμη επειδή παρέχει μία ορατή αναπαράσταση των σχετικών διαφορών μεταξύ των πλεονεκτημάτων (από τις αληθείς θετικές) και μειονεκτημάτων/κόστη (εσφαλμένες θετικές) της ταξινόμησης σε σχέση με τις κατανομές δεδομένων.

Οι καμπύλες ROC χρησιμοποιούνται για τον καθορισμό μίας σύντομης τιμής (cutoffvalue) ιατρικών κλινικών δοκιμών. Για παράδειγμα, στην περίπτωση της δοκιμής PSA (ProstateSpecificAntigen) για καρκίνο προστάτη η ενδεικτική τιμή 4.0 ng/ml έχει καθοριστεί ως οριακή τιμή, με τιμή μικρότερη του 4.0 να θεωρείται κανονική, ενώ τιμή μεγαλύτερου του 4.0 να θεωρείται μη-κανονική (abnormal). Σημειώνεται ότι ασθενείς με τιμές PSA μικρότερες του 4.0 μπορεί να είναι μη-κανονικές (falsenegative) και τιμές μεγαλύτερες του 4.0 μπορεί να είναι κανονικές (falsepositive).

Συνήθως χρησιμοποιείται ένας πίνακας σύγχυσηςο (confusionmatrix), γνωστός ως και πίνακας συνάφειας ή πίνακας σφάλματος, που είναι μια συγκεκριμένη διάταξη πίνακα η οποία απεικονίζει την απόδοση ενός αλγορίθμου, συνήθως με επιβλεπόμενη μάθηση (στη μάθηση χωρίς επίβλεψη συνήθως ονομάζεται πίνακας ταιριάσματος (matchingmatrix)). Κάθε στήλητου πίνακα αντιπροσωπεύει τις περιπτώσεις σε μια προβλεπόμενη τάξη, ενώ κάθε σειρά αντιπροσωπεύει τις περιπτώσεις σε μια πραγματική κλάση. Το όνομα προέρχεται από το γεγονός ότι καθιστά εύκολο να διαπιστωθεί αν το σύστημα συγχέει δύο κλάσεις(δηλ. υπάρχει εσφαλμένη επισήμανση ως προς ένα άλλο).

Οι σημαντικότεροι βασικοί παράγοντες τεχνικών μετρικής είναι η ακρίβεια (accuracy) και η τιμή σφάλματος (errorrate). Αν θεωρήσουμε ένα βασικό πρόβλημα ταξινόμησης δύο τάξεων και {p,n} είναι ετικέτες της αληθούς θετικής ρκαι της αρνητικής ντάξης, τότε μία αναπαράσταση της απόδοσης ταξινόμησης μπορεί να δοθεί από ένα πίνακα σύγχυσης(confusionmatrix)(HeandGarcia, 2009). Συνήθως ένας πίνακας διαστάσεων (2x2)αναφέρεται στον αριθμό των ψευδώς θετικών (falsepositives), ψευδώς αρνητικών (falsenegatives),αληθώς θετικών (truepositives), αληθώς αρνητικών (truenegatives) αποτελεσμάτων. Αυτό επιτρέπει να έχουμε τις σχετικές αναλογίες για μια σωστή πρόβλεψη. Σημειώνεται ότι αυτό δεν είναι αρκετό διότι η ακρίβεια που βρίσκουμε δεν είναι αξιόπιστη, αν το σύνολο των δεδομένων μας είναι ασύμμετρο, δηλαδή ανομοιογενές.

Ένα μοντέλο ταξινόμησης (ταξινομητής ή διάγνωση) είναι μία απεικόνιση περιστατικών ορισμένων τάξεων και ομάδων. Ο ταξινομητής ή αποτέλεσμα διάγνωσης μπορεί να είναι μία πραγματική τιμή (συνεχές στοιχείο εξόδου), οπότε το σύνορο του ταξινομητή μεταξύ τάξεων

πρέπει να καθοριστεί από μια οριακή τιμή (thresholdvalue), π.χ. για τον καθορισμό εάν ένα άτομο έχει υπέρταση που βασίζονται σε μετρήσεις της πίεσης αίματος. Μπορεί να είναι μια διακριτή ετικέτα τάξης, που δείχνει μια από τις τάξεις.

4.2.3 Αποτίμηση Μοντέλου Καμπυλών ROC

Η καμπύλη ROC επ΄΄ισης δείχνει τις ενδείξεις TPR[TruePositiveRate] (στον άξονα των y) και FPR [FalsePositiveRate] (στον άξονα x). Η απόδοση κάθε ταξινομητή αναπαριστάται ως ένα σημείο στην καμπύλη ROC

Η αποτίμηση του μοντέλου ROC σχετίζεται με τις ακόλουθες εξισώσεις :

TPR = TP/(TP +FN) [θετικές τιμές] (4.1)

FPR = FP / (TN +FP) [αρνητικές τιμές] (4.2)

Η τεχνική αξιολόγησης με καμπύλες ROC χρησιμοποιεί την αναλογία δύο απλών στηλών που βασίζονται σε μετρικές αξιολογήσεις, δηλαδή την αληθή-θετική τιμή (TP_rate) και την εσφαλμένη θετική τιμή (FP_rate), οι οποίες ορίζονται με τον ακόλουθο τρόπο:

TP _ rate= TP / Pc και FP _ rate =FP / Nc . (4.3)

		ΠΡΟΒΛΕΨΗ		
		ΘΕΤΙΚΟ (POSITIVE)	ΑΡΝΗΤΙΚΟ (NEGATIVE)	
ΠΡΑΓΜΑΤΙΚΟ (ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΕΣΤ)	ΑΡΝΗΤΙΚΟ (NEGATIVE)	TRUE POSSITIVE (TP)	FALSE POSITIVE (FP)	ΠΡΟΒΛΕΨΗ (PRECISION)
	ΘΕΤΙΚΟ (POSITIVE)	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)	ΑΡΝΗΤΙΚΗ ΠΡΟΒΛΕΠΟΜΕΝΗ ΤΙΜΗ
		ΕΥΑΙΣΘΗΣΙΑ (SENSITIVITY)	ΕΙΔΙΚΟΤΗΤΑ (SPECIFICITY)	ΑΚΡΙΒΕΙΑ (ACCURACY)

Εικόνα 4.1: Πίνακας σύγχυσης για αναπαράσταση απόδοσης ταξινόμησης.

Η ακρίβεια και η τιμή σφάλματος δίνονται αντίστοιχα από τους τύπους:

$$\text{Accuracy} = (TP + TN) / (P_C + N_C) \quad (4.4)$$

$$\text{Error Rate} = 1 - \text{Accuracy}, \quad (4.5)$$

όπουTP (True Positive), TN (True Negative) και P_C , N_C (ColumnCounts).

Όσον αφορά την ακρίβεια (accuracy), άλλοι για την αποτίμηση μετρικών για την αξιολόγηση προβλημάτων ανομοιογενούς μάθησης είναι η ακρίβεια προσέγγισης (precision), η ανάκληση (recall), το μέτρο -F (F-measure) και ο μέσος όρος -G (G-mean). Οι παράγοντες αυτοί ορίζονται με τον ακόλουθο τρόπο:

$$\text{Precision} = TP / (TR+FP) \quad (4.6)$$

$$\text{Recall} = TP / (TP +FN) \quad (4.7)$$

$$\text{F- Measure} = [(1+ b)^2 * \text{Recall} * \text{Precision}] /[b^2 * \text{Recall} + \text{Precision}] \quad (4.8)$$

$$\text{G- Mean} = [TP /(TP+ FN) * TN /(TN+ FP)]^{1/2}, \quad (4.9)$$

ΌπουFN (FalseNegatives); FP (FalsePositives) και b ένας συντελεστής για τον καθορισμότης σχετικής σημασίας της ακρίβειας προσέγγισης έναντι της ανάκλησης (Guoand Viktor, 2004; Weiss, 2004, Provostetal. 1998; Sunetal. 2007).

Το γράφημα της καμπύλης ROC σχηματίζεται σχεδιάζοντας την TP_rate πάνω από την FP_rate, και κάθε σημείο της ROC αντιστοιχεί στην απόδοση ενός απλού ταξινομητή σε μια δεδομένη κατανομή .(Fawcett, 2003;2006).

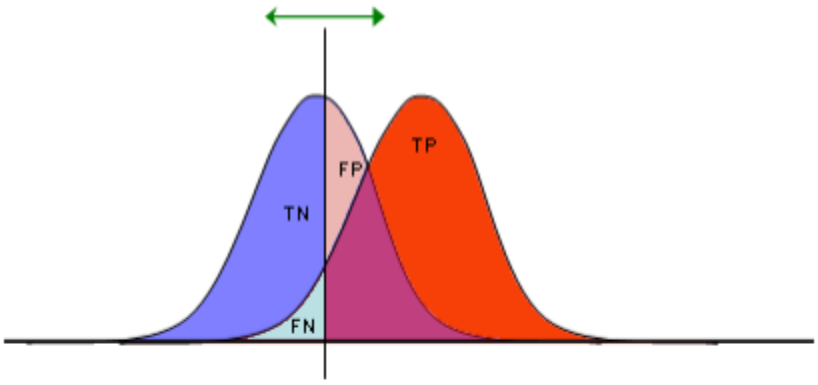
Αν όλα τα μη ταξινομημένα σύνολα είχαν τα ίδια βάρη, οι ετικέτες κλάσεων που εμφανίζονταν θα ήταν λιγότερο συχνές από τις τιμές στόχους. Κατασκευάζοντας ένα μοντέλο επιτυγχάνεται παράλληλα ένα πολύ μικρό συνολικό ποσοστό σφάλματος . Για την βελτίωση της ταξινόμησης των δέντρων αποφάσεων και για να αποκτήσουμε καλύτερα μοντέλα από ανομοιογενή δεδομένα, το ευρετήριο του δέντρου δημιουργεί αυτόματα ένα κατάλληλο πίνακα κόστους για την εξισορρόπιση της κατανομής των ετικετών των τάξεων όταν ένα δέντρο απόφασης εκπαιδεύεται. Ο πίνακας κόστους μπορεί να ρυθμιστεί και κατά βούληση.

	Προβλεψη (κλάση με πρόβλεψη)		
Πραγματική κλάση	$C(i/j)$	ΚΛΑΣΗ =+	ΚΛΑΣΗ=-
	ΚΛΑΣΗ =+	C(+,+)	C(+,-)
	ΚΛΑΣΗ=-	C(-,+)	C(-,-)

Εικόνα 4.2¨: Πίνακας κόστους, όπου Cij (/) είναι το κόστος λανθασμένης ταξινόμησης ενός δεδομένου της κλάσης i δεδομένου της κλάσεως j.

Οι καμπύλες ROC (Receiver Operating Characteristic) μπορούν να χρησιμοποιηθούν για συνολική απόδοση ταξινομητών που αναφέρονται σε ένα φάσμα ανταλλαγών μεταξύ αληθών θετικών τιμών σφαλμάτων και μη-αληθών θετικών τιμών σφαλμάτων (Swets, 1988).

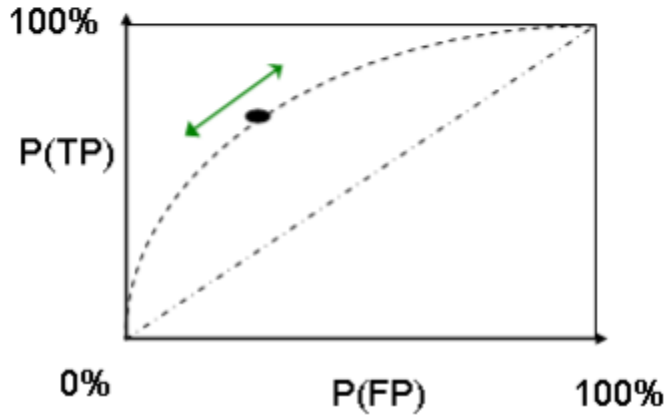
Η περιοχή κάτω από την καμπύλη (AreaUndertheCurve [AUC]) αποτελεί ένα αποδεκτό κριτήριο μέτρησης της απόδοσης για μια καμπύλη ROC (Bradley, 1997). Οι καμπύλες ROC θεωρούνται ότι αντιπροσωπεύουν την οικογένεια των καλύτερων οριακών αποφάσεων (bestdecisionboundaries) για το σχετικό κόστος των TP και FP.



Σχήμα (a)

TP	FP
FN	TN
1	1

Σχήμα (b)



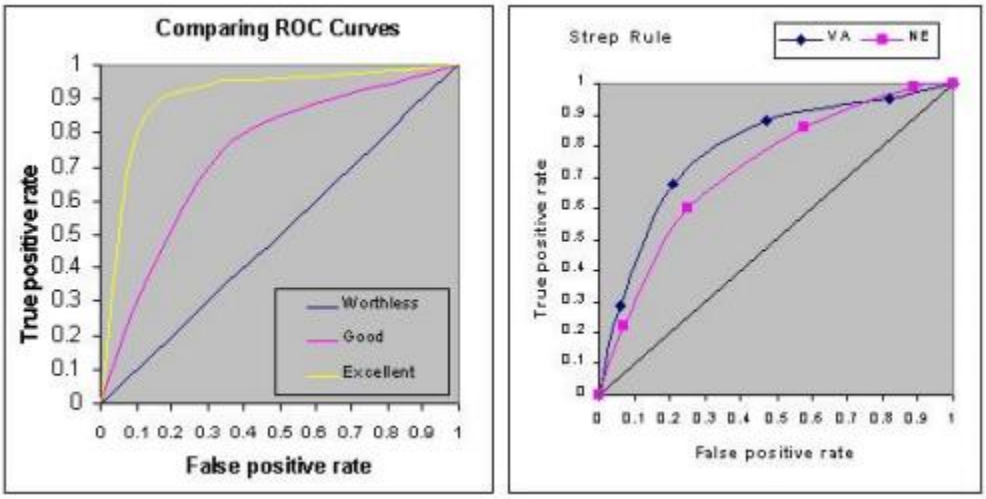
Σχήμα (c)

Εικόνα 4.3: Στο σχήμα (a) βλέπουμε την ταξινόμηση των μεταβλητών σε κανονικές κατανομές, στ σχήμα (b),ο πίνακας κόστους παρουσιάζει το κόστος λανθασμένης κατηγοριοποίησης ενός παραδείγματος (το βάρος) ως μέτρο εκτίμησης, και στο τελευταίο σχήμα (c) αναπαριστάται ως ένα σύνολο σημείων στην καμπύλη ROC.

Στην συνέχεια δίδεται παράδειγμα ερμηνείας σχήματος (α) σε σχέση με την καμπύλη ROC σχήμα (c). Υποθέστε ότι τα επίπεδα πρωτεϊνών στο αίμα σε νοσούντες και υγιείς ανθρώπους κυμαίνονται συνήθως σε 2 g / dL και 1 g / dL αντίστοιχα. Μια ιατρική εξέταση μπορεί να μετρήσει το επίπεδο της συγκεκριμένης πρωτεΐνης σε ένα δείγμα αίματος και να τα ταξινομήσει οποιοδήποτε αριθμό πάνω από ένα ορισμένο κατώφλι ως ένδειξη της νόσου. Ο πειραματιστής

μπορεί να προσαρμόσει το κατώφλι (μαύρη κάθετη γραμμή στο σχήμα), η οποία με την σειρά της θα αλλάξει το ποσοστό ψευδών θετικών αποτελεσμάτων. Αυξάνοντας το κατώφλι θα μπορούσε να οδηγήσει σε λιγότερες ψευδές θετικές (falsepositives) και αληθώς θετικές (truepositive) που αντιστοιχεί σε μια κίνηση προς τα αριστερά της καμπύλης. Το πραγματικό σχήμα της καμπύλης καθορίζεται από το πόσο επικαλύπτονται οι δύο κατανομές.

Σημειώνεται ότι ε μια καμπύλη ROC ο οριζόντιος άξων X αντιπροσωπεύει τιμές %FP=FP/(FP+TN), ενώ ο κάθετος άξων Y αντιπροσωπεύει τις τιμές %TP=TP/(TP+FN). Το ιδανικό σημείο της καμπύλης ROC είναι το σημείο (0,100) όπου όλα τα θετικά παραδείγματα ταξινομούνται σωστά και τα μη-αρνητικά παραδείγματα ταξινομούνται εσφαλμένα ως θετικά. Ένας τρόπος με το οποίο μπορεί μια καμπύλη ROC να σαρωθεί είναι να χρησιμοποιηθεί το ισοζύγιο τω δοκιμαστικών δειγμάτων για κάθε τάξη στο δοκιμαστικό σύνολο. Το κυρτό μέρος της καμπύλης ROC μπορεί να χρησιμοποιηθεί ως μια αποδοτική (εύρωστη) μέθοδος για την αναγνώριση δυναμικά βέλτιστων ταξινομητών (ProvostandFawcett, 2001).



4.4: Γραφική παράσταση και συμπεριφορά της καμπύλης ROC.

Για την ταξινόμηση της ακρίβειας μπορούμε να χρησιμοποιήσουμε μια κλίμακα [0,1], όπου από το 0.9-1 έχουμε εξαιρετικά αποτελέσματα, ενώ για τις τιμές κλίμακα0.5-0.6 προκύπτει ένα τεστ χωρίς αξία.

4.3 Καμπύλες Ανάκλησης Ακρίβειας (PRCurves)

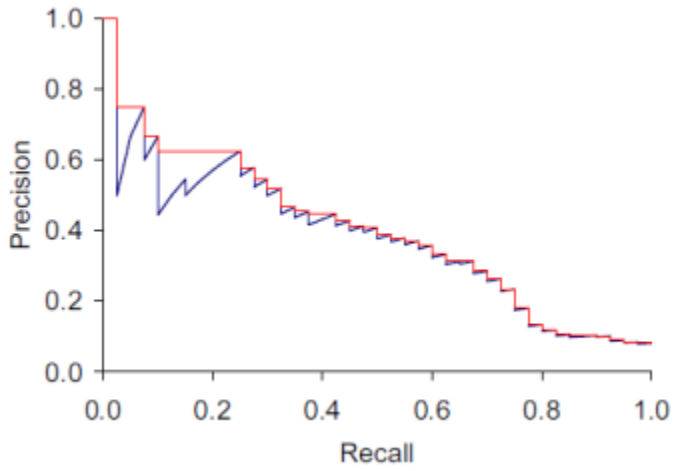
Οι καμπύλες PR , που αντιστοιχούν στις καμπύλες ROC , μπορούν να δώσουν μια πληρέστερη πληροφοριακή αναπαράσταση της αξιολόγησης απόδοσης από τις καμπύλες ROC . Μια

καμπύλη ROC ορίζεται σχεδιάζοντας την τιμή ακριβείας (precisionrate) πάνω από την τιμή ανάκλησης (recallrate). Μια καμπύλη κυριαρχεί στο χώρο ROC αν και μόνο αν κυριαρχεί στο χώρο PR (DavisandGoadrich, 2006; Bunescuetal., 2005).

Οι καμπύλες ROC μπορούν να δώσουν μια εποπτική αξιολόγηση της απόδοσης της μεθόδου, αλλά παρουσιάζουν μια αισιόδοξη άποψη της απόδοσης του αλγορίθμου. Σε τέτοιες περιπτώσεις οι καμπύλες ανάκλησης ακρίβειας (PrecisionRecall [PR] Curves) μπορούν να δώσουν πληρέστερες πληροφοριακές παραστάσεις για την αξιολόγηση απόδοσης των τεχνικών καμπυλών PR (DavisandGoadrich, 2006).

Η τιμή ακριβείας στον άξονα Y υπολογίζεται από τον αριθμό των σχετικών στοιχείων που ανακτώνται προς τα σχετικά στοιχεία. Συνηθίζεται να παρουσιάζεται μια γραφική παράσταση με διακριτά 'οδοντωτό' σχήμα.

Για παράδειγμα, εάν έχουμε το (κ+1)- οστό στοιχείο που ανακτάται και δεν είναι σχετικό, παρατηρούμε ότι η ανάκληση είναι η ίδια ,όμως η ακρίβεια μειώνεται .Εαν έχουμε το (κ+ 1) - οστό στοιχείο που ανακτάται και είναι σχετικό, τότε η ανάκληση και η ακρίβεια αυξάνονται και η καμπύλη δημιουργεί απότομες γωνίες προς τα δεξιά.



Εικόνα 4.5: Σχέση ακρίβειας (precision) και ανάκλησης (recall)

Για να αποφύγουμε αυτές τις απότομες γωνίες μπορεί να χρησιμοποιηθεί η τιμή ακριβειας με παρεμβολή (interpolatedprecision) pinter, όπου σε ένα ορισμένο επίπεδο ανάκλησης r ορίζεται ως υψηλότερη ακρίβεια που συναντάμε για κάθε επίπεδο ανάκλησης r’ ≥r με σχετικό τύπου

$$P_{int\,ep}(r)=max_{r' \geq r}\, p(r') . \qquad (4.10)$$

Στη συνέχεια παρουσιάζονται οι καμπύλες κόστους

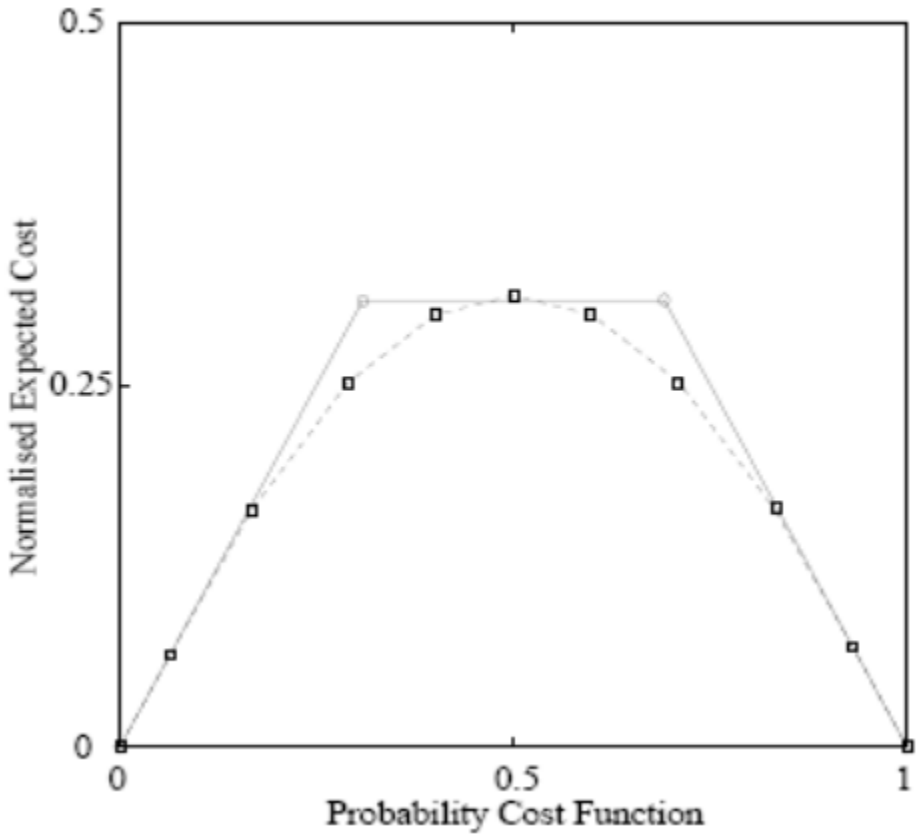
4.4 Καμπύλες Ακρίβειας / Κοστους

Οι καμπύλες κόστους παρέχουν μια κατανοητή αποτίμηση μετρικής για την απόδοση ταξινομητών σε μια μεταβαλλόμενη τάξη πιθανοτήτων ή κόστη εσφαλμένης ταξινόμησης (HolteandDrummond, 2005-2006).

Οι καμπύλες ROC αδυνατούν να δώσουν διαστήματα εμπιστοσύνης για την απόδοση ενός ταξινομητή και δεν μπορουν να συνάγουν την στατιστική σημασία της απόδοσης διαφόρων ταξινομητών (HolteandDrummond, 2005-2006). Η καμπύλη κόστους που έχει την ικανότητα να εκφράσει μια τεχνικη αποτίμησης με ευαισθησία κόστους που έχει τη ικανότητα να εκφράσει άμεσα την απόδοση ενός ταξινομητή για μεταβαλλόμενα κόστη εσφαλμένης τοποθέτησης και τάξης κατανομών σε ένα οπτικό περίγραμμα (visualformat). Η μέθοδος καμπύλης κόστους έχει τα χαρακτηριστικά αναπαράστασης της ανάλυσης ROC και προσφέρει διευρυμένες πληροφορίες για την απόδοση ταξινόμησης (HolteandDrummond, 2000- 2005-2006). Το αναμενόμενο κόστος ενός ταξινομητή μπορεί να παρασταθεί άμεσα από την καμπύλη , που είναι εύκολο να κατανοηθεί.

Η καμπύλη κόστους επιτρέπει την άμεση παρατήρηση του διαστήματος κόστους και καταδεικνύει αν ενας ταξινομητής είναι καλύτερος και ποσοτικά πόσο καλύτερος από άλλους ταξινομητές . Το αναμενόμενο κόστους ενός ταξινομητή για όλες τις πιθανές επιλογές εσφαλμένου

κόστους και κατανομών τάξεων εμφανιζεται στο ακόλουθο διάγραμμα .



Εικόνα 4.6: Αναμενόμενο Κόστος Ταξινομητών

Στο σχήμα 4.6 ο άξονας X είναι η πιθανότητα συνάρτησης κόστους για θετικά παραδείγματα και ορίζεται ως

$$PCF(+) = w_+/(w_+ + w_-) \quad (4.11)$$

Και ο άξονας Y είναι το αναμενόμενο κανονικοποιημένο κόστους ως προς το κόστος που προκύπτει όταν κάθε παράδειγμα είναι ταξινομημένο εσφαλμένα και ορίζεται ως

$$NE[C]=(1-TP)w_+ + FPww_+ + w- \quad (4.12)$$

Σημειώνεται ότι

$$w_{+}=p(+)C(-|+) \text{ και } w_{-}=p(-)C(+|-) \quad (4.13)$$

με P(α) : η πιθανότητα ενός δεδομένου παραδείγματος να είναι σε μια τάξη a και C(a|b) : το κόστος που προκύπτει αν ένα παράδειγμα σε μια τάξη b ταξινομηθεί εσφαλμένα ότι ανήκει στην τάξη α.

4.5 Μετρικές Αξιολόγησης για Πολυταξική Ανομοιογενή Μάθηση(Assessment Metrics for Multiclass Imbalanced Learning)

Τα πολυταξικά γραφήματα ROC μπορούν να χρησιμοποιηθούν σε πολυταξικά προβλήματα ανομοιογενούς μάθησης (multiclassimbalancedlearningproblems) (Fawcett, 2006). Για το πρόβλημα n – τάξεων , ο πίνακας σύγχυσης είναι ένας nxn πίνακας με n σωστές ταξινομήσης(τα στοιχεία της κύριας διαγώνιας) και (n2 –n) σφάλματα (τα στοιχεία εκτός διαγωνίου) (HandandTill, 2001; Abeetal., 2004; Sunetal., 2006; Fawcett, 2006). Για τον υπολογισμό πολιταξικών τιμών περιοχών –κάτω –από την καμπύλη (AreasUnderCurve [AUC]) μπορεί να χρησιμοποιηθεί η ακόλουθη πιθανοθεωρητική υπολογιστική προσέγγιση

(a) η καμπύλη ROC για κάθε τάξγ αναφορών δημιουργείται και υπολογίζονται οι αντίστοιχες AUC

(b) Όλες οι AUC συνδιάζονται με έναν συντελεστή βάρους σύμφωνα με την υπερίσχυση της τάξης αναφορών στα δεδομένα

Μια γενικευμένη προσέγγιση χρησιμοποιεί το μέτρο-M (M-measure, που συναθροίζει όλα τα ζεύγη τάξεων τα οποία βασίζονται εγγενή χαρακτηριστικά των AUC . Η μέθοδος αυτή δεν επηρεάζεται από την κατανομή τάξεων και τα σφάλματα κόστους (Chawlaetal., 2002).

Αναφέρεται ότι η μάθηση με ευαισθησία κόστους μπορεί να χρησιμοποιήσει κόστη εσφαλμένης ταξινόμησης για την αξιολόγηση απόδοσης πολυταξικών ανομοιογενών προβλημάτων (Abeetal., 2004; LiuandZhou, 2006).Ειπρόσθετα με την ανάλυση για την ανάκληση τιμών κάθε τάξης για την πολυταξική ανομοιογενή μάθηση (Sunetal. 2006, Chanetal. 1999).

Θεμελιώδη και κριτικής σημασίας ερβτήματα ηια την ανομοιογενή μάθηση, που θα έχει σημαντικές επιδράσεις στην πρόοδο της μηχανικής μάθησης και γενικότερα της μηχανικής δεδομένων (dataengineering),, είναι τα ακόλουθα βασικά ερωτήματα:

(i) Σε ποιό βαθμό οι μέθοδοι ανομοιογενούς μάθησης βοηθούν τις ικανότητες μάθησης ;

(ii) Πώς οι αλγόρθμοι μηχανικής μάθησης μπορούν να επιδράσουν καλύτερα με οποιαδήποτε δεδομένα δοθούν; (Provost, 2000; Weiss and Provost, 2003; Estabrooks et al., 2004; He and Garcia, 2009).

Στα πεδία ανακάλυψης και μηχανικής δεδομένων υπάρχουν αρκετές ειδικές δοκιμασίες επιδόσεων (benchmarks) για τν αποτίμηση της επίδρασης διαφόρων αλγορίθμων και εργαλείων δεδομένων μάθησης μηχανικής, όπως η αποθήκη UCI (UCIrvineMachineLearningRepository, 2009), οι επιστημονικές και τεχνικές βάσεις δεδομένων NIST (NISTScientificandTechnicalDatabases, 2009), αλλά υπάρχει πολύ περιορισμένος αριθμός δοκιμασιών επιδόσεων αποκλειστικά αφιερωμένων σ προβλήματα ανομοιογενούς μάθησης. Σημειώνεται επίσης ότι πολλά σύνολα δεδομένων απαιτούν επιπρόσθετους χειρισμούς πριν εφαρμοσθούν σε τεχνικές ανομοιογενούς μάθησης (HeandGarcia, 2009).

Σε πολλά περιβάλλοντα ρεαλιστικών εφαρμογών, όπως η εξόρυξη δεδομένων από δίκτυα (web), δίκτυα αισθητήρων (sensornetworks) , συστήματα πολυμέσων κ.α., τα πρωτογενή δεδομένα είναι διαθέσιμα συνεχώς σε κάποια χρόνο-διαστήματα μάθησης. Για τέτοια δεδομένα μάθησης χρειάζονται νέες αρχές, μεθοδολογίες , αλγόριθμοι και εργαλεία για να μετασχηματίσουν πρωτογενή δεδομένα σε χρήσιμες πληροφορίες και αναπαράσταση γνώσεων για να υποστηρίξουν διαδικασίες λήψης αποφάσεων (HeandChen, 2008).

Οι αλγόριθμοι από ανομοιογενή δεδομένα , με τη διαθεσιμότητα τεράστιων ποσών πρωτογενών δεδομένων σε πολλές σύγχρονες εφαρμογές πραγματικού κόσμου, παίζουν κρτικό ρόλο σε πολλές διάφορες περιοχές της επιστήμης και τεχνολογίας.

4.6 Μέθοδοι Ευαισθησίας Κόστους για Ανομοιογενή Μάθηση

Οι δειγματοληπτικές μέθοδοι συνήθως ομογενοποιούν κατανομές θεωρώντας αντιπροσωπευτικές αναλογίες παραδειγμάτων τάξεων στην κατανομή, ενώ οι μέθοδοι μάθησης με ευαισθησία κόστους θεωρούν τα κόστη που συνδέονται με παραδείγματα εσφαλμένης ταξινόμησης (Elkan, 2001). Η μάθηση με ευαισθησία κόστους αντιμετωπίζει το πρόβλημα ανομοιογενούς μάθησης χρησιμοποιώντας διαφορετικούς πίνακες κόστους που περιγράφουν τα κόστη για εσφαλμένη ταξινόμηση κάθε ιδιαίτερου παραδείγματος δεδομένων. Αναφέρεται ότι οι θεμελιώδεις αρχές και οι αλγόριθμοι μεθόδων με ευαισθησία κόστους μπορούν να εφαρμοστούν σε προβλήματα ανομοιογενούς μάθησης (Chawla, 2004) και ότι σε ορισμένες περιπτώσεις οι μέθοδοι

μάθησης με κόστος ευαισθησίας αποδεικνύονται υπέρτερες από τις μεθόδους δειγματοληψίας (LiuandZhou, 2006).

Η έννοια του πίνακα είναι θεμελιώδης για τις μεθοδολογίες μάθησης με ευαισθησία κόστους . Ο πίνακας κόστους μπορεί να θεωρηθεί ως μια αριθμητική αναπαράσταση του μειονεκτήματος (penalty) παραδειγμάτων ταξινόμησης από τη μία τάξη στην άλλη.

Η υλοποίηση μεθόδων μάθησης με ευαισθησία κόστους περιλαμβάνει τις ακόλουθες κατηγορίες (i) η πρώτη κατηγορία εφαρμόζει κόστη εσφαλμένης ταξινόμησης σε σύνολα δεδομένων ως ένας τύπος βάρους χώρο-δεδομένων (dataspace), (ii) η δεύτερη κατηγορία εφαρμόζει τεχνικές ελαχιστοποίησης κόστους (μετατεχνικές) σε συνδυαστικά σχήματα συνολικών μεθόδων (Domingos, 1999), και (iii) η τρίτη κατηγορία περιλαμβάνει συναρτήσεις με κόστος ευαισθησίας ή χαρακτηριστικά άμεσα σε παραδείγματα ταξινόμησης για να προσαρμόσουν το πλαίσιο ευαισθησίας κόστους σε αυτούς τους ταξινομητές.

Βιβλιογραφικές Πηγές

AIZERMAN M., BRAVERMAN E. and ROZONOER L. (1964): Theoretical foundation of the potential function method in pattern recognition learning, Automation and Remote Control 25, 821-837

ANDERSON J.R. (1980): Cognitive Psychology and its implications, Freeman W.H., NY

ANDERSON J.R. (1983): The architecture of cognition, Harvard University Press, Cambridge, Mass.

AHO A.V., HOPCROFT J.E. and ULLMAN J.D. (1974): The design and analysis of computer algorithms, Addison-Wesley Pub. Co., section 1.3

BERLINSKI D.(2000): The Advent of the Algorithm: The 300-Year Journey from an Idea to the Computer, Harcourt, Inc., San Diego,

BISHOP C.M. (1995): Neural Networks for Pattern Recognition, Oxford University Press, Oxford, UK

BROADBENT D.E. (1958): Perception and Communication, Pergamon, Oxford, UK

BROOKS R.A. (1989): Engineering approach to building complete,intelligewnt beings, Procs of the SPIE-the international society for optical engineering 1002, 618-625

CHAWLA N. V., BOWYER K. W., HALL L. O. and KEGELMEYER P. (2002): SMOTE: Synthetic Minority Oversampling Technique,, Journal of Artificial Intelligence Research 16, 321-357.

COHEN P.R. (1995): Empirical methods for artificial intelligence, MIT Press, Cambridge, Mass.

COHEN W.W. and PAGE C.D. (1995): Learnability in inductive logic programming: Methods and Results, New Generation Computing 13 (3-4), 369-409.

DIETTERICH T. (1990): Machine Learning, Annual Review of Computer Science 4, 255-306

DOMINGOS P. (1999): Metacost: A General Method for Making Classifiers Cost-sensitive, In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 155-164, San Diego, CA, ACM Press.

FAWCETT T. (2004): ROC Graphs: Notes and Practical Considerations for Researchers, Pattern Recognition Letters **27**(8), 882-891.

FAWCETT T. (2006): An Introduction to ROC Analysis," Pattern Recognition Letters 27 (8), 861-874.

GOLDREICH O.: Computational Complexity: A Conceptual Perspective, Cambridge University Press, 2010.

GONEN M.: Receiver Operating Characteristic (ROC) Curves, Memorial Sloan-Kettering Cancer Center

GUREVICH Y. (2000): Sequential Abstract State Machines Capture Sequential Algorithms, ACM Transactions on Computational Logic 1 (1), 77–111.

HAIBO H. and GARCIA E.A. (2009): Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering 21 (9)

HAND D.J. (2009): Measuring classifier performance: A coherent alternative to the area under the ROC curve, Machine Learning 77, 103–123

HARLEN W. (2007a): Assessment of Learning, London: Sage.

HARLEN W. (2007b): The Quality of Learning: assessment alternatives for primary Education, (Primary Review Research Survey 3/4), Cambridge: University of Cambridge.

HROMKOVIC J. (2004): Theoretical computer science: introduction to Automata, computability, complexity, algorithmics, randomization, communication, and cryptography, Springer, pp. 177–178.

JAPKOWICZ N. (2000): The Class Imbalance Problem: Significance and Strategies, In Proceedings of the 2000 International Conference on Artificial Intelligence (IC-A1'2000): Special Track on Inductive Learning, Las Vegas, Nevada.

JAPKOWICZ N. and STEPHEN S. (2002): The Class Imbalance Problem: A Systematic Study, Intelligent Data Analysis 6 (5), 429-449.

LIU X.Y., WU J. and ZHOU Z.H. (2006): Exploratory Under Sampling for Class Imbalance Learning, Proc. Int'l Conf. Data Mining, 965-969.

ΑΔΑΜΟΣ Α. (2006) Αλγόριθμοι ταξινόμησης υπερφασματικής απεικόνισης για την ανίχνευση, τμηματοποίηση και ταυτοποίηση χαρακτηριστικών διαγνωστικής σημασίας, Διπλωματική Εργασία, Πολυτεχνείο Κρήτης, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών.

ΒΛΑΧΑΒΑΣ Ι, ΚΕΦΑΛΑΣ Π., ΒΑΣΙΛΕΙΑΔΗΣ Ν.,ΚΟΚΚΟΡΑΣ Φ. και ΣΑΚΕΛΛΑΡΙΟΥ Η.: Μηχανική Μάθηση (Machine Learning), Τεχνητή Νοημοσύνη –Β Έκδοση.

ΣΑΡΔΗΣ Γ. (2009):Ιεραρχική Ταξινόμηση Δεδομένων Πολλαπλών Ετικετών, Πτυχιακή Εργασία, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Σχολή Θετικών Επιστημών, Τμήμα Πληροφορικής.

ΤΡΟΓΚΑΝΗΣ Ν.Α (2006): Μέθοδοι εκμάθησης ταξινομητών από θετικά παραδείγματα με αριθμητικά χαρακτηριστικά, Διπλωματική Εργασία, ΕΜΠ , Τομέας Τεχνολογίας Πληροφορικής Υπολογιστών, Αθήνα.

http://nemertes.lis.upatras.gr/jspui/bitstream/10889/8630/1/THESIS-LIPITAKI-FINAL.pdf

http://archive.ics.uci.edu/ml/index.php

http://aibook.csd.auth.gr/include/ch18.pdf

http://aibook.csd.auth.gr/include/slides/Chap18.pdf

https://repository.kallipos.gr/bitstream/11419/3382/1/02_chapter_04.pdf

https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/

https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

https://vvvvw.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-001.pdf

https://link.springer.com/chapter/10.1007%2F11538059_91?LI=true

https://www.researchgate.net/profile/Taeho_Jo/publication/220542017_A_Multiple_Resampling_Method_for_Learning_from_Imbalanced_Data_Sets/links/53fe8cf40cf283c3583bdd19.pdf

https://ocs.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-003.pdf

http://www.csi.uottawa.ca/~hguo028/papers/KDDExplorations2004.pdf

http://sci2s.ugr.es/keel/dataset/includes/catImbFiles/2004-Batista-SIGKDD.pdf

http://sci2s.ugr.es/keel/pdf/specific/congreso/akbani_svm_2004.pdf

http://machinelearning.wustl.edu/mlpapers/paper_files/icml2007_HulseKN07.pdf

http://pages.stern.nyu.edu/~fprovost/Papers/skew.PDF

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.308.9315&rep=rep1&type=pdf

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.309.942&rep=rep1&type=pdf

https://pdfs.semanticscholar.org/4a61/1badb9b31b82cdeabfa77d6bd4ce17b194b0.pdf

https://s3.amazonaws.com/academia.edu.documents/3252371/kdd03-talk.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1505689024&Signature=NkE26W6aXGoMSfVv5yxx%2BJk0Dok%3D&response-content-disposition=inline%3B%20filename%3DMining_Concept-Drifting_Data_Streams_Usi.pdf

http://www.ulb.ac.be/di/map/adalpozz/pdf/poster_unbalanced.pdf

http://worldcomp-proceedings.com/proc/p2013/DMI8016.pdf

https://www.researchgate.net/figure/220637823_fig6_Fig-15-AdaCost-algorithm

https://uwspace.uwaterloo.ca/bitstream/handle/10012/3000/thesis.pdf?sequence=1&isAllowed=y

https://pdfs.semanticscholar.org/9ddf/bc2cc5c1b13b80a1a487b9caa57e80edd863.pdf

https://pdfs.semanticscholar.org/36bc/a41eba5a7cea8d69a89ee7bc24923bc380ba.pdf

https://pdfs.semanticscholar.org/3c0c/776156d537fe438af2fb25623fdc8816cda1.pdf

https://pdfs.semanticscholar.org/36bc/a41eba5a7cea8d69a89ee7bc24923bc380ba.pdf

https://pdfs.semanticscholar.org/fb76/4c7bdaa19550e520f7f0eeb8003c11b1b0fb.pdf

https://s3.amazonaws.com/academia.edu.documents/41732527/A_Multiple_Model_Cost-Sensitive_Approach20160129-11311-f1um8b.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1505689306&Signature=hAuYKieCt6wMxq1W8Y4f0PMnQwY%3D&response-content-disposition=inline%3B%20filename%3DA_Multiple_Model_Cost-Sensitive_Approach.pdf

ftp://nozdr.ru/biblio/kolxo3/Cs/CsLn/M/Machine%20Learning..%20ECML%202000,%2011%20conf.(LNCS1810,%20Springer,%202000)(ISBN%203540676023)(468s)_CsLn_.pdf#page=421

https://pdfs.semanticscholar.org/e957/9d7a0a446137fab0c2031c255bafc64a02ef.pdf

http://www.gipsa-lab.fr/~jocelyn.chanussot/publis/ieee_grsl_14_xia_rotation.pdf

https://www.researchgate.net/profile/Kun_Hong_Liu/publication/5457299_Cancer_classification_using_Rotation_Forest/links/5783014d08ae69ab88286b53.pdf