Πολυτεχνείο Κρήτης
Τμήμα Ηλεκτρονικών Μηχανικών και
Μηχανικών Ηλεκτρονικών Υπολογιστών
Τομέας Τηλεπικοινωνιών

Συνδυασμός Ταξινομητών χρησιμοποιώντας Μήτρες Αποφάσεων (Decision Templates) με εφαρμογή στην Ταξινόμηση Καρκινικών Δεδομένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ευαγγελία Γ. Δουγαλή

Επιβλέπων Καθηγητής: *Καθ. Μιχάλης Ζερβάκης*

Εξεταστική Επιτροπή: *Καθ. Μιχάλης Ζερβάκης*
Επ. Καθ. Αθανάσιος Λιάβας
Αν. Καθ. Κωνσταντίνος Μπάλας

Χανιά 2008

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΧΑΡΙΣΤΙΕΣ

Στο σημείο αυτό, θα ήθελα να ευχαριστήσω όλους αυτούς που με βοήθησαν στην εκπόνηση της διπλωματικής μου εργασίας.

Αρχικά, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον καθηγητή μου κ. Μιχάλη Ζερβάκη, για την πολύτιμη βοήθεια και καθοδήγησή του σε όλη τη διάρκεια αυτής της εργασίας.

Θα ήθελα, επίσης, να ευχαριστήσω τον μεταπτυχιακό φοιτητή Γιώργο Μανίκη, για την εποικοδομητική συνεργασία μας καθώς και τους καθηγητές κ. Λιάβα Αθανάσιο και κ. Μπάλα Κωνσταντίνο για την συμμετοχή τους στην παρουσίαση και αξιολόγηση αυτής της εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου για την αμέριστη συμπαράστασή τους όλα αυτά τα χρόνια των προπτυχιακών μου σπουδών.

Δουγαλή Ευαγγελία

Χανιά, Μάρτιος 2008

ΠΕΡΙΛΗΨΗ

Στις μέρες μας, ο καρκίνος αποτελεί μία από τις σοβαρότερες ασθένειες του ανθρώπου και ταυτόχρονα μία από τις κυριότερες αιτίες θανάτου. Μπορεί να αναπτυχθεί σε όλα σχεδόν τα όργανα και τους ιστούς του ανθρώπινου σώματος και βασική προϋπόθεση για την θεραπεία του, αποτελεί η έγκαιρη διάγνωσή του. Για τον σκοπό αυτό, τα τελευταία χρόνια παρατηρείται μία συνεχής προσπάθεια δημιουργίας μαθηματικών μοντέλων (ταξινομητών) που θα μπορούν να δρουν βοηθητικά στην σωστή διάγνωση του καρκίνου. Οι ταξινομητές λειτουργούν είτε ως μεμονωμένα στοιχεία είτε συνδυάζονται για να πετύχουν πιθανώς καλύτερη απόδοση. Αντικείμενο αυτής της εργασίας αποτελεί η μελέτη και η υλοποίηση του συνδυασμού αυτών των ταξινομητών με διάφορες μεθόδους, όπως με τα decision templates, καθώς και η εφαρμογή τους σε τέσσερα datasets. Από τα τέσσερα datasets, δύο αφορούν την οξεία μυελοειδή λευχαιμία και τα άλλα δύο αφορούν τον καρκίνο του στήθους.

ΕΙΣΑΓΩΓΗ

Με τη συνεχή μείωση της θνησιμότητας από την καρδιακή νόσο, ο καρκίνος θα μπορούσε να αναδειχθεί σήμερα σε κύρια αιτία θανάτου στον ανεπτυγμένο κόσμο. Ο καρκίνος εμφανίζεται με ποικίλες μορφές και προσβάλλει ποικίλα όργανα. Μερικές από τις πιο συνηθισμένες μορφές του είναι ο καρκίνος του εγκεφάλου, του πνεύμονα, του δέρματος, των οστών και άλλων ζωτικής για τον άνθρωπο σημασίας οργάνων. Δυστυχώς, ακόμη και στις μέρες μας, παρόλο που η ιατρική έχει παρουσιάσει αλματώδη ανάπτυξη στην σωστή διάγνωση και ίαση πολλών ασθενειών, που μέχρι πριν από μερικά χρόνια θεωρούνταν ανίατες, τα αποτελέσματα των ερευνών δεν είναι το ίδιο θεαματικά και στην περίπτωση του καρκίνου. Στην περίπτωση αυτή, σημαντικός παράγοντας για την πλήρη ίαση αποτελεί η έγκαιρη διάγνωσή του. Για τον λόγο αυτό, εκτός από τις έρευνες που γίνονται στο πεδίο της ιατρικής, γίνονται και προσπάθειες υλοποίησης μοντέλων σε υπολογιστές, τα οποία θα δρουν συμπληρωματικά προς τις διαγνώσεις των γιατρών και θα έχουν σαν σκοπό την μείωση των λανθασμένων διαγνώσεων.

Η εργασία αυτή, έχει σαν σκοπό την μελέτη, την υλοποίηση και τη σύγκριση τέτοιων μοντέλων καθώς και την εξαγωγή κάποιων συμπερασμάτων από την εφαρμογή τους σε τέσσερα διαφορετικά σύνολα δεδομένων που αφορούν τον καρκίνο. Τα μοντέλα που υλοποιούνται αφορούν τον συνδυασμό ταξινομητών με διάφορες μεθόδους, όπως για παράδειγμα Majority Voting, Decision Templates κ.ά..

Στο πρώτο κεφάλαιο αυτής της εργασίας, θα διατυπωθούν κάποιες βασικές έννοιες σχετικά με την ταξινόμηση. Θα αναλυθεί ο βασικός τρόπος λειτουργίας ενός μοντέλου ταξινόμησης καθώς και κάποιες μέθοδοι οι οποίες συμβάλλουν ακόμη και στο αρχικό αυτό στάδιο στην βελτίωση της απόδοσης των ταξινομητών. Επιπλέον, θα παρουσιαστούν τα μέτρα με τα οποία υπολογίζεται η απόδοση των ταξινομητών.

Στο δεύτερο κεφάλαιο, θα περιγραφεί συνοπτικά η λειτουργία κάποιων ταξινομητών ως μεμονωμένων στοιχείων.

Στο τρίτο κεφάλαιο, θα διατυπωθεί η έννοια της διαφορετικότητας (diversity) μεταξύ ταξινομητών και σε τι αυτή μας χρησιμεύει. Στη συνέχεια, θα μελετηθούν οι μέθοδοι με τις οποίες μπορούμε να υπολογίζουμε την διαφορετικότητα. Ακόμη, θα μελετηθούν οι διάφορες μέθοδοι συνδυασμού των ταξινομητών και το αν και πώς αυτές συμβάλλουν στην βελτιστοποίηση της απόδοσης της ταξινόμησης.

Στο τέταρτο κεφάλαιο, θα παρουσιαστεί αναλυτικά η υλοποίηση καθώς και τα αποτελέσματα που προκύπτουν από τις μεθόδους που έχουν περιγραφεί στα προηγούμενα κεφάλαια.

Στο πέμπτο κεφάλαιο, θα διατυπωθούν τα συμπεράσματα με βάση τα αποτελέσματα του τέταρτου κεφαλαίου και επιπλέον τι θα μπορούσε να γίνει περαιτέρω σε αυτήν την εργασία.

Τέλος, στο Παράρτημα θα δοθεί μία σύντομη περιγραφή των χαρακτηριστικών του κάθε dataset.

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΙΚΕΣ ΕΝΝΟΙΕΣ

1.1 Εισαγωγή στην αναγνώριση προτύπων

Η αναγνώριση προτύπων (pattern recognition) έχει σαν αντικείμενο την ταξινόμηση (classification) ενός συνόλου αντικειμένων (πρότυπα) σε κάποιες κατηγορίες από ένα σύστημα. Η διαδικασία αυτή, είναι εύκολη για τον άνθρωπο, μιας και έχει την ικανότητα να διακρίνει τα αντικείμενα, να παρατηρεί τις όποιες ομοιότητες ή διαφορές στα χαρακτηριστικά τους και επομένως μπορεί να τα ομαδοποιεί. Δεν συμβαίνει το ίδιο όμως και σε ένα σύστημα, το οποίο, σε αντιδιαστολή με τον άνθρωπο, παρόλο που έχει την δυνατότητα ανάλυσης περισσότερων δεδομένων στο ίδιο χρονικό διάστημα, πρέπει να εκπαιδευτεί για αυτήν την διαδικασία. Επομένως, για να επιτευχθεί η εκπαίδευση ενός συστήματος σε προβλήματα ταξινόμησης, τα αντικείμενα περιγράφονται από ένα σύνολο χαρακτηριστικών, που προέρχονται συνήθως από μετρήσεις και οργανώνονται σε κάποιον πίνακα. Η αναγνώριση προτύπων αντιμετωπίζει την πρόκληση της επίλυσης προβλημάτων της πραγματικής ζωής και για τον λόγο αυτόν, παρόλο που υπάρχουν δεκάδες αποδοτικές μελέτες, οι τελευταίες συνυπάρχουν σε κάποιο βαθμό και με την διαίσθηση.

1.2 Μάθηση

Ένα σύστημα, όπως αναφέρθηκε και προηγουμένως, θα πρέπει να εκπαιδευτεί για να μπορεί να διαχωρίζει και να ταξινομεί τα πρότυπα. Η εκπαίδευση έγκειται στο να μάθει το σύστημα αρχικά, να διακρίνει κάποια πρότυπα και στη συνέχεια να μπορεί να εφαρμόσει την γνώση αυτή και σε άλλα νέα πρότυπα. Ανάλογα με τον είδος της εκπαίδευσης, τα προβλήματα αναγνώρισης προτύπων διακρίνονται σε δύο κύριες

κατηγορίες: στη μάθηση με επίβλεψη (supervised learning) και στη μάθηση χωρίς επίβλεψη (unsupervised learning).

1.2.1 Μάθηση με επίβλεψη (Supervised Learning)

Στη μάθηση με επίβλεψη, κάθε πρότυπο στο σύνολο των δεδομένων έχει ήδη αντιστοιχηθεί σε μία κατηγορία (κλάση). Με άλλα λόγια, το σύστημα ταξινόμησης γνωρίζει για τα δεδομένα που δέχεται στο στάδιο της εκπαίδευσης την έξοδό τους (σε ποια κλάση ανήκουν). Σε αυτήν την περίπτωση, ο σκοπός της αναγνώρισης προτύπων είναι η εκπαίδευση του ταξινομητή στην λογική και στατιστική αντιστοίχηση νέων αντικειμένων σε κλάσεις. Η γνώση της ταξινόμησης που έχει αποκτήσει το σύστημα σε αυτήν την διαδικασία μπορεί να είναι αμυδρή, αλλά η ακρίβεια της αναγνώρισης του ταξινομητή, θα είναι το στοιχείο που θα κρίνει την ορθότητά του.

1.2.2 Μάθηση χωρίς επίβλεψη (Unsupervised learning)

Στη μάθηση χωρίς επίβλεψη, το πρόβλημα έγκειται στην ανακάλυψη της δομής του συνόλου δεδομένων, εφόσον υπάρχει κάποια. Αυτό σημαίνει ότι ο χρήστης θέλει να γνωρίζει κατά πόσο υπάρχουν ομάδες στα δεδομένα και ποια είναι τα χαρακτηριστικά που κάνουν τα δεδομένα στην ομάδα να είναι όμοια μεταξύ τους και διαφορετικά μεταξύ άλλων ομάδων. Υπάρχουν πολλοί αλγόριθμοι ομαδοποίησης για την μάθηση χωρίς επίβλεψη. Η επιλογή του αλγορίθμου είναι καθαρά θέμα του σχεδιαστή και γι’ αυτόν τον λόγο διαφορετικοί αλγόριθμοι μπορεί να δημιουργήσουν διαφορετικές ομάδες για τα ίδια δεδομένα. Ακόμη, δεν υπάρχει κάποια σωστή λύση με την οποία μπορούν να συγκριθούν οι διάφοροι αλγόριθμοι οπότε και η μόνη ένδειξη του πόσο καλό είναι το αποτέλεσμα, εξαρτάται από την υποκειμενική εκτίμηση του χρήστη.

1.3 Χαρακτηριστικά (features)

Όπως αναφέρθηκε και προηγουμένως, τα αντικείμενα (πρότυπα ή δεδομένα ή δείγματα) περιγράφονται από κάποια χαρακτηριστικά, τα οποία είναι συνήθως αποτέλεσμα μετρήσεων και είναι χρήσιμα για την αναπαράσταση των δεδομένων. Τα χαρακτηριστικά μπορεί να είναι ποσοτικά ή ποιοτικά. Τα ποσοτικά, μπορούν να παίρνουν είτε διακριτές τιμές (π.χ. αριθμός κατοίκων μιας πόλης) είτε συνεχείς (π.χ. μήκος). Τα ποιοτικά χαρακτηριστικά, είναι εκείνα με μικρό αριθμό πιθανών τιμών και τα οποία μπορεί να έχουν διαβαθμίσεις (π.χ. σειρά κατάταξης) ή να είναι ονομαστικά (π.χ. όνομα ασθένειας).

Η στατιστική αναγνώριση προτύπων λειτουργεί με αριθμητικά χαρακτηριστικά (π.χ. πίεση του αίματος, ηλικία κλπ). Για ένα αντικείμενο x, που περιγράφεται από n χαρακτηριστικά, η απεικόνισή του γίνεται με ένα n-διάστατο διάνυσμα $\mathbf{x} = [x_1, .., x_n]^T \in \Re^n$. Τα στοιχεία x_i προέρχονται από τις μετρήσεις των χαρακτηριστικών του αντικειμένου. Το πραγματικό διάστημα $\Re^n$ ονομάζεται χώρος χαρακτηριστικών (feature space), όπου ο κάθε άξονάς του αντιστοιχεί σε ένα φυσικό χαρακτηριστικό.

1.4 Κλάσεις και ετικέτες κλάσεων

Διαισθητικά, μία κλάση περιέχει όμοια αντικείμενα, ενώ αντικείμενα από διαφορετικές κλάσεις είναι ανόμοια. Μερικές κλάσεις έχουν ξεκάθαρο νόημα και στην απλούστερη περίπτωση είναι αμοιβαία αποκλειόμενες. Για παράδειγμα, στην πιστοποίηση υπογραφών, μία υπογραφή είναι είτε αυθεντική είτε πλαστή. Η πραγματική κλάση είναι μία από τις δύο, ανεξάρτητα από το αν υπάρχει περίπτωση να μην προβλεφθεί σωστά από την παρατήρηση της υπογραφής. Σε άλλα προβλήματα, μπορεί να είναι δύσκολο να καθοριστούν οι κλάσεις, όπως στις κλάσεις των δεξιόχειρων και των αριστερόχειρων ανθρώπων. Σε ιατρικά προβλήματα, υπάρχει μεγάλη δυσκολία στην αναπαράσταση των δεδομένων λόγω της ποικιλομορφίας του αντικειμένου της μελέτης. Για παράδειγμα, είναι συχνά επιθυμητό να γίνει διαχωρισμός του χαμηλού, μεσαίου και υψηλού κινδύνου εμφάνισης κάποιας

ασθένειας, αλλά είναι πολύ δύσκολο να οριστούν τα ακριβή κριτήρια που θα μπορέσουν να διαχωρίσουν αυτές τις κατηγορίες.

Σε ένα πρόβλημα ταξινόμησης, με την υπόθεση ότι υπάρχουν c κλάσεις, οι κλάσεις αυτές ορίζονται ως $\omega_1,...,\omega_c$ και αποτελούν στοιχεία του συνόλου των κλάσεων $\Omega = \{\omega_1,...,\omega_c\}$. Τα $\omega_1,...,\omega_c$ ονομάζονται class labels (ετικέτες κλάσεων). Επίσης πρέπει να τονιστεί ότι κάθε αντικείμενο του συνόλου των δεδομένων ανήκει σε μία και μόνο κλάση.

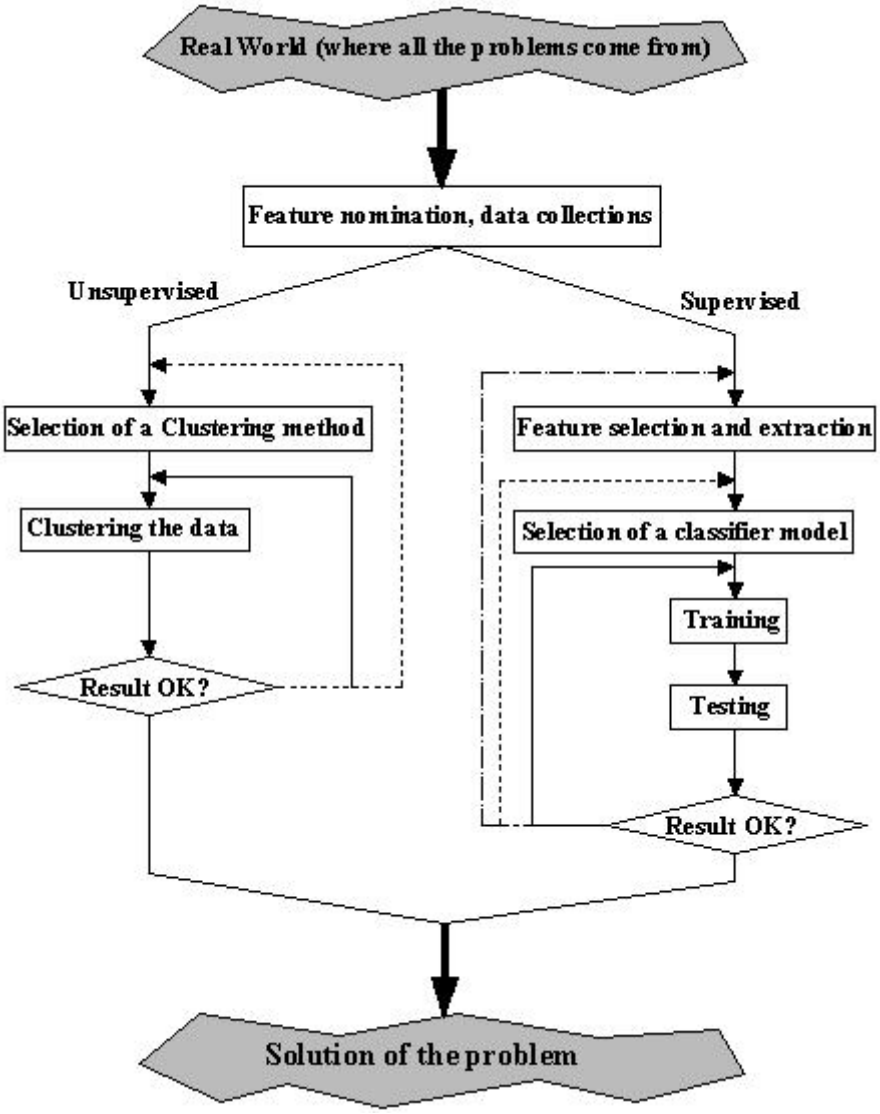
1.5 Σύνολο δεδομένων (Dataset)

Τα δεδομένα (πρότυπα) με τα οποία πρόκειται να γίνει ο σχεδιασμός του ταξινομητή αποτελούν ένα σύνολο που ονομάζεται σύνολο δεδομένων (dataset). Το σύνολο δεδομένων με N πρότυπα ορίζεται ως $Z = \{z_1,\ldots,z_N\}$, $z_j \in \Re^n$. Ο συμβολισμός z_j χρησιμοποιείται για να οριστούν τα αντικείμενα. Η ετικέτα της κλάσης (class label) του αντικειμένου z_j δηλώνεται ως $l(z_j) \in \Omega$, j = 1,…,N.

1.6 Διαδικασία ταξινόμησης

Η διαδικασία της αναγνώρισης προτύπων ακολουθεί πάντοτε κάποια βασικά βήματα [1]. Τα βήματα αυτά έχουν κάποιες διαφοροποιήσεις μεταξύ του supervised και του unsupervised learning, τα οποία θα αναλυθούν στην συνέχεια. Στο Σχήμα 1 φαίνονται οι βασικές λειτουργίες και τα στάδια της αναγνώρισης προτύπων. Με βάση το πρόβλημα και τα δεδομένα του χρήστη, ο σκοπός της αναγνώρισης προτύπων είναι να διευκρινίσει το πρόβλημα, να το μεταφράσει στην ορολογία της αναγνώρισης προτύπων και να το λύσει. Πιο συγκεκριμένα, θα πρέπει αρχικά να οριστεί το πρόβλημα και να σχεδιαστούν τα περαιτέρω βήματα που θα πρέπει να ακολουθηθούν προκειμένου να επιτευχθεί η λύση του. Στη συνέχεια, εάν το σύνολο των δεδομένων (dataset) δεν είναι εκ των προτέρων γνωστό, θα πρέπει να συλλεχθεί με την εξαγωγή κάποιων πειραμάτων. Το σύνολο των χαρακτηριστικών θα πρέπει να είναι όσο το

δυνατόν επαρκέστερο, ακόμη και να περιέχει χαρακτηριστικά τα οποία δεν φαίνονται να είναι τόσο σχετικά σε αυτό το στάδιο, γιατί μπορεί να είναι σχετικά σε συνδυασμό με άλλα χαρακτηριστικά. Τα όρια για τη συλλογή δεδομένων αφορούν την οικονομική πλευρά του προβλήματος. Βεβαίως, μεγάλος αριθμός χαρακτηριστικών απαιτεί και μεγάλο αριθμό δεδομένων και στις περιπτώσεις αυτές, εφαρμόζονται τεχνικές feature reduction, προκειμένου να μειωθεί ο αριθμός των χαρακτηριστικών. Ένα άλλο ενδεχόμενο όριο, μπορεί να είναι για παράδειγμα χαρακτηριστικά τα οποία είναι δύσκολο να μετρηθούν.



Σχήμα 1: Η διαδικασία ταξινόμησης σε supervised και unsupervised learning προβλήματα.

Τα χαρακτηριστικά δεν είναι όλα το ίδιο σχετικά. Μερικά από αυτά μπορεί να είναι σημαντικά μόνο σε συνδυασμό με κάποια άλλα, ενώ κάποια άλλα

χαρακτηριστικά μπορεί να μην είναι καθόλου χρήσιμα για το συγκεκριμένο πρόβλημα. Η επιλογή και η εξαγωγή των χαρακτηριστικών χρησιμοποιούνται για να βελτιώσουν την ποιότητα της περιγραφής των αντικειμένων.

Η επιλογή των χαρακτηριστικών (feature selection), η εκπαίδευση (training) και ο έλεγχος (testing) του μοντέλου του ταξινομητή αποτελούν τον πυρήνα της αναγνώρισης προτύπων με επίβλεψη. Όπως φαίνεται και στο προηγούμενο σχήμα από τις διακεκομμένες γραμμές, η επανάληψη που κάνουμε για να ρυθμίσουμε τον ταξινομητή μπορεί να γίνει σε διάφορα σημεία. Έτσι, μπορούμε να επιλέξουμε να χρησιμοποιήσουμε το ίδιο μοντέλο του ταξινομητή και να επαναλάβουμε τη διαδικασία της εκπαίδευσης με διαφορετικές αυτή τη φορά παραμέτρους ή και να αλλάξουμε το μοντέλο του ταξινομητή. Κάποιες φορές η επιλογή των χαρακτηριστικών μπορεί να περιλαμβάνεται στην επαναληπτική αυτή διαδικασία με σκοπό την ρύθμιση του μοντέλου. Όταν έχουμε μία ικανοποιητική λύση τότε μπορούμε να χρησιμοποιήσουμε αυτό το μοντέλο για περαιτέρω έλεγχο και εφαρμογές.

1.7 Διαχωρισμός του συνόλου δεδομένων

Σε προβλήματα ταξινόμησης μπορούμε να αξιολογήσουμε την απόδοση του μοντέλου με βάση το ποσοστό του σφάλματος (error rate), που είναι ο λόγος του αριθμού των αντικειμένων του dataset που ταξινομήθηκαν λάθος προς τον συνολικό αριθμό των αντικειμένων.

$$error\_rate = \frac{N_{misclassified\_objects}}{N_{objects}}$$

Ο λόγος για τον οποίο δημιουργούμε ένα μοντέλο ταξινόμησης είναι η κατηγοριοποίηση νέων δεδομένων και έτσι ενδιαφερόμαστε πρωτίστως για την απόδοση του μοντέλου στα νέα δεδομένα (τα οποία δεν τα έχει «δει» προηγουμένως ο ταξινομητής κατά τη διάρκεια της εκπαίδευσης). Για τον λόγο αυτόν, το ποσοστό του σφάλματος για το σύνολο των δεδομένων με το οποίο εκπαιδεύουμε το μοντέλο δεν

αποτελεί το κατάλληλο κριτήριο για την εκτίμηση της απόδοσης του ταξινομητή. Επιπλέον, εάν χρησιμοποιήσουμε τα ίδια δεδομένα για την εκπαίδευση και τον έλεγχο, αυτό θα έχει σαν πιθανό αποτέλεσμα το overtrain του ταξινομητή, δηλαδή το να μάθει απέξω την διαθέσιμη πληροφορία και να αποτυγχάνει σε πληροφορία την οποία δεν έχει δει κατά τη διάρκεια της εκπαίδευσης [2]. Με βάση τα παραπάνω, συμπεραίνουμε πως είναι σημαντικό να έχουμε ξεχωριστά δεδομένα για την εκπαίδευση και τον έλεγχο του ταξινομητή έτσι ώστε η τελική απόδοση να αντιστοιχεί στην πραγματικότητα.

Για να μπορέσουμε επομένως να μετρήσουμε την απόδοση του ταξινομητή όσο το δυνατό πιο δίκαια, χωρίζουμε το σύνολο με τα δεδομένα (dataset) σε τρία σύνολα: το σύνολο εκπαίδευσης (training set), το σύνολο επικύρωσης (validation set) και το σύνολο ελέγχου (test set). Ο ταξινομητής εκπαιδεύεται χρησιμοποιώντας το training set. Με το validation set αποφασίζουμε το σημείο εκείνο στο οποίο θα σταματήσουμε την εκπαίδευση ρυθμίζοντας ταυτόχρονα και κάποιες παραμέτρους του μοντέλου με σκοπό την καλύτερη επίδοση του ταξινομητή. Η εκπαίδευση του μοντέλου συνήθως σταματάει όταν η αύξηση της απόδοσης του training set δεν συνεπάγεται και ταυτόχρονη αύξηση της απόδοσης στο validation set και αυτό για να αποφύγουμε το overtrain του ταξινομητή. Στη συνέχεια, χρησιμοποιούμε το test set για να υπολογίσουμε την τελική απόδοση του μοντέλου που υλοποιήσαμε. Στο σημείο αυτό πρέπει να επισημάνουμε πως τόσο το training set όσο και το validation set δεν μπορούν να συμμετέχουν στον υπολογισμό της απόδοσης του ταξινομητή και αυτό γιατί περιέχουν δείγματα τα οποία έχει «μάθει» ο ταξινομητής να τα κατηγοριοποιεί και έτσι η απόδοση του μοντέλου θα ήταν μεγαλύτερη από ότι στην πραγματικότητα.

Από τα όσα αναφέρθηκαν πιο πάνω, είναι εύκολο να αντιληφθεί κανείς αφενός ότι όσο μεγαλύτερο είναι το σύνολο των δεδομένων με το οποίο θα εκπαιδεύσουμε τον ταξινομητή τόσο πιο καλό θα είναι το μοντέλο, αφού τόσο καλύτερη θα είναι η κάλυψη του χώρου των χαρακτηριστικών και αφετέρου ότι όσο πιο πολλά δεδομένα έχουμε στην διάθεση μας για τον έλεγχο του ταξινομητή τόσο πιο αντιπροσωπευτική θα είναι η εκτίμηση της απόδοσης του μοντέλου. Έτσι λοιπόν, παρόλο που με τον διαχωρισμό του αρχικού συνόλου δεδομένων σε τρία υποσύνολα επιτυγχάνεται καλύτερη εκπαίδευση του μοντέλου, στην πράξη η παραπάνω διαδικασία είναι

δύσκολο πολλές φορές να εφαρμοστεί και αυτό γιατί συνήθως έχουμε στην διάθεσή μας μικρά datasets.

Το πρόβλημα της ύπαρξης μικρών datasets λύνεται μερικώς με τον διαχωρισμό του αρχικού συνόλου δεδομένων σε δύο υποσύνολα: το training set και το test set. Έτσι, θα έχουμε στη διάθεση μας περισσότερα δεδομένα για την εκπαίδευση και τον έλεγχο του ταξινομητή. Επιπλέον, με τη μέθοδο αυτή, αντί να δεσμεύσουμε δεδομένα για το validation set (με το οποίο ουσιαστικά ρυθμίζουμε κάποιες παραμέτρους του μοντέλου και αποφασίζουμε το σημείο εκείνο στο οποίο τερματίζουμε την εκπαίδευση), χρησιμοποιούμε κάποιες μεθόδους με τις οποίες ελέγχουμε την εξέλιξη της πορείας της εκπαίδευσης. Συνήθως, από το αρχικό σύνολο δεδομένων το 70% χρησιμοποιείται ως training set και το υπόλοιπο 30% ως test set. Η αναλογία όμως αυτή δεν είναι απόλυτη αλλά εξαρτάται και από την κρίση του προγραμματιστή.

Και τα δύο υποσύνολα του αρχικού dataset (training και test set) πρέπει να περιέχουν αντιπροσωπευτικά δείγματα των δεδομένων στα οποία θα εφαρμοστεί το μοντέλο ταξινόμησης. Για τον διαχωρισμό του αρχικού dataset σε training set και σε test set ακολουθούνται διάφορες μέθοδοι, οι οποίες θα αναπτυχθούν στην συνέχεια.

1.8 Μέθοδοι διαχωρισμού του συνόλου δεδομένων

1.8.1 Resubstitution (R-method)

Με την R-method ουσιαστικά σχεδιάζουμε τον ταξινομητή D με την πληροφορία όλου του dataset Z και ελέγχουμε τον ταξινομητή με τα ίδια ακριβώς δεδομένα. Η μέθοδος αυτή, όπως αναφέρθηκε και στα προηγούμενα, δεν είναι καλό να χρησιμοποιείται γιατί ενέχει τον κίνδυνο του overtrain του ταξινομητή.

1.8.2 Hold-out (H-method)

Σύμφωνα με τη συγκεκριμένη μέθοδο, το dataset χωρίζεται σε δύο ίσα μέρη και το ένα μέρος χρησιμοποιείται για την εκπαίδευση του ταξινομητή και το άλλο μέρος για τον έλεγχο της απόδοσής του. Ακόμη, υπάρχει η δυνατότητα, να εναλλαχθούν τα δύο μέρη και να υπολογιστεί ξανά η απόδοση του ταξινομητή. Η τελική απόδοση θα είναι ο μέσος όρος των δύο επιμέρους αποδόσεων. Μία άλλη εκδοχή αυτής της μεθόδου είναι η επανάληψη της παραπάνω διαδικασίας για L φορές χωρίζοντας κάθε φορά το dataset τυχαία και υπολογίζοντας τον μέσο όρο των L επιμέρους υπολογισμών της απόδοσης.

1.8.3 Cross-validation

Με την συγκεκριμένη μέθοδο, γίνεται πιο αποδοτική η χρήση των δεδομένων ακόμη και σε μικρά datasets. Σύμφωνα με την cross-validation μέθοδο, επιλέγεται ένας ακέραιος k, ο οποίος είναι προτιμητέο να είναι παράγοντας του αριθμού N των αντικειμένων του dataset Z. Στη συνέχεια, το Z χωρίζεται σε k υποσύνολα (folds) μεγέθους N/k το καθένα. Ο ταξινομητής εκπαιδεύεται στα k-1 folds, ενώ το άλλο χρησιμοποιείται για τον έλεγχο της απόδοσης. Η διαδικασία επαναλαμβάνεται k φορές, έτσι ώστε όλα τα folds να χρησιμοποιηθούν για τον έλεγχο του ταξινομητή. Η συνολική απόδοση υπολογίζεται ως ο μέσος όρος των αποδόσεων των k test sets. Έτσι η μέθοδος αυτή, χρησιμοποιεί αποτελεσματικά όλα τα διαθέσιμα δεδομένα τόσο για την εκπαίδευση όσο και για τον έλεγχο του ταξινομητή.

1.8.4 Leave-one-out

Η μέθοδος αυτή είναι μία παραλλαγή της cross-validation. Πιο συγκεκριμένα, αν στην cross-validation ο αριθμός των folds (k) γίνει ίσος με τον αριθμό των δεδομένων (N), προκύπτει η leave-one-out μέθοδος. Στην περίπτωση αυτή, το test set αποτελείται από ένα μόνο δείγμα (το οποίο θα ταξινομηθεί είτε σωστά είτε λάθος), ενώ τα υπόλοιπα δείγματα χρησιμοποιούνται για την εκπαίδευση του ταξινομητή. Η

διαδικασία επαναλαμβάνεται Ν φορές ώστε όλα τα δείγματα του dataset να αποτελέσουν test set. Η απόδοση του ταξινομητή υπολογίζεται ως ο μέσος όρος των Ν μεμονωμένων αποδόσεων. Το πλεονέκτημα αυτής της μεθόδου είναι η μέγιστη χρήση του training set (χρησιμοποιούνται Ν-1 δείγματα κάθε φορά). Το κύριο μειονέκτημα αυτής της μεθόδου είναι η δυσκολία εφαρμογής της σε πολύ μεγάλα datasets. Σε αυτές τις περιπτώσεις, απαιτείται μεγάλος αριθμός επαναλήψεων και το υπολογιστικό κόστος είναι υψηλό. Επιπλέον, τις περισσότερες φορές με τη συγκεκριμένη μέθοδο γίνεται υπερεκτίμηση της απόδοσης του ταξινομητή.

1.8.5 Stratified k-fold cross validation

Στις προηγούμενες μεθόδους, τα δείγματα λαμβάνονται τυχαία. Αυτό μπορεί να επηρεάσει αρνητικά το μοντέλο, αφού ο διαχωρισμός του dataset σε training και σε test set μπορεί και να μην είναι αντιπροσωπευτικός. Για παράδειγμα, υπάρχει περίπτωση με τον τυχαίο διαχωρισμό των υποσυνόλων, στο training set να μην εμφανίζεται ούτε ένα δείγμα από μία συγκεκριμένη κλάση. Αυτό θα έχει σαν αποτέλεσμα, ο ταξινομητής να μην μάθει να κατηγοριοποιεί αντικείμενα αυτής της κλάσης. Την λύση στο πρόβλημα αυτό, δίνει η χρήση της stratified k-fold cross validation μεθόδου. Με την μέθοδο αυτή, βρίσκονται οι αναλογίες των κλάσεων στο αρχικό σύνολο δεδομένων και ο περαιτέρω διαχωρισμός τους γίνεται όπως ακριβώς και στην cross-validation αλλά με τον περιορισμό σε κάθε fold να διατηρούνται οι αρχικές αναλογίες των κλάσεων.

1.9 Μέτρηση της απόδοσης του ταξινομητή

Στην παραπάνω ενότητα μελετήθηκαν οι διάφοροι τρόποι με τους οποίους μπορεί να χωριστεί το σύνολο των δεδομένων σε ένα υποσύνολο για την εκπαίδευση και σε ένα δεύτερο υποσύνολο για τον έλεγχο του μοντέλου ταξινόμησης. Όπως αναφέρθηκε και προηγουμένως, ο υπολογισμός της απόδοσης του ταξινομητή αναφέρεται μόνο στο κομμάτι εκείνο των δεδομένων που χρησιμοποιείται για τον έλεγχο του μοντέλου και όχι στο σύνολο τους γενικά. Τον περιορισμό αυτό τον

λαμβάνουμε υπόψη μας γιατί τα δεδομένα του συνόλου ελέγχου είναι άγνωστα στον ταξινομητή έως το πέρας της εκπαίδευσής του και έτσι θα μπορούμε να έχουμε μία εικόνα της απόδοσης του μοντέλου ταξινόμησης, η οποία να ανταποκρίνεται στην πραγματικότητα.

Η απόδοση ενός ταξινομητή είναι ένα σύνθετο χαρακτηριστικό, του οποίου το πιο κύριο στοιχείο είναι η ακρίβεια της ταξινόμησης (classification accuracy). Άλλα μέτρα που χρησιμοποιούνται για την εκτίμηση της απόδοσης είναι τα Precision, Recall, Specificity, Sensitivity, το εμβαδό της καμπύλης ROC κ.ά. τα οποία θα αναλυθούν στις επόμενες ενότητες. Εάν ο ταξινομητής μπορούσε να ελεγχθεί σε όλα τα πιθανά δείγματα εισόδου θα μπορούσε να υπολογιστεί η ακριβής απόδοσή του. Επειδή κάτι τέτοιο δεν είναι εφικτό, χρησιμοποιείται μία εκτίμηση της ακρίβειάς του.

Η ακρίβεια της ταξινόμησης δείχνει το κατά πόσο ένας ταξινομητής λειτουργεί σωστά και υπολογίζεται ως ο λόγος των αντικειμένων που ταξινομήθηκαν σωστά προς τον συνολικό αριθμό των αντικειμένων:

$$classification\_accuracy = \frac{N_{successful\_classifications}}{N_{classifiacations}}$$

Η ακρίβεια της ταξινόμησης είναι πολύ εύκολο να υλοποιηθεί. Για να είναι όμως, τα αποτελέσματα που δίνει αντιπροσωπευτικά, θα πρέπει το test set να διατηρεί τις αναλογίες των κλάσεων του αρχικού dataset. Πολλές φορές, είναι σημαντικό, εκτός από την ακρίβεια του ταξινομητή να είναι γνωστή και η κατανομή των λανθασμένων ταξινομήσεων στις κλάσεις. Στην επόμενη ενότητα, θα οριστεί η έννοια του confusion matrix, ο οποίος περιγράφει τη γενική κατανομή των ταξινομήσεων.

1.9.1 Confusion Matrix

Όπως αναφέρθηκε και στα προηγούμενα, σε ένα πρόβλημα ταξινόμησης είναι πολλές φορές χρήσιμο εκτός από την γενική εικόνα της ακρίβειας του ταξινομητή, να υπάρχει πληροφορία και για την κατανομή των εσφαλμένων ταξινομήσεων στις

κλάσεις. Η κατανομή αυτή μπορεί να οργανωθεί σε έναν πίνακα, που ονομάζεται confusion matrix. Η δομή ενός confusion matrix για ένα πρόβλημα ταξινόμησης με δύο κλάσεις φαίνεται στον Πίνακα 1:

Πραγματική κλάση	Απόφαση του ταξινομητή	
	ω_1	ω_2
ω_1	$\omega_1\ \omega_1$	$\omega_1\ \omega_2$
ω_2	$\omega_2\ \omega_1$	$\omega_2\ \omega_2$

Πίνακας 1: Confusion matrix

Κάθε στοιχείο a_{ij} του confusion matrix δηλώνει τον αριθμό των στοιχείων του test set του οποίου η πραγματική κλάση είναι η ω_i και ο ταξινομητής D το αντιστοίχισε στην κλάση ω_j. Προφανώς, μπορεί να υπολογιστεί εύκολα η ακρίβεια της ταξινόμησης από τον παραπάνω πίνακα από τη σχέση:

$$classification\_accuracy = \frac{\omega_1\omega_1 + \omega_2\omega_2}{\omega_1\omega_1 + \omega_1\omega_2 + \omega_2\omega_1 + \omega_2\omega_2}$$

Ο confusion matrix μπορεί να επεκταθεί και για περισσότερες κλάσεις από δύο, που φαίνονται στον παραπάνω πίνακα. Μία άλλη μορφή του confusion matrix, που χρησιμοποιείται κυρίως σε βιοϊατρικά προβλήματα ταξινόμησης, που το ζητούμενο είναι η ταξινόμηση ασθενών που είτε νοσούν από κάποια ασθένεια (κλάση ω_1) είτε όχι (κλάση ω_2), είναι η αυτή του Πίνακα 2:

	Positive	Negative
True	TP	TN
False	FP	FN

Πίνακας 2: Μία εναλλακτική μορφή του confusion matrix

Τα στοιχεία του παραπάνω πίνακα μπορούν να μεταφραστούν ως εξής:

- **True Positive (TP)**: αναπαριστά τους ανθρώπους που νοσούν από την ασθένεια και που το μοντέλο ανίχνευσε (σωστά) την ύπαρξη της ασθένειας
- **True Negative (TN)**: αναπαριστά τους ανθρώπους που δεν νοσούν από την ασθένεια και το μοντέλο ανίχνευσε (σωστά) την μη ύπαρξη της ασθένειας
- **False Positive (FP)**: αναπαριστά τους ανθρώπους που δεν νοσούν από την ασθένεια ενώ το μοντέλο ανίχνευσε (λανθασμένα) την ύπαρξη της ασθένειας
- **False Negative (FN)**: αναπαριστά τους ανθρώπους που νοσούν από την ασθένεια ενώ το μοντέλο ανίχνευσε (λανθασμένα) την μη ύπαρξη της ασθένειας

Με βάση τα παραπάνω, ορίζονται κάποιες έννοιες, οι οποίες είναι πολύ σημαντικές στην εκτίμηση των μοντέλων ταξινόμησης:

- **True Positive Rate ή Sensitivity (TPR)**: είναι το ποσοστό των θετικών περιπτώσεων που ταξινομήθηκαν σωστά

$$TPR = \frac{TP}{FN + TP}$$

- **True Negative Rate ή Specificity (TNR)**: είναι το ποσοστό των αρνητικών περιπτώσεων που ταξινομήθηκαν σωστά

$$TNR = \frac{TN}{TN + FP}$$

- **False Positive Rate (FPR)**: είναι το ποσοστό των αρνητικών περιπτώσεων που εσφαλμένα ταξινομήθηκαν ως θετικές

$$FPR = \frac{FP}{TN + FP}$$

- **False Negative Rate (FNR)**: είναι το ποσοστό των θετικών περιπτώσεων που εσφαλμένα ταξινομήθηκαν ως αρνητικές

$$FNR = \frac{FN}{FN + TP}$$

- **Precision**: είναι ο λόγος των σωστά ταξινομημένων θετικών περιπτώσεων προς όλες τις περιπτώσεις που ταξινομήθηκαν ως θετικές

$$\Pr ecision = \frac{TP}{TP + FP}$$

- **Recall**: είναι ο λόγος των σωστά ταξινομημένων θετικών περιπτώσεων προς όλες τις θετικές περιπτώσεις

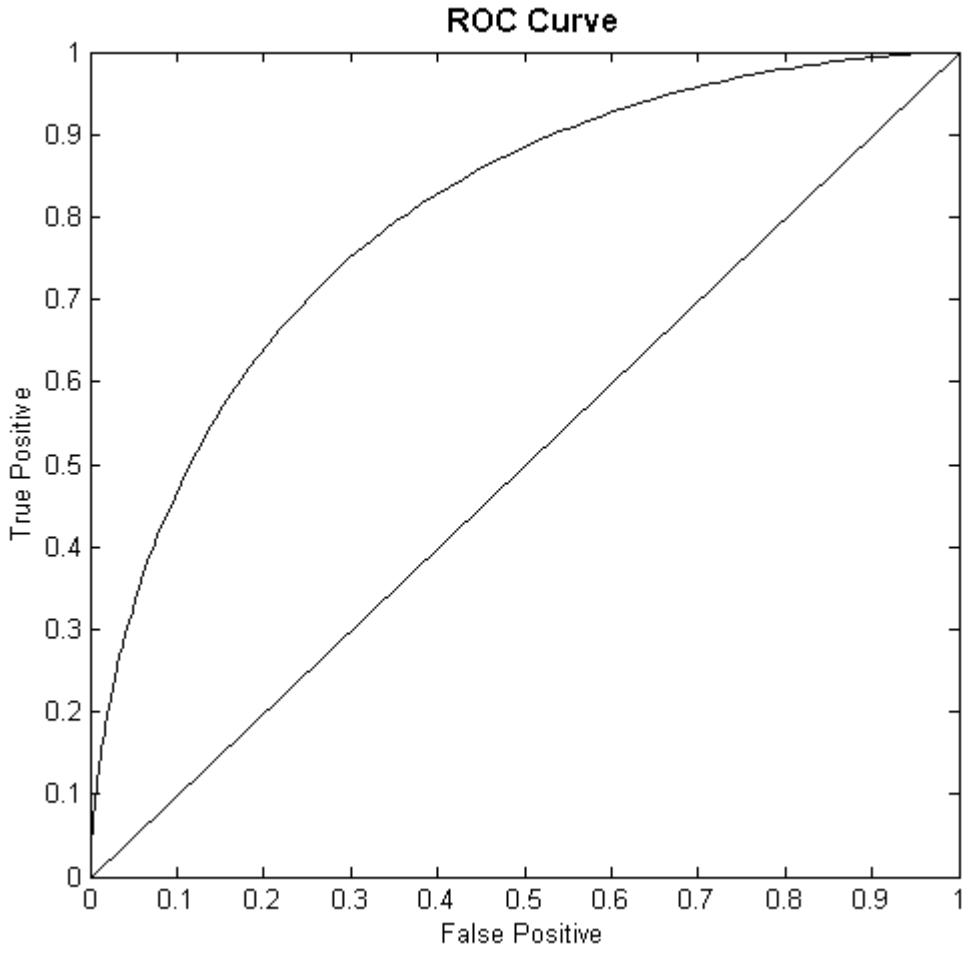
$$\mathrm{Re} call = \frac{TP}{TP + FN}$$

1.9.2 Receiver Operating Characteristic (ROC)

Η ακρίβεια ενός ταξινομητή μετρημένη ως το ποσοστό των σωστών (ή λάθος) ταξινομήσεων, δεν αποτελεί πάντοτε μία αντικειμενική ένδειξη της απόδοσης του ταξινομητή. Για παράδειγμα, έστω ότι σε ένα πρόβλημα ταξινόμησης δύο κλάσεων, το ποσοστό των αντικειμένων της πρώτης κλάσης είναι 5% του συνολικού αριθμού των αντικειμένων και ένας ταξινομητής μπορεί να ταξινομήσει σωστά τα αντικείμενα μόνο της δεύτερης κλάσης. Στην περίπτωση αυτή, ο ταξινομητής θα εμφανίζει μία ακρίβεια της τάξης του 95% χωρίς όμως να μπορεί να ταξινομήσει έστω και ένα αντικείμενο από την πρώτη κλάση.

Την λύση στο πρόβλημα αυτό, δίνει η χρήση των Receiver Operating Characteristic (ROC) καμπύλων, οι οποίες απεικονίζουν την απόδοση ενός ταξινομητή, ανεξάρτητα της κατανομής των κλάσεων ή του κόστους των σφαλμάτων ([4], [5]). Οι ROC καμπύλες αναπτύχθηκαν την δεκαετία του 1940 για την θεωρία ανίχνευσης σήματος (signal detection theory) για την ανάλυση θορύβου στα σήματα, χρησιμοποιούνται όμως ευρέως από την δεκαετία του 1970 για την αναπαράσταση

των αποτελεσμάτων σε βιοϊατρικά προβλήματα. Ένα παράδειγμα καμπύλης ROC φαίνεται στο Σχήμα 2 που ακολουθεί.



Σχήμα 2: Παράδειγμα καμπύλης ROC. Η διαγώνιος αναπαριστά την περίπτωση της τυχαίας πρόβλεψης.

Όπως φαίνεται και στο παραπάνω σχήμα μία ROC καμπύλη αποτελεί την γραφική αναπαράσταση του sensitivity (TPR) συναρτήσει του 1-specifity (FPR). Υπάρχουν τρία σημεία στην ROC καμπύλη, στα οποία πρέπει να δοθεί ιδιαίτερη έμφαση. Το σημείο με συντεταγμένες (0,0) αναπαριστά την περίπτωση που όλα τα αντικείμενα ταξινομούνται ως αρνητικά. Το σημείο με συντεταγμένες (1,1) αναπαριστά την περίπτωση που όλα τα αντικείμενα ταξινομούνται ως θετικά. Το σημείο με συντεταγμένες (1,0) αναπαριστά την ιδανική περίπτωση (όπου δηλαδή υπάρχει τέλειος διαχωρισμός μεταξύ των δύο κλάσεων). Η διαγώνιος (η ευθεία γραμμή που ενώνει τα σημεία (0,0) και (1,1)) στην ROC καμπύλη αναπαριστά την περίπτωση της τυχαίας πρόβλεψης.

Αφού διατυπώθηκαν ορισμένα σημαντικά σημεία για τις ROC καμπύλες, αξίζει να περιγραφούν και τα βήματα που ακολουθούνται για την κατασκευή τους. Καταρχήν, θα πρέπει ο ταξινομητής να παράγει για κάθε δείγμα μία τιμή, η οποία θα δηλώνει την πιθανότητα το δείγμα να ανήκει στην θετική κλάση. Στη συνέχεια, οι τιμές αυτές ταξινομούνται κατά φθίνουσα σειρά και ακολουθούνται τα παρακάτω βήματα:

1. Αρχικά σημειώνεται το σημείο με συντεταγμένες (0,0), στο οποίο TPR = 0 και FPR = 0. Στο σημείο αυτό το κατώφλι έχει την τιμή 1. Δηλαδή τα δείγματα που ταξινομούνται στην θετική κλάση είναι αυτά που έχουν πιθανότητα μεγαλύτερη ή ίση του 1 και άρα όλα τα δείγματα ταξινομούνται ως αρνητικά.
2. Στη συνέχεια για κάθε πιθανότητα, θεωρείται ως κατώφλι η τιμή της πιθανότητας και επομένως όσα δείγματα έχουν πιθανότητα να ανήκουν στην θετική κλάση μεγαλύτερη ή ίση της τιμής του κατωφλίου ταξινομούνται στην θετική κλάση ενώ τα υπόλοιπα στην αρνητική κλάση. Στη συνέχεια, με βάση και τις πραγματικές κλάσεις των δειγμάτων μετρούνται τα TP, FP, TN, FN, και στη συνέχεια υπολογίζονται τα TPR, FPR (TPR = TP/(TP+FN) και FPR = FP/(FP+TN)). Για κάθε κατώφλι, το ζεύγος (FPR, TPR) δίνει τις συντεταγμένες του αντίστοιχου σημείου, τα οποία θα σχηματίσουν την καμπύλη ROC.
3. Ως τελικό σημείο της καμπύλης ROC προκύπτει το σημείο με συντεταγμένες (1,1) στο οποίο όλα τα δείγματα ταξινομούνται ως θετικά.

Είναι προφανές ότι στην ιδανική περίπτωση ενός ταξινομητή που έχει την ικανότητα να ταξινομεί σωστά όλα τα δείγματα, η ROC καμπύλη πρώτα θα ανέβει προς τα πάνω στον άξονα των TPR και στη συνέχεια θα μετακινηθεί προς τα δεξιά, παράλληλα με τον άξονα των FPR.

Ένα χρήσιμο στοιχείο για την εκτίμηση της απόδοσης ενός ταξινομητή αλλά κυρίως για τη σύγκρισή της με τις αποδόσεις άλλων ταξινομητών, είναι το εμβαδό της περιοχής κάτω από την καμπύλη ROC, AUROC (Area Under the ROC Curve). Στην περίπτωση το βέλτιστου ταξινομητή το AUROC έχει την τιμή 1. Στην περίπτωση που ο ταξινομητής κάνει τυχαίες αντιστοιχίσεις (και όπως αναφέρθηκε και προηγουμένως σε αυτήν την περίπτωση η ROC καμπύλη είναι η διαγώνιος) το AUROC έχει την τιμή

0.5. Το εμβαδό κάτω από την ROC καμπύλη είναι πολύ χρήσιμο κυρίως σε περιπτώσεις που οι ROC καμπύλες δύο ή και περισσότερων ταξινομητών τέμνονται, και έτσι δεν μπορεί να εκτιμηθεί με το μάτι ποιος από αυτούς έχει την καλύτερη απόδοση. Σε αυτήν την περίπτωση, καλύτερη απόδοση εμφανίζει ο ταξινομητής με τη μεγαλύτερη τιμή της AUROC.

ΚΕΦΑΛΑΙΟ 2

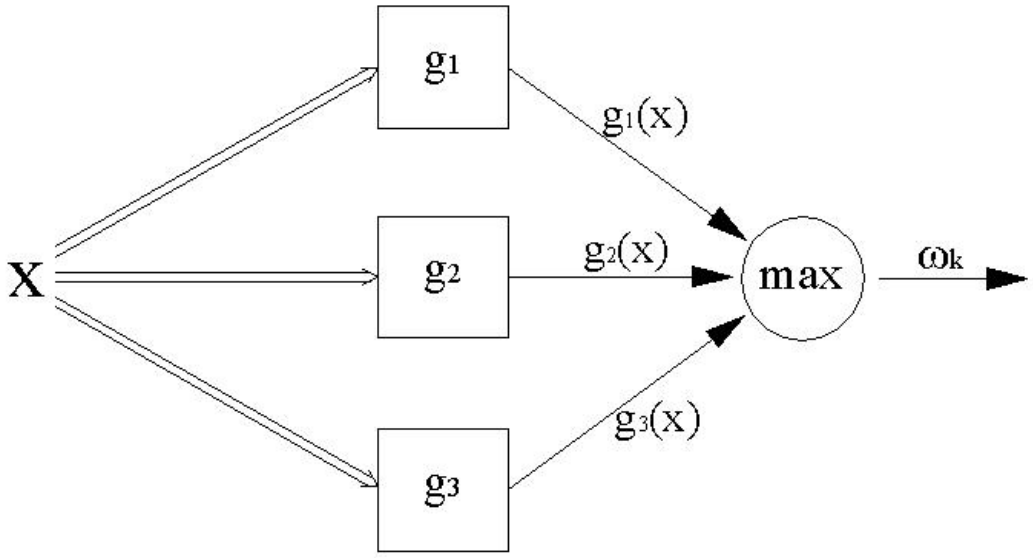
ΒΑΣΙΚΗ ΘΕΩΡΙΑ ΤΑΞΙΝΟΜΗΤΩΝ

2.1 Εισαγωγή

Στο κεφάλαιο αυτό, θα γίνει μία σύντομη αναφορά στη γενική λειτουργία των βασικών ταξινομητών (base classifiers) πρώτου επιπέδου. Οι ταξινομητές αυτοί, λειτουργούν μεμονωμένα (πρώτο επίπεδο) και τα αποτελέσματά τους, μπορούν να συνδυαστούν με διάφορες μεθόδους (δεύτερο επίπεδο), με σκοπό τη βελτίωση της απόδοσης του μοντέλου. Οι μέθοδοι συνδυασμού των ταξινομητών, θα μελετηθούν στο επόμενο κεφάλαιο.

Έστω ότι ένα αντικείμενο $\mathbf{x}$ περιγράφεται από n χαρακτηριστικά (features) και ανήκει σε μία κλάση του συνόλου Ω. Ο όρος «ταξινομητής» (classifier) αναφέρεται σε μία οποιαδήποτε συνάρτηση, η οποία αντιστοιχίζει το αντικείμενο $\mathbf{x} \in \Re^n$ σε μία κλάση $\omega \in \Omega$. Δηλαδή, ένας ταξινομητής D, αντιστοιχίζει ένα στοιχείο του συνόλου του διαστήματος χαρακτηριστικών $\Re^n$, σε ένα στοιχείο του συνόλου Ω ($\mathbf{D}: \Re^n \to \Omega$).

Σε ένα πρόβλημα ταξινόμησης με c κλάσεις, ένας ταξινομητής δημιουργεί c συναρτήσεις διαχωρισμού (discriminant functions) $G = \{g_1(\mathbf{x}), \ldots, g_c(\mathbf{x})\}$. Κάθε μία από τις συναρτήσεις διαχωρισμού αποδίδει την πιθανότητα να ανήκει το αντικείμενο στη συγκεκριμένη κλάση. Το αντικείμενο αντιστοιχίζεται τελικά στην κλάση με την μεγαλύτερη πιθανότητα. Σε περίπτωση που δύο ή παραπάνω κλάσεις έχουν ίσες πιθανότητες αντιστοίχησης, το αντικείμενο αντιστοιχίζεται σε μία από αυτές τυχαία. Η παραπάνω διαδικασία φαίνεται στο Σχήμα 3.

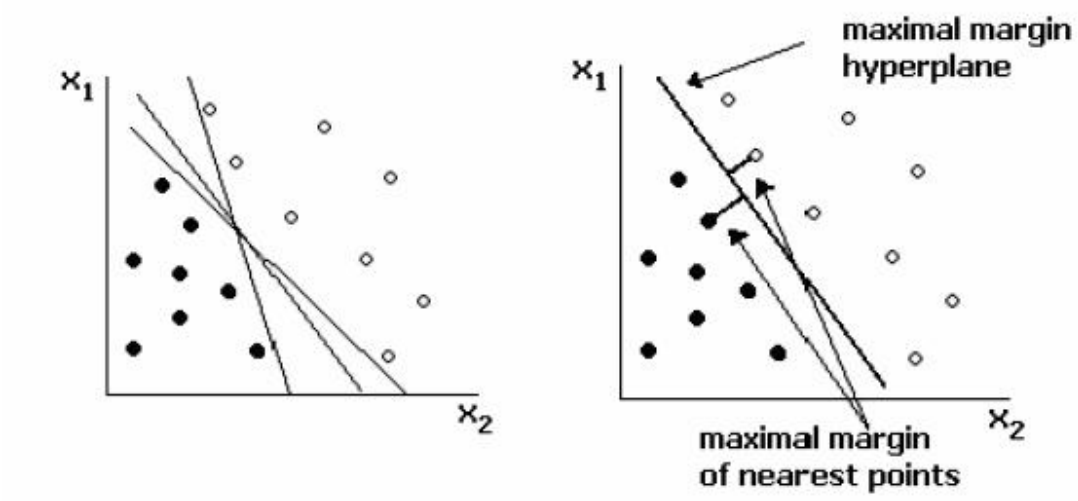


Σχήμα 3: Σχηματικό Διάγραμμα της γενικής λειτουργίας ενός ταξινομητή. Τα διπλά βέλη δηλώνουν ότι η είσοδος είναι ένα n-διάστατο διάνυσμα x.

2.2 Support Vector Machines (SVMs)

Τα Support Vector Machines (SVMs) είναι αποτελεσματικές μηχανές μάθησης που χρησιμοποιούνται σε supervised learning προβλήματα ταξινόμησης [6]. Η φιλοσοφία τους βασίζεται στην κατασκευή ενός υπερεπίπεδου (hyperplane) που θα διαχωρίζει τις κλάσεις και που ταυτόχρονα θα μεγιστοποιεί το περιθώριο ανάμεσά τους, γεγονός που τους δίνει μεγαλύτερη ικανότητα γενίκευσης. Για τον λόγο αυτό, είναι γνωστά και ως ταξινομητές μέγιστου περιθωρίου (maximum margin classifiers).

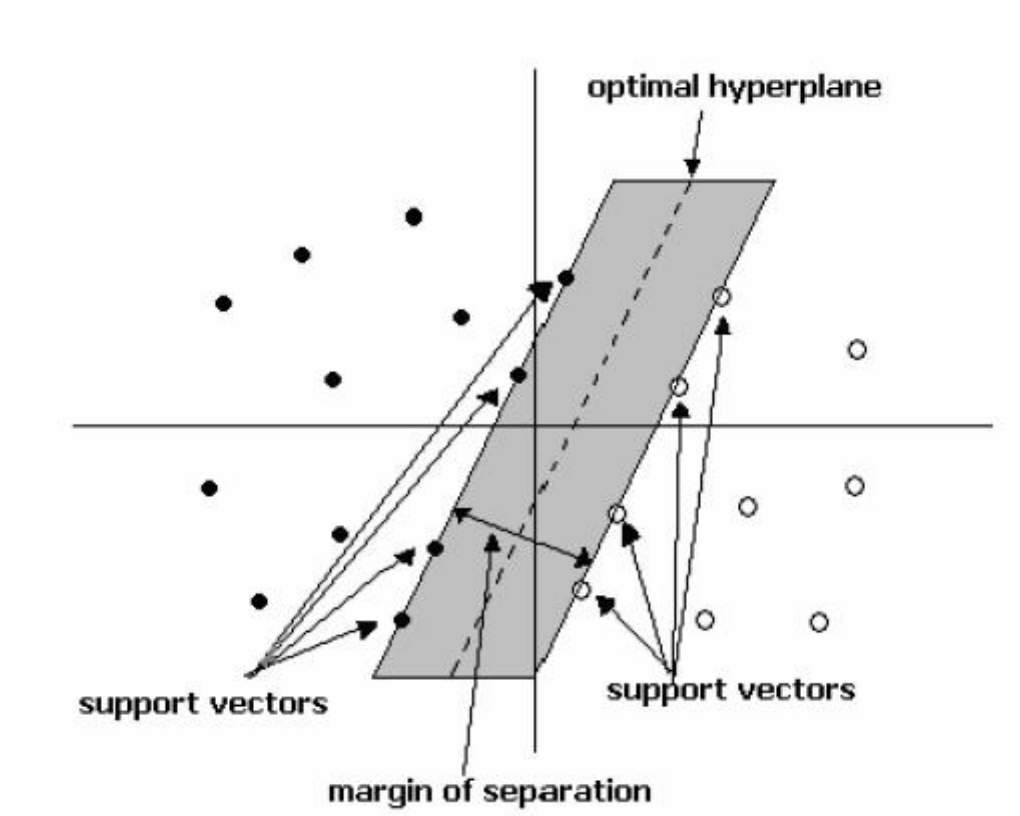
Τα SVMs αντιστοιχίζουν τα διανύσματα εισόδου (διανύσματα που φέρουν τις τιμές των χαρακτηριστικών που περιγράφουν τα αντικείμενα) σε έναν πολυδιάστατο χώρο, όπου θα κατασκευαστεί το υπερεπίπεδο. Για γραμμικά διαχωρίσιμα δεδομένα, όπως φαίνεται και στο Σχήμα 4, είναι προφανές ότι δεν υπάρχει ένα μοναδικό υπερεπίπεδο που να διαχωρίζει τις κλάσεις, εκτός κι αν υπάρχουν περιορισμοί που πρέπει να ικανοποιούνται.



Σχήμα 4: Για γραμμικά διαχωρίσιμα δεδομένα δεν υπάρχει ένα μοναδικό υπερεπίπεδο που να διαχωρίζει τις κλάσεις, εκτός κι αν τεθούν κάποιοι περιορισμοί.

Η εύρεση του βέλτιστου υπερεπιπέδου, βασίζεται στην εύρεση υπερεπιπέδων με μεγάλο περιθώριο. Η ιδέα αυτή προήλθε από την έννοια της μάθησης, όπου η αναγνώριση κάποιας κλάσης βασίζεται στην εξέταση των σημείων που αναπαριστούν τα δεδομένα εκπαίδευσης (training points) της συγκεκριμένης κλάσης. Τα νέα δεδομένα (που αναπαρίστανται και αυτά ως σημεία) λογικά θα βρίσκονται κάπου κοντά στα ήδη γνωστά training points. Έτσι, η επιλογή του υπερεπιπέδου πρέπει να είναι τέτοια, ώστε μία μικρή απόκλιση στα δεδομένα να μην προκαλεί λάθος στην ταξινόμηση των νέων δεδομένων και έτσι να μειώνεται η πιθανότητα του σφάλματος ταξινόμησης. Για το λόγο αυτό προτιμάται ένα υπερεπίπεδο με μεγάλο περιθώριο.

Βασισμένοι σε αυτήν τη λογική οι Vapnik και Chervonenkis, πρότειναν έναν αλγόριθμο μάθησης για προβλήματα που μπορούν να χωριστούν από υπερεπίπεδα. Απεδείχθη, ότι από όλα τα υπερεπίπεδα που διαχωρίζουν τα δεδομένα, υπάρχει ένα μοναδικό βέλτιστο, το οποίο είναι αυτό, στο οποίο οι δυο κλάσεις διαχωρίζονται με το μέγιστο περιθώριο (Σχήμα 5). Τα σημεία που βρίσκονται πάνω στα όρια του περιθωρίου, δημιουργούν δύο παράλληλα υπερεπίπεδα ως προς το βέλτιστο, τα οποία ονομάζονται support vectors.



Σχήμα 5: Το βέλτιστο υπερεπίπεδο που διαχωρίζει τις κλάσεις είναι αυτό που βασίζεται στην μεγιστοποίηση του περιθωρίου.

Έστω ότι το σύνολο των δεδομένων απεικονίζεται ως τα σημεία $\{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_N, c_N)\}$, όπου το c_i είναι -1 ή 1 και δηλώνει την κλάση στην οποία ανήκει το δείγμα $\mathbf{x}_i$. Κάθε δείγμα $\mathbf{x}_i$ είναι ένα n-διάστατο διάνυσμα (αφού περιγράφεται από n χαρακτηριστικά). Το σύνολο αυτών των σημείων μπορεί να θεωρηθεί ως το training set με το οποίο θα εκπαιδευτεί το SVM. Η εκπαίδευση γίνεται με τη βοήθεια του υπερεπιπέδου που παίρνει τη μορφή:

$$w \cdot x - b = 0 .$$

Το διάνυσμα w είναι κάθετο στο διαχωριστικό υπερεπίπεδο και προσθέτοντας την παράμετρο b δίνεται η δυνατότητα αύξησης του περιθωρίου. Από τη στιγμή που αυτό που έχει κυρίως σημασία είναι το μέγιστο περιθώριο, η προσοχή στρέφεται στα παράλληλα υπερεπίπεδα (support vectors). Μπορεί να αποδειχθεί ότι αυτά τα support vectors μπορούν να περιγραφούν από τις παρακάτω σχέσεις:

$$w \cdot x - b = 1 \quad \text{και} \quad w \cdot x - b = -1$$

Εάν τα δεδομένα εκπαίδευσης είναι γραμμικά διαχωρίσιμα, τα υπερεπίπεδα μπορούν να επιλεγούν έτσι ώστε να μην υπάρχει κανένα σημείο ανάμεσά τους και στην συνέχεια να γίνει προσπάθεια να μεγιστοποιηθεί η μεταξύ τους απόσταση. Μπορεί να βρεθεί ότι η απόσταση μεταξύ των δύο υπερεπιπέδων είναι $\frac{2}{|w|}$ και έτσι για να μεγιστοποιηθεί αυτή η απόσταση, χρειάζεται να ελαχιστοποιηθεί ο όρος $|w|$. Για να αποκλειστεί η πιθανότητα ύπαρξης σημείων μεταξύ των δύο υπερεπιπέδων θα πρέπει για όλα τα σημεία i να ισχύει μία από τις παρακάτω ανισότητες:

$$w \cdot x_i - b \geq 1 \quad \text{ή} \quad w \cdot x_i - b \leq -1 .$$

Το οποίο μπορεί να γραφτεί και ως : $c_i(w \cdot x_i - b) \geq 1,\ 1 \leq i \leq N$. Το πρόβλημα πλέον βρίσκεται στην ελαχιστοποίηση του όρου $|w|$ αλλά χωρίς την παραβίαση της παραπάνω ανισότητας.

Εκτός, όμως, από την περίπτωση των γραμμικά διαχωρίσιμων δεδομένων τα SVMs μπορούν να λειτουργήσουν και σε περιπτώσεις που τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα στον χώρο που αναπαρίστανται. Αυτό, γίνεται εύκολα, κάνοντας χρήση του λεγόμενου kernel trick στα υπερεπίπεδα μέγιστου περιθωρίου. Στην περίπτωση αυτή, τα δεδομένα θα μπορούσαν να αντιστοιχηθούν σε ένα χώρο χαρακτηριστικών (feature space) υψηλών διαστάσεων (μέσω μιας συνάρτησης φ(x)), στον οποίο θα κατασκευαστεί το βέλτιστο υπερεπίπεδο που θα διαχωρίζει γραμμικά τα δεδομένα. Στη συνέχεια, θα μπορούσε και πάλι να αντιστοιχηθεί το υπερεπίπεδο στον αρχικό χώρο (με την αντίστροφη της φ(x)), όπου θα έχει μη γραμμική μορφή. Οι δύο αυτές αντιστοιχίσεις των δεδομένων από τον αρχικό χώρο στον χώρο χαρακτηριστικών και έπειτα πάλι στον αρχικό χώρο θα γίνονταν με το εσωτερικό γινόμενο $\phi(x_i)^T \phi(x_j)$. Οι πράξεις όμως σε ένα χώρο πολυδιάστατο είναι αρκετά πολύπλοκες και έτσι με τη χρήση του kernel trick ουσιαστικά το εσωτερικό γινόμενο $\phi(x_i)^T \phi(x_j)$ αντικαθίσταται με μη γραμμικές kernel συναρτήσεις $(k(x_i, x_j))$ και έτσι

αποφεύγονται οι υπολογισμοί στον χώρο υψηλών διαστάσεων. Ισχύει δηλαδή η σχέση:

$$\left(k\left(x_i, x_j\right)\right) = \phi\left(x_i\right)^T \phi\left(x_j\right)$$

Δύο από τους πιο συνηθισμένους kernels είναι :

- Ο πολυωνυμικός kernel : $k\left(x, x'\right) = \left(x \cdot x' + 1\right)^d$
- Η Radial Basis Function : $k\left(x, x'\right) = \exp\left(-\frac{\left\|x - x'\right\|^2}{2\sigma^2}\right)$

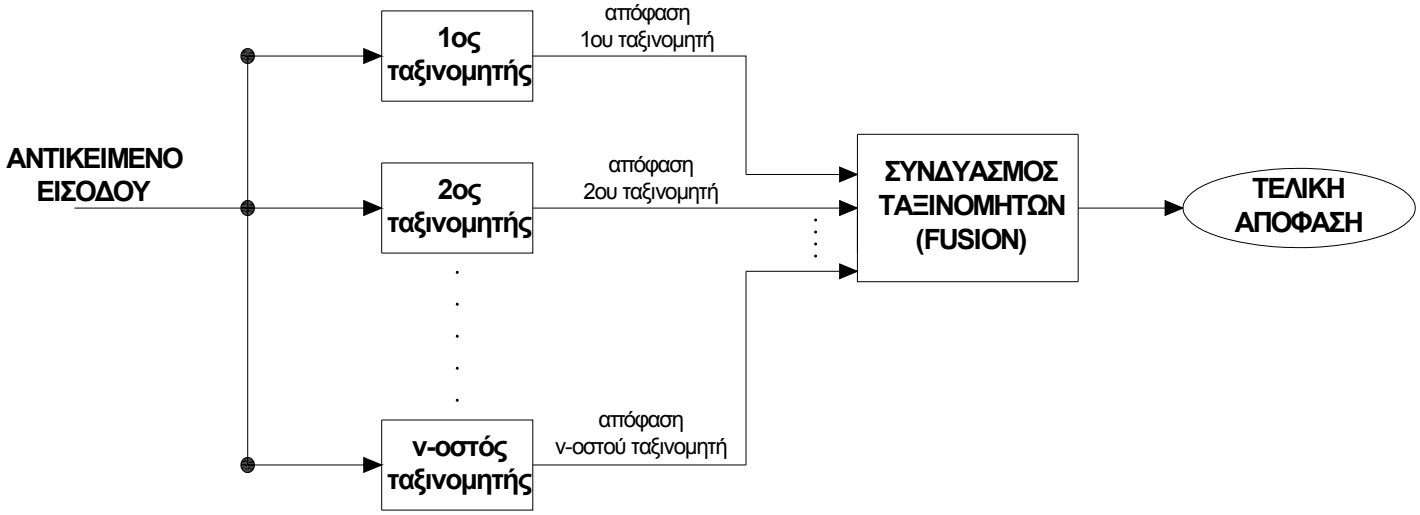
ΚΕΦΑΛΑΙΟ 3

ΣΥΝΔΥΑΣΜΟΣ ΤΑΞΙΝΟΜΗΤΩΝ

3.1 Συνδυασμός ταξινομητών (Classifier Fusion)

Στο κεφάλαιο 2, δόθηκε μία σύντομη περιγραφή της λειτουργίας των ταξινομητών. Οι ταξινομητές αυτοί μπορούν να λειτουργήσουν μεμονωμένα. Αυτό σημαίνει πως μπορούν να εκπαιδευτούν με ένα σύνολο δεδομένων (training set) και στη συνέχεια μπορούν να δεχθούν νέα δεδομένα και να τα αντιστοιχίσουν σε μία από τις κλάσεις.

Τα μεμονωμένα μοντέλα ταξινόμησης άρχισαν να αντικαθίστανται με συνδυασμούς ταξινομητών, οι οποίοι σε πολλές περιπτώσεις εμφανίζουν καλύτερη απόδοση [7]. Στο Σχήμα 6 παρουσιάζεται το σχηματικό διάγραμμα ενός μοντέλου το οποίο βασίζεται στον συνδυασμό ταξινομητών. Όπως φαίνεται, το δείγμα το οποίο θέλουμε να ταξινομήσουμε, αρχικά μπαίνει ως είσοδος στον κάθε classifier ξεχωριστά, ο οποίος αποφασίζει σε ποια κλάση θα το αντιστοιχίσει. Στη συνέχεια, οι αποφάσεις των μεμονωμένων ταξινομητών συνδυάζονται σύμφωνα με την μέθοδο που εφαρμόζεται και έτσι έχουμε την τελική απόφαση ταξινόμησης του συνδυασμού.



Σχήμα 6: Σχηματικό διάγραμμα συνδυασμού ταξινομητών

Από τη μία μεριά, είναι προφανές ότι ο συνδυασμός ίδιων ταξινομητών δεν θα προσφέρει τίποτα ως προς την απόδοση του συνολικού μοντέλου αλλά επιπλέον θα αυξήσει την πολυπλοκότητα του συστήματος. Από την άλλη, διαφορετικοί αλλά με πολύ χειρότερη απόδοση ταξινομητές, είναι απίθανο να επιφέρουν καλή απόδοση ως συνδυασμός. Έτσι, ούτε η μεμονωμένη απόδοση αλλά ούτε η διαφορετικότητα από μόνες τους παρέχουν ένα αξιόπιστο διαγνωστικό εργαλείο, το οποίο είναι σε θέση να εντοπίσει πότε ο συνδυασμός των classifiers υπερτερεί του καλύτερου μεμονωμένου classifier. Αντίθετα, επικρατεί η αντίληψη ότι οι βέλτιστοι συνδυασμοί ταξινομητών θα πρέπει να έχουν ταυτόχρονα καλές μεμονωμένες αποδόσεις και επαρκή διαφορετικότητα. Όλες αυτές οι διαπιστώσεις οδηγούν στο συμπέρασμα της χρήσης της απόδοσης του συνδυασμού ως ένα κατάλληλο κριτήριο επιλογής για τον συγκεκριμένο συνδυασμό.

Σίγουρα, η πιο αξιόπιστη τακτική είναι ο υπολογισμός όσο το δυνατόν περισσότερων, αν όχι όλων, διαφορετικών συνδυασμών και στη συνέχεια η επιλογή του συνδυασμού με την καλύτερη απόδοση. Μία δυσκολία στην συγκεκριμένη τακτική είναι το συνήθως μεγάλο φάσμα υπολογισμών. Για να γίνει αυτό αντιληπτό, αρκεί να σημειωθεί ότι υποθέτοντας έναν συγκεκριμένο αριθμό ταξινομητών, ο υπολογισμός όλων των συνδυασμών είναι μία διαδικασία που μεγαλώνει εκθετικά με τον αριθμό των ταξινομητών. Έτσι, η τακτική αυτή αυξάνει σημαντικά το υπολογιστικό κόστος για μεγάλο σύνολο ταξινομητών, οπότε και η διαφορετικότητα (diversity) είναι χρήσιμη για τέτοιες περιπτώσεις. Σε περιπτώσεις, επομένως, όπου ο αριθμός των διαθέσιμων ταξινομητών για την εφαρμογή του συνδυασμού τους είναι μεγάλος, προτείνεται ο υπολογισμός της διαφορετικότητας μεταξύ όλων των πιθανών συνδυασμών και στη συνέχεια η εφαρμογή των fusion μεθόδων στους συνδυασμούς με τη μεγαλύτερη διαφορετικότητα [8]. Στην περίπτωση που είτε ο αριθμός των διαθέσιμων ταξινομητών είναι μικρός είτε αυτό που ενδιαφέρει είναι η μέγιστη απόδοση ανεξαρτήτως υπολογιστικού κόστους, η τακτική που θα πρέπει να εφαρμόζεται είναι η εφαρμογή των fusion μεθόδων σε όλους τους πιθανούς συνδυασμούς των διαθέσιμων ταξινομητών και στη συνέχεια η επιλογή του συνδυασμού που επιφέρει την μεγαλύτερη απόδοση.

Υπάρχουν διάφορα κριτήρια υπολογισμού της διαφορετικότητας μεταξύ των ταξινομητών, τα οποία θα αναπτυχθούν στην επόμενη ενότητα. Πριν την ανάλυση

αυτών των κριτηρίων θα πρέπει να κάνουμε κάποιες υποθέσεις. Έστω λοιπόν ότι $D = \{D_1, D_2, ..., D_L\}$ είναι το σύνολο από L ταξινομητές, $\Omega = \{\omega_1, ..., \omega_c\}$ είναι το σύνολο από c κλάσεις και έστω ότι τα δείγματα περιγράφονται από n χαρακτηριστικά. Έτσι κάθε ταξινομητής δέχεται σαν είσοδο ένα διάνυσμα $\mathbf{x} \in \Re^n$, το οποίο το αντιστοιχεί σε μία κλάση από το σύνολο Ω ($\mathbf{x} \in \Re^n \to \Omega$). Συνήθως ο ταξινομητής δεν αντιστοιχεί απλά το δείγμα σε μία κλάση, αλλά δίνει για κάθε κλάση την πιθανότητα να ανήκει το δείγμα σε αυτήν. Έτσι η έξοδος του ταξινομητή για το δείγμα x αναπαρίσταται σε ένα c-διάστατο διάνυσμα της μορφής:

$$D_i(x) = \begin{bmatrix} d_{i,1}(x) \\ ... \\ d_{i,c}(x) \end{bmatrix},$$

όπου $i = 1, ..., L$ και $d_{i,j}(x)$ να είναι η πιθανότητα το δείγμα x να ανήκει στην κλάση j, όπου $j = 1, ..., c$. Τα $d_{i,j}(x)$ κανονικοποιούνται στο διάστημα [0,1]. Όταν τα $d_{i,j}(x)$ παίρνουν οποιαδήποτε τιμή μέσα στο διάστημα [0,1] τότε θεωρούνται ως soft labels. Κάποιες φορές όμως, χρειάζεται να είναι γνωστό μόνο εάν ένα δείγμα ανήκει ή δεν ανήκει σε μία κλάση. Σε αυτήν την περίπτωση τα $d_{i,j}(x)$ παίρνουν τιμές διακριτές και συγκεκριμένα όταν αναφερόμαστε στο διάστημα [0,1] παίρνουν τις τιμές 0 ή 1 και θεωρούνται ως crisp labels. Γενικότερα, πάντως, τα soft labels δίνουν περισσότερη πληροφορία και επιπλέον μπορούν να μετατραπούν σε crisp labels με την εφαρμογή ενός κατωφλίου. Για παράδειγμα, αν έχουμε δύο κλάσεις, τις ω_1 και ω_2, με $d_{i,1}(x) = 0.3$ και $d_{i,2}(x) = 0.7$, με εφαρμογή του κατωφλίου t=0.5 το δείγμα x αντιστοιχίζεται στην κλάση ω_2 και τα αντίστοιχα crisp labels είναι $d_{i,1}(x) = 0$ και $d_{i,2}(x) = 1$.

3.2 Υπολογισμός της διαφορετικότητας των ταξινομητών (diversity)

Όπως αναφέρθηκε και προηγουμένως, ένα σημαντικό στοιχείο για να επιτευχθεί η βέλτιστη απόδοση, όταν πρόκειται να συνδυαστούν οι μεμονωμένοι

ταξινομητές, είναι η διαφορετικότητα μεταξύ αυτών των ταξινομητών (diverse classifiers). Υπάρχουν διάφοροι τρόποι για να υπολογίσει κανείς την διαφορετικότητα, οι οποίοι διακρίνονται σε δύο κύριες κατηγορίες: αυτούς που εφαρμόζονται σε ζεύγη ταξινομητών και στην βιβλιογραφία παρουσιάζονται ως pairwise diversity measures και αυτούς που εφαρμόζονται σε όλο το σύνολο των ταξινομητών και αναφέρονται ως non-pairwise diversity measures ([9], [10]). Αν επιλεχθούν να χρησιμοποιηθούν τα pairwise κριτήρια, τότε για την εύρεση της διαφορετικότητας μεταξύ περισσότερων των δύο ταξινομητών, υπολογίζεται με τα κριτήρια αυτά η διαφορετικότητα για όλα τα πιθανά ζεύγη των ταξινομητών και στη συνέχεια χρησιμοποιείται ο μέσος όρος αυτών. Η τακτική αυτή έχει ακολουθηθεί και στην παρούσα διπλωματική εργασία.

3.2.1 Pairwise Diversity Measures

Έστω ότι έχουμε δύο ταξινομητές, τον D_i και τον D_k. Ως προς την ταξινόμηση ενός δείγματος οι D_i και D_k μπορούν να βρεθούν σε τέσσερις καταστάσεις:

- και οι δύο έχουν ταξινομήσει σωστά το δείγμα
- ο D_i έχει ταξινομήσει σωστά ενώ ο D_k λάθος το δείγμα
- ο D_i έχει ταξινομήσει λάθος ενώ ο D_k σωστά το δείγμα
- και οι δύο έχουν ταξινομήσει λάθος το δείγμα.

Οι καταστάσεις αυτές μαζί με τις αντίστοιχες πιθανότητες φαίνονται στον Πίνακα 3. Προφανώς, ισχύει $a+b+c+d=1$. Τα a, b, c, d χρησιμοποιούνται για τον ορισμό των κριτηρίων που θα μελετηθούν στην συνέχεια. Για ένα σύνολο με L ταξινομητές θα δημιουργηθούν $\frac{L\cdot(L-1)}{2}$ τιμές για κάθε pairwise κριτήριο διαφορετικότητας. Η τελική τιμή του κριτηρίου προκύπτει από τον μέσο όρο των παραπάνω τιμών.

	D_k correct (1)	D_k wrong (0)
D_i correct (1)	a	b
D_i wrong (0)	c	d

Πίνακας 3: Κατανομή ταξινομήσεων μεταξύ δύο ταξινομητών

3.2.1.1 Q-statistic (Q)

Το Q-statistic ορίζεται για δύο ταξινομητές D_i και D_k ως:

$$Q_{i,k} = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}$$

Για στατιστικά ανεξάρτητους ταξινομητές ισχύει ότι $Q_{i,k} = 0$. Η τιμή του Q παίρνει τιμές ανάμεσα στο -1 και στο 1. Ταξινομητές που έχουν την τάση να αντιστοιχίζουν σωστά τα ίδια δείγματα θα έχουν θετικές τιμές του Q, ενώ αυτοί που κάνουν λάθος ταξινομήσεις, αλλά σε διαφορετικά δείγματα, θα έχουν αρνητικό Q. Όσο μεγαλύτερη είναι η τιμή του Q τόσο μικρότερη είναι η διαφορετικότητα (diversity) ανάμεσα στους ταξινομητές.

3.2.1.2 Correlation coefficient (ρ)

Ο συντελεστής συσχέτισης ορίζεται για δύο ταξινομητές D_i και D_k ως:

$$\rho_{i,k} = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)}}$$

Όταν οι ταξινομητές είναι ανεξάρτητοι και ασυσχέτιστοι ισχύει ότι $\rho_{i,k} = 0$. Όσο μεγαλύτερη είναι η τιμή του συντελεστή συσχέτισης, τόσο μικρότερη είναι η διαφορετικότητα ανάμεσα στους ταξινομητές. Για οποιουσδήποτε δύο ταξινομητές, το

Q και το ρ έχουν το ίδιο πρόσημο και ακόμη μπορεί να αποδειχθεί ότι ισχύει $|\rho| \leq |Q|$.

3.2.1.3 Disagreement measure (D)

Το κριτήριο διαφωνίας (disagreement measure) ορίζεται για δύο ταξινομητές D_i και D_k ως:

$$D_{i,k} = b + c$$

Το κριτήριο διαφωνίας, εκφράζει την πιθανότητα ο ένας ταξινομητής να αντιστοίχισε σωστά το δείγμα και ο άλλος λανθασμένα, δηλαδή, την πιθανότητα οι δύο ταξινομητές να διαφωνούν στην απόφασή τους. Όσο μεγαλύτερη είναι η τιμή του κριτηρίου διαφωνίας (D), τόσο μικρότερη είναι η συσχέτιση μεταξύ των δύο ταξινομητών, δηλαδή, τόσο μεγαλύτερη είναι η διαφορετικότητα ανάμεσα τους.

3.2.1.4 Double-fault measure (DF)

Το κριτήριο διπλού σφάλματος (double fault measure) ορίζεται για δύο ταξινομητές D_i και D_k ως:

$$DF_{i,k} = d$$

Το κριτήριο διπλού σφάλματος, εκφράζει την πιθανότητα και οι δύο ταξινομητές να έχουν αντιστοιχίσει λανθασμένα το δείγμα (ταυτόχρονα). Όσο μεγαλύτερη είναι η τιμή του διπλού σφάλματος (DF) για τους δύο ταξινομητές, τόσο μεγαλύτερη είναι και η μεταξύ τους συσχέτιση, δηλαδή τόσο μικρότερη είναι η διαφορετικότητα ανάμεσα τους. Το κριτήριο αυτό είναι αρκετά σημαντικό, γιατί γενικότερα είναι προτιμητέο να γνωρίζουμε την πιθανότητα του ταυτόχρονου

σφάλματος και από τους δύο ταξινομητές, παρά την πιθανότητα της ταυτόχρονης σωστής ταξινόμησης.

3.2.1.5 Kappa Statistic

Το Kappa Statistic ορίζεται για δύο ταξινομητές D_i και D_k ως:

$$k_{i,k} = \frac{2 \cdot (a \cdot d - b \cdot c)}{(a+b) \cdot (b+d) + (a+c) \cdot (c+d)}$$

Το κριτήριο αυτό εκφράζει το βαθμό που συμφωνούν οι δύο ταξινομητές. Όταν η τιμή του k είναι μικρότερη του 0.40, τότε υπάρχει μικρή συμφωνία μεταξύ των δύο ταξινομητών, η οποία δεν βασίζεται στην τύχη, όταν είναι ανάμεσα στο 0.40 και στο 0.75 υπάρχει μέτρια συμφωνία, ενώ για τιμές του k μεγαλύτερες του 0.75 η συμφωνία των δύο ταξινομητών είναι πολύ μεγάλη.

3.2.2 Non-pairwise Diversity Measures

Στην προηγούμενη ενότητα, μελετήθηκαν κριτήρια με τα οποία υπολογίζεται η διαφορετικότητα μεταξύ ζευγών ταξινομητών. Στην ενότητα αυτή, θα μελετηθούν κριτήρια διαφορετικότητας, τα οποία εφαρμόζονται στο σύνολο των ταξινομητών. Όπως και προηγουμένως, θεωρείται ότι L είναι ο αριθμός των ταξινομητών. Επιπλέον, με N συμβολίζεται ο αριθμός των αντικειμένων του dataset και με $y_{j,i}$ η ορθότητα της απόφασης του ταξινομητή i για το αντικείμενο j (αν ο ταξινομητής αντιστοίχισε σωστά το αντικείμενο $y_{j,i} = 1$, ενώ αν το αντιστοίχισε λάθος $y_{j,i} = 0$).

3.2.2.1 Entropy measure

Το κριτήριο εντροπίας για το σύνολο των L ταξινομητών ορίζεται ως:

$$E = \frac{1}{N} \cdot \sum_{j=1}^{N} \frac{1}{\left(L - \frac{L}{2} - 1 \right)} \min \left\{ \sum_{i=1}^{L} y_{j,i}, L - \sum_{i=1}^{L} y_{j,i} \right\}$$

Η τιμή της εντροπίας παίρνει τιμές από 0 έως 1, όπου το 0 δηλώνει μεγάλη συσχέτιση μεταξύ των ταξινομητών, ενώ το 1 δηλώνει ότι η διαφορετικότητα ανάμεσά τους είναι μεγάλη. Το κριτήριο της εντροπίας παίρνει τη μεγαλύτερη τιμή (E=1) για ένα συγκεκριμένο αντικείμενο x_j, όταν οι μισοί (L/2) ταξινομητές το αντιστοιχίσουν στη σωστή κλάση ($y_{j,i} = 1$) και οι υπόλοιποι μισοί το ταξινομήσουν λάθος ($y_{j,i} = 0$). Επιπλέον, το κριτήριο εντροπίας παίρνει τη μικρότερη τιμή όταν όλοι οι ταξινομητές παίρνουν την ίδια απόφαση για όλα τα αντικείμενα, γεγονός που υποδηλώνει ότι δεν είναι διαφορετικοί.

3.2.2.2 Kohavi-Wolpert variance

Η διακύμανση Kohavi-Wolpert για το σύνολο των L ταξινομητών ορίζεται ως:

$$kw = \frac{1}{NL^2} \cdot \sum_{j=1}^{N} l(z_j) \cdot \left(L - l(z_j) \right),$$

όπου με $l(z_j)$ συμβολίζεται ο αριθμός των ταξινομητών που αντιστοίχισαν στη σωστή κλάση το z_j.

Στον παρακάτω πίνακα (Πίνακας 4), συνοψίζονται τα κριτήρια διαφορετικότητας που μελετήθηκαν παραπάνω, ο τύπος τους, καθώς και το πώς επηρεάζουν την διαφορετικότητα. Το βέλος δείχνει κατά πόσο η διαφορετικότητα μεγαλώνει όταν η τιμή του κριτηρίου μειώνεται (↓) ή αυξάνεται (↑).

Όνομα	Σύμβολο	↑/↓	Pairwise?
Q-statistic	Q	↓	√
Correlation Coefficient	ρ	↓	√
Disagreement measure	D	↑	√
Double-fault measure	DF	↓	√
Kappa statistic	k	↓	√
Entropy measure	E	↑	×
Kohavi-Wolpert variance	kw	↑	×

Πίνακας 4: Συνοπτικός πίνακας των Diversity Measures.

3.3 Μέθοδοι συνδυασμού των ταξινομητών

Όπως αναφέρθηκε και στα προηγούμενα, υπάρχουν διάφορες μέθοδοι που μπορεί να ακολουθήσει κανείς για να συνδυάσει τα αποτελέσματα των ταξινομητών, με σκοπό την βελτιστοποίηση της απόδοσης του μοντέλου ταξινόμησης. Δεν είναι πάντοτε ξεκάθαρο ποια μέθοδος θα επιφέρει την βέλτιστη απόδοση, γι'αυτό συνήθως συνδυάζουμε τους ταξινομητές με αρκετές μεθόδους και κρατάμε την βέλτιστη από πλευράς απόδοσης.

Υπάρχουν δύο βασικές κατηγορίες στις οποίες χωρίζονται οι μέθοδοι συνδυασμού των ταξινομητών, ως προς το είδος της εισόδου που δέχονται. Στην πρώτη κατηγορία ανήκουν οι μέθοδοι που δέχονται crisp labels ως εισόδους, ενώ στην δεύτερη κατηγορία οι μέθοδοι απαιτούν soft labels [10].

3.3.1 Crisp Labeling μέθοδοι συνδυασμού

Οι μέθοδοι αυτές, αφού έχουν εκπαιδευτεί οι μεμονωμένοι ταξινομητές, χρησιμοποιούν τα crisp labels των δειγμάτων που έχουν δημιουργηθεί και εξάγουν την τελική απόφαση για κάθε δείγμα, η οποία είναι και αυτή σε crisp labeling.

3.3.1.1 Majority Vote

Εφαρμόζοντας τη μέθοδο του Majority Vote το δείγμα x αντιστοιχίζεται στην κλάση, στην οποία το αντιστοίχισε η πλειοψηφία των μεμονωμένων ταξινομητών [7]. Η μέθοδος αυτή για περιττό αριθμό ταξινομητών δεν παρουσιάζει κανένα πρόβλημα. Στην περίπτωση άρτιου αριθμού ταξινομητών και όταν υπάρχει ισοψηφία στην απόφασή τους, το δείγμα x αντιστοιχίζεται τυχαία σε μία κλάση.

3.3.2 Soft labeling μέθοδοι συνδυασμού

Οι μέθοδοι που ακολουθούν, χρησιμοποιούν ως εισόδους τα soft labels των μεμονωμένων ταξινομητών. Υποθέτουμε όπως και πριν ότι $D = \{D_1, \ldots, D_L\}$ είναι το σύνολο των L ταξινομητών και $\Omega = \{\omega_1, \ldots, \omega_c\}$ το σύνολο με τις c κλάσεις. Επίσης με $\mu_j(x)$ συμβολίζεται η πιθανότητα αντιστοίχησης του αντικειμένου x στην κλάση j από τον συνδυασμό των ταξινομητών.

3.3.2.1 Minimum, Maximum, Average, Product rule

Οι απλές αυτές μέθοδοι συνδυασμού των ταξινομητών, υπολογίζουν την πιθανότητα αντιστοίχησης για την κλάση ω_j, $\mu_j(x)$, χρησιμοποιώντας γενικά τον κανόνα :

$$\mu_j(\mathbf{x}) = F\left[d_{1,j}(\mathbf{x}), \ldots, d_{L,j}(\mathbf{x})\right], \; j = 1,...,c$$

και όπου F είναι η συνάρτηση συνδυασμού (F={minimum, maximum, average, product}). Έτσι, τα Minimum, Maximum, Average, Product rule δίνονται αντίστοιχα από τους τύπους:

- Minimum Rule: $\mu_j(x) = \min\left(d_{1,j}(x),...,d_{L,j}(x)\right)$
- Maximum Rule: $\mu_j(x) = \max\left(d_{1,j}(x),...,d_{L,j}(x)\right)$
- Average Rule: $\mu_j(x) = \frac{1}{L} \cdot \sum_{i=1}^{L} d_{i,j}(x)$
- Product Rule: $\mu_j(x) = \prod_{i=1}^{L} d_{i,j}(x)$

Για κάθε κλάση j υπολογίζεται το $\mu_j(x)$ και στη συνέχεια, το δείγμα x αντιστοιχίζεται στην κλάση που έχει τη μεγαλύτερη τιμή $\mu_j(x)$.

3.3.2.2 Decision Templates

Όπως αναφέρθηκε και προηγουμένως, έστω ότι $D = \{D_1, \ldots, D_L\}$ το σύνολο με τους L ταξινομητές, $\Omega = \{\omega_1, \ldots, \omega_c\}$ το σύνολο με τις c κλάσεις και η έξοδος του κάθε ταξινομητή για το δείγμα x να περιγράφεται από το c-διάστατο διάνυσμα:

$$D_i(x) = \begin{bmatrix} d_{i,1}(x) \\ ... \\ d_{i,c}(x) \end{bmatrix},$$

όπου $i = 1,...,L$ και $d_{i,j}(x)$ η πιθανότητα το δείγμα x να ανήκει στην κλάση j, όπου $j = 1,...,c$. Οργανώνοντας τις εξόδους από όλους τους ταξινομητές για το δείγμα x, σε έναν πίνακα, προκύπτει το Decision Profile του δείγματος [11], το οποίο είναι της μορφής:

$$DP(x) = \begin{bmatrix} d_{1,1}(x) & \ldots & d_{1,j}(x) & \ldots & d_{1,c}(x) \\ \vdots & & \vdots & & \vdots \\ d_{i,1}(x) & \ldots & d_{i,j}(x) & \ldots & d_{i,c}(x) \\ \vdots & & \vdots & & \vdots \\ d_{L,1}(x) & \ldots & d_{L,j}(x) & \ldots & d_{L,c}(x) \end{bmatrix}$$

Είναι φανερό ότι ο πίνακας αυτός έχει διάσταση $L \times c$. Οι στήλες του πίνακα αναπαριστούν για κάθε κλάση, την πιθανότητα που δίνει ο κάθε ταξινομητής, το δείγμα x να ανήκει σε αυτήν, ενώ οι γραμμές του πίνακα αφορούν την έξοδο του κάθε ταξινομητή. Είναι προφανές πως το άθροισμα κάθε γραμμής είναι ίσο με 1.

Σε αυτήν την μέθοδο, δημιουργείται ένα Decision Template (DT) για κάθε κλάση. Το DT κατασκευάζεται από τη μέση τιμή των DPs όλων των δειγμάτων που ανήκουν στην κλάση, σύμφωνα με την εξίσωση:

$$DT_j = \frac{1}{N_j} \cdot \sum_{\substack{z_i \in \omega_j \\ z_i \in Z}} DP(z_i) \ , \quad j = 1,...,c$$

Το DT είναι όπως και το DP ένας πίνακας $L \times c$ και μπορεί να θεωρηθεί πως είναι το decision profile της κάθε κλάσης. Αφού δημιουργηθούν τα decision templates, στη συνέχεια συγκρίνεται το $DP(x)$ κάθε δείγματος με τα DT της κάθε κλάσης με τη χρήση κάποιου κριτηρίου ομοιότητας S:

$$\mu_j(x) = S\left(DP(x), DT_j\right) \ , \ j = 1,...,c$$

Το DT της κλάσης που ταιριάζει καλύτερα με το $DP(x)$, σύμφωνα με το κριτήριο S, καθορίζει και την κλάση που θα αντιστοιχιστεί το δείγμα. Δύο από τα κριτήρια ομοιότητας που χρησιμοποιούνται είναι η ευκλείδεια απόσταση και η συμμετρική διαφορά:

- **Ευκλείδεια απόσταση**

Η πιθανότητα, που προκύπτει από τον συνδυασμό των ταξινομητών, να ανήκει το δείγμα x στην κλάση j είναι :

$$\mu_j(x) = 1 - \frac{1}{L \cdot c} \cdot \sum_{i=1}^{L} \sum_{k=1}^{c} \left(DT_j(i,k) - d_{i,k}(x)\right)^2 \ ,$$

όπου το $DT_j(i,k)$ είναι το (i,k)-στοιχείο του DT_j.

- **Συμμετρική απόσταση**

Η πιθανότητα, που προκύπτει από τον συνδυασμό των ταξινομητών, να ανήκει το δείγμα x στην κλάση j είναι :

$$\mu_j(x) = 1 - \frac{1}{L \cdot c} \cdot \sum_{i=1}^{L} \sum_{k=1}^{c} \max\left\{\min\left\{DT_j(i,k), \left(1 - d_{i,k}(x)\right)\right\}, \min\left\{\left(1 - DT_j(i,k)\right), d_{i,k}(x)\right\}\right\}$$

ΚΕΦΑΛΑΙΟ 4

ΥΛΟΠΟΙΗΣΗ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 Υλοποίηση και Πειραματική διαδικασία

4.1.1 Αντικείμενο της εργασίας

Στην εργασία αυτή, μελετήθηκαν και υλοποιήθηκαν διάφορες μέθοδοι συνδυασμού των ταξινομητών, οι οποίες στην συνέχεια εφαρμόστηκαν σε τέσσερα datasets, προκειμένου να διεξαχθούν συμπεράσματα, σχετικά με το κατά πόσο η απόδοση των συνδυαστικών μοντέλων παρουσίασε αύξηση, σε σχέση με την απόδοση των ταξινομητών όταν εφαρμόστηκαν μεμονωμένα. Τα δύο datasets αφορούν ασθενείς με οξεία μυελοειδή λευχαιμία (Acute Myeloid Leukemia, AML), ενώ τα άλλα δύο αφορούν ασθενείς με καρκίνο του μαστού (Breast Cancer Recursion και Breast Cancer Diagnosis dataset). Το AML dataset, αφορά ασθενείς που πάσχουν από οξεία μυελοειδή λευχαιμία και οι οποίοι έχουν υποβληθεί σε θεραπεία [13]. Έτσι, ελέγχοντας την πρόοδο τους την $30^{η}$ και την $90^{η}$ μέρα από την έναρξη της θεραπείας το αρχικό dataset εξετάζεται με δύο διαφορετικούς τρόπους (αντίστοιχα, AML Short Term Analysis και AML Long Term Analysis). Το Breast Cancer Recursion dataset, αφορά ασθενείς που είχαν εμφανίσει στο παρελθόν όγκο στο στήθος και είχαν υποβληθεί σε θεραπεία και οι οποίοι τώρα επανεξετάζονται για την επανεμφάνιση ή μη του όγκου. Το Breast Cancer Diagnosis dataset, αφορά ασθενείς με καρκίνο στο στήθος και στους οποίους γίνεται διάγνωση για το αν πρόκειται για καλοήθη ή κακοήθη όγκο [14]. Περισσότερες πληροφορίες σχετικά με τις κλάσεις του κάθε dataset, την αναλογία τους, όπως και τα χαρακτηριστικά που περιγράφουν το κάθε δείγμα περιέχονται στο Παράρτημα.

4.1.2 Προετοιμασία των δεδομένων και εκπαίδευση των ταξινομητών

Όπως αναφέρθηκε και στο κεφάλαιο 1, πριν από το στάδιο του συνδυασμού των ταξινομητών (level 2), προηγείται το στάδιο (level 1) που περιλαμβάνει τον διαχωρισμό του dataset σε training set και σε test set, την επιλογή των χαρακτηριστικών με τα οποία θα περιγράφονται τα δείγματα, την επιλογή των ταξινομητών που θα χρησιμοποιηθούν καθώς και την εκπαίδευσή τους. Το στάδιο αυτό, υλοποιήθηκε από τον μεταπτυχιακό φοιτητή Γιώργο Μανίκη και στην συνέχεια θα γίνει μία πολύ σύντομη αναφορά στη διαδικασία που ακολουθήθηκε. Στο σημείο αυτό, πρέπει να αναφερθεί, πως η υλοποίηση έγινε στο MATLAB, με τη χρήση δύο επιπλέον toolboxes, των PRTOOLS 4.0 και LS-SVMlab 1.5 [12].

Καταρχήν, όπως αναφέρθηκε και στο κεφάλαιο 1, ένας σημαντικός παράγοντας για την επίτευξη καλής απόδοσης στην ταξινόμηση, είναι η χρήση όσο το δυνατόν περισσότερων δεδομένων του dataset για την εκπαίδευση των ταξινομητών και επίσης όσο το δυνατόν περισσότερων δεδομένων για τον έλεγχο της απόδοσής τους. Όμως, η χρήση όλων των δεδομένων του dataset τόσο για την εκπαίδευση όσο και για τον έλεγχο του ταξινομητή δεν συνίσταται, γιατί υπάρχει το ενδεχόμενο του overtrain, δηλαδή του να μάθει ο ταξινομητής να κατηγοριοποιεί σωστά όλα τα δείγματα του dataset και στη συνέχεια να αποτυγχάνει στα νέα δεδομένα. Έτσι, είναι απαραίτητο, να υπάρχουν δύο ξεχωριστά σύνολα για την εκπαίδευση και τον έλεγχο της απόδοσης του ταξινομητή. Για το λόγο αυτό, το αρχικό dataset χωρίστηκε με αναλογία 80% - 20% σε ένα training και σε ένα test set, αντίστοιχα, φροντίζοντας να διατηρούνται οι αναλογίες των κλάσεων του αρχικού dataset. Για να είναι τα αποτελέσματα που θα προκύψουν, όσο το δυνατό πιο δίκαια και αντικειμενικά, ο διαχωρισμός του dataset σε training και test set επαναλήφθηκε 20 φορές. Επιπλέον, το training set χρησιμοποιήθηκε για την εκπαίδευση των ταξινομητών και την επιλογή των κατάλληλων παραμέτρων για κάθε ταξινομητή. Για τον σκοπό αυτό, χρησιμοποιήθηκε η 5-folds cross-validation μέθοδος. Το test set δεν συμμετείχε καθόλου στη διαδικασία της εκπαίδευσης των ταξινομητών και χρησιμοποιήθηκε αποκλειστικά στην τελική αξιολόγηση της απόδοσής τους.

Οι ταξινομητές που εκπαιδεύτηκαν και χρησιμοποιήθηκαν στο level 1, ήταν τα Support Vector Machines (SVMs), τα Least Square Support Vector Machines (LS-SVMs) και τα Hidden Space Support Vector Machines (HS-SVMs). Εκτός από αυτές τις παραλλαγές των SVMs, χρησιμοποιήθηκαν επίσης και ταξινομητές που περιέχονται στο PRTOOLS toolbox, όπως οι: Fisher's Linear Classifier (Fisher), Linear Discriminant Classifier (LDC), Naïve Bayes Classifier (NBC), Quadratic Discriminant Classifier (QDC), K- Nearest Neighbor Classifier (KNNC), Radial Basis Function Neural Network Classifier (RBNC), Random Neural Network Classifier (RNNC).

Τέλος, στα Least Square Support Vector Machines έγινε χρήση του Radial Basis Function (RBF) kernel, στα Support Vector Machines χρησιμοποιήθηκε ο Polynomial kernel και στα Hidden Space Support Vector Machines χρησιμοποιήθηκαν σε σειρά ο Radial Basis Function, ο Polynomial και ένας Linear kernel.

4.1.3 Συνδυασμός των ταξινομητών (Classifier Fusion)

Όπως αναφέρθηκε και στο κεφάλαιο 3, δοθέντος ενός συνόλου ταξινομητών, υπάρχουν διάφορες μέθοδοι συνδυασμού (fusion) των αποτελεσμάτων τους στο level 2. Υπάρχουν δύο εναλλακτικοί τρόποι που ακολουθούνται πριν την εφαρμογή των fusion μεθόδων, οι οποίοι αφορούν το πλήθος των ταξινομητών που είναι διαθέσιμοι για το level 2. Ο πρώτος τρόπος εφαρμόζεται όταν το σύνολο περιέχει πολλούς ταξινομητές και είναι ιδιαίτερα χρονοβόρο να υπολογιστεί κάθε πιθανός συνδυασμός τους με όλες τις μεθόδους. Επιπλέον, το να συνδυαστούν ταξινομητές οι οποίοι μετά την εκπαίδευσή τους επιφέρουν όμοια αποτελέσματα ταξινόμησης στα νέα δεδομένα, δεν μπορεί να προκαλέσει αύξηση της απόδοσης του συνδυαστικού μοντέλου παρά μόνο επιπλέον υπολογιστικό κόστος στην εφαρμογή. Έτσι, στην περίπτωση που το σύνολο αποτελείται από μεγάλο αριθμό ταξινομητών, αρχικά με κάποια diversity measures υπολογίζεται η διαφορετικότητα (diversity) μεταξύ των πιθανών συνδυασμών των ταξινομητών και στη συνέχεια στον συνδυασμό που έχει τη μεγαλύτερη διαφορετικότητα, εφαρμόζονται οι διάφορες fusion μέθοδοι. Ο δεύτερος

εναλλακτικός τρόπος είναι η εύρεση όλων των πιθανών συνδυασμών ταξινομητών και η εφαρμογή των fusion μεθόδων σε όλους τους συνδυασμούς.

Στην παρούσα εργασία, υλοποιήθηκαν και οι δύο τρόποι που περιγράφηκαν παραπάνω (εξαντλητική αναζήτηση και επιλογή με βάση τα diversity measures). Πιο αναλυτικά, οι fusion μέθοδοι που μελετήθηκαν και στη συνέχεια υλοποιήθηκαν είναι : Majority vote, Minimum rule, Maximum rule, Average rule, Product rule και Decision Templates. Για την υλοποίηση αυτών των μεθόδων, τα αποτελέσματα του level 1 οργανώθηκαν σε Decision Templates, όπως έχει περιγραφεί στο Κεφάλαιο 3.

Τα μέτρα που χρησιμοποιήθηκαν για τον υπολογισμό της διαφορετικότητας (diversity measures) μεταξύ κάθε πιθανού συνδυασμού ταξινομητών είναι τα: Q statistic, correlation coefficient, disagreement measure, double fault measure και K statistic. Έτσι, αφού υπολογίστηκε για κάθε πιθανό συνδυασμό η τιμή του καθενός από τα παραπάνω μέτρα, στη συνέχεια οι τιμές κανονικοποιήθηκαν στο διάστημα [0,1] με το 0 να δηλώνει την απόλυτη ομοιότητα μεταξύ των ταξινομητών του συνδυασμού και το 1 να δηλώνει την μέγιστη διαφορετικότητα. Τέλος, για πιο ακριβή και αντικειμενικά αποτελέσματα υπολογίστηκε και χρησιμοποιήθηκε ο μέσος όρος αυτών των μετρήσεων. Πρέπει να τονιστεί, ότι η τιμή όλων των diversity measures για κάθε συνδυασμό προέκυψε από τον μέσο όρο των measures από κάθε διαχωρισμό του αρχικού dataset, που έγινε επαναληπτικά 20 φορές, όπως εξηγήθηκε στην προηγούμενη ενότητα.

4.2 Αποτελέσματα

Στις επόμενες ενότητες, παρουσιάζονται τα αποτελέσματα για το AML Long Term Analysis, AML Short Term Analysis, Breast Cancer Recursion και για το Breast Cancer Diagnosis datasets. Προηγουμένως, θα πρέπει να δοθούν ορισμένες διευκρινήσεις σχετικά με το μέτρο που χρησιμοποιήθηκε για την αξιολόγηση της απόδοσης των fusion μεθόδων ταξινόμησης καθώς και για τη σημασία των παραμέτρων που εμφανίζονται στους πίνακες.

Τα dataset που χρησιμοποιήθηκαν δεν είχαν ίση κατανομή των δειγμάτων ανά κλάση. Για την ακρίβεια, σε όλα τα dataset εκτός του Breast Cancer Diagnosis, η μία κλάση είχε πολύ περισσότερα δείγματα από την άλλη. Για τον λόγο αυτό, και προκειμένου να υπάρχει μία καλύτερη εικόνα σχετικά με την συμπεριφορά των fusion μεθόδων στα τέσσερα datasets, για κάθε μέθοδο υπολογίζεται η κατανομή των σωστών και λανθασμένων ταξινομήσεων ανά κλάση. Έτσι υπολογίζονται οι λόγοι TPR (ή sensitivity), FPR, TNR (ή specificity), FNR, όπως ορίστηκαν στο κεφάλαιο 1. Για να μπορεί να γίνει καλύτερα η σύγκριση σχετικά με το πώς λειτούργησαν οι fusion μέθοδοι μεταξύ των τεσσάρων datasets, ως Positives θεωρείται πάντα η κλάση με τα περισσότερα δείγματα και αντίστοιχα ως Negatives θεωρείται η κλάση με τα λιγότερα δείγματα (θα μπορούσε να θεωρηθεί και το αντίθετο). Τέλος, ως μέτρο αξιολόγησης της συνολικής απόδοσης των fusion μεθόδων χρησιμοποιήθηκε το εμβαδό της καμπύλης ROC (το οποίο εμφανίζεται στους πίνακες ως AUROC).

Όπως αναφέρθηκε και στην προήγουμενη ενότητα, για κάθε dataset οι fusion μέθοδοι εφαρμόστηκαν όχι μόνο στους συνδυασμούς που εμφάνισαν την μεγαλύτερη διαφορετικότητα αλλά σε όλο το σύνολο των πιθανών συνδυασμών των ταξινομητών. Επείδη με 9 διαθέσιμους ταξινομητές οι πιθανοί συνδυασμοί φτάνουν τους 502 είναι αδύνατο και ανούσιο να παρουσιαστούν τα αποτελέσματα από όλους αυτούς τους συνδυασμούς. Για τον λόγο αυτό, σε κάθε dataset τα αποτελέσματα παρουσιάζονται σε τρεις πίνακες. Στον πρώτο πίνακα, φαίνεται για κάθε fusion μέθοδο η καλύτερη απόδοση που έχει επιτευχθεί και από ποιον συνδυασμό. Στον δεύτερο πίνακα, φαίνονται για κάθε fusion μέθοδο οι αντίστοιχες αποδόσεις του συνδυασμού με το μεγαλύτερο diversity. Τέλος, στον τρίτο πίνακα, φαίνονται οι αποδόσεις του συνδυασμού με όλους τους ταξινομητές για κάθε fusion μέθοδο. Και στους τρεις πίνακες για κάθε συνδυασμό φαίνεται και το αντίστοιχο πεδίο που δηλώνει την διαφορετικότητα μεταξύ των ταξινομητών του συνδυασμού. Η διαφορετικότητα, όπως έχει αναφερθεί και προηγουμένως, εκφράζεται ως ο μέσος όρος των 5 diversity measures που χρησιμοποιήθηκαν. Επιπλέον, για κάθε dataset υπάρχει και μία γραφική παράσταση στην οποία φαίνονται 9 confidence intervals. Το πρώτο αντιστοιχεί στην περίπτωση της απόδοσης όταν οι ταξινομητές λειτουργούν μεμονωμένα. Το δεύτερο αντιστοιχεί στην περίπτωση της απόδοσης όταν οι ταξινομητές συνδυάζονται ανά 2, το τρίτο όταν οι ταξινομητές συνδυάζονται ανά 3 κ.ο.κ.. Με τον τρόπο αυτό, μπορεί να διαπιστωθεί η αύξηση ή όχι της απόδοσης του μοντέλου όταν ο αριθμός των

ταξινομητών που αποφασίζουν από κοινού αυξάνεται. Στις επόμενες ενότητες παρατίθενται τα αποτελέσματα για κάθε dataset.

4.2.1 Αποτελέσματα για το AML Long Term Analysis dataset

Στην ενότητα αυτή, παρουσίαζονται τα αποτελέσματα για το AML Long Term Analysis dataset. Στο dataset αυτό, οι ασθενείς πάσχουν από οξεία μυελοειδή λευχαιμία και αφού έχουν υποβληθεί σε θεραπεία, την 90η μέρα από την έναρξη της θεραπείας χωρίζονται σε δύο κατηγορίες ανάλογα με το εάν θεραπεύτηκαν πλήρως (κλάση 2) ή όχι (κλάση 1).

Classifier Fusion με βάση το καλύτερο AUROC	Μέθοδος Fusion	Ταξινομητές 1ου επιπέδου	Sensitivity	Specificity	FPR	FNR	AUROC	Diversity
	Majority	NBC-HSSVM-LSSVM	73,62%	68,33%	31,67%	26,38%	70,98%	0,329
	Minimum	RBNC-HSSVM-LSSVM	73,28%	67,00%	33,00%	26,72%	70,14%	0,617
	Maximum	RBNC-HSSVM-LSSVM	73,28%	67,00%	33,00%	26,72%	70,14%	0,617
	Average	LDC-HSSVM-LSSVM	72,93%	67,33%	32,67%	27,07%	70,13%	0,222
	Product	LDC-HSSVM-LSSVM	72,93%	67,33%	32,67%	27,07%	70,13%	0,222
	Decision Templates	KNNC-LDC-QDC-RNNC-Fisher-HSSVM-LSSVM	74,14%	66,67%	33,33%	25,86%	70,40%	0,391
		KNNC-LDC-QDC-RNNC-RBNC-Fisher-HSSVM-LSSVM	74,14%	66,67%	33,33%	25,86%	70,40%	0,450

Πίνακας 5: Αποτελέσματα της καλύτερης απόδοσης (με βάση το AUROC) όλων των fusion μεθόδων για το AML Long Term Analysis.

Στον παραπάνω πίνακα, παρουσιάζονται για το AML Long Term Analysis dataset, οι συνδυασμοί που πέτυχαν το μέγιστο AUROC για κάθε fusion μέθοδο. Από όλες τις μεθόδους, την καλύτερη απόδοση παρουσίασε το Majority Voting με $AUROC_{majority} = 70{,}98\%$, ενώ την χειρότερη οι Average και Product Rule με $AUROC_{Average,\Pr oduct} = 70{,}13\%$. Ο συνδυασμός NBC-HSSVM-LSSVM, που εμφανίζεται στο Majority Voting, έχει υψηλότερο diversity από τον συνδυασμό LDC-HSSVM-LSSVM που εμφανίζεται στους Average και Product Rule. Υπάρχουν όμως

και συνδυασμοί, όπως ο RBNC-HSSVM-LSSVM με μεγαλύτερο diversity αλλά που πετύχαν μικρότερη απόδοση από τον NBC-HSSVM-LSSVM. Επομένως, δεν παρατηρείται κάποια αντιστοιχία μεταξύ του diversity και της απόδοσης. Επιπλέον, όπως φαίνεται και στον παραπάνω πίνακα, οι LSSVM και HSSVM συμμετέχουν σε όλους τους συνδυασμούς. Το γεγονός αυτό είναι αναμενόμενο, καθώς και οι δύο αυτοί ταξινομητές, έχουν υψηλότερο ποσοστό Specificity στο level 1 σε σχέση με τους υπόλοιπους και έτσι συμβάλλουν σημαντικά στην αύξηση του AUROC του συνδυασμού στον οποίο συμμετέχουν.

Classifier Fusion με βάση το καλύτερο diversity	Μέθοδος Fusion	Ταξινομητές 1ου επιπέδου	Sensitivity	Specificity	FPR	FNR	AUROC	Diversity
	Majority	RBNC-LSSVM	83,45%	37,67%	62,33%	16,55%	60,56%	0,847
	Minimum	RBNC-LSSVM	75,35%	63,67%	36,33%	24,67%	69,51%	0,847
	Maximum	RBNC-LSSVM	75,35%	63,67%	36,33%	24,67%	69,51%	0,847
	Average	RBNC-LSSVM	75,35%	63,67%	36,33%	24,67%	69,51%	0,847
	Product	RBNC-LSSVM	75,35%	63,67%	36,33%	24,67%	69,51%	0,847
	Decision Templates	RBNC-LSSVM	69,31%	70,67%	29,33%	30,69%	69,99%	0,847

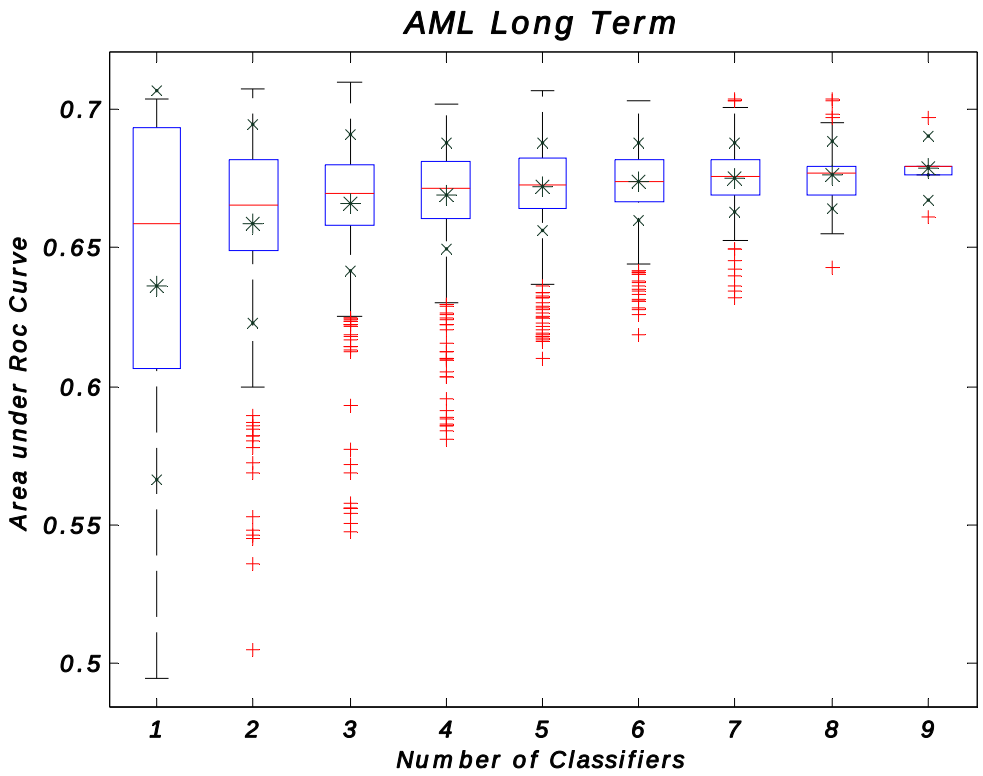
Πίνακας 6: Αποτελέσματα των fusion μεθόδων για τον συνδυασμό των ταξινομητών που εμφανίζει το μέγιστο diversity για το AML Long Term Analysis.

Στον παραπάνω πίνακα, παρουσιάζονται για το AML Long Term Analysis dataset, τα αποτελέσματα του συνδυασμού με το μεγαλύτερο diversity για όλες τις fusion μεθόδους. Τον συνδυασμό αποτελούν οι ταξινομητές RBNC και LSSVM και το diversity τους είναι 0,847. Από όλες τις μεθόδους, τα Decision Templates πέτυχαν την υψηλότερη απόδοση με $AUROC_{D.T.} = 69{,}99\%$ ενώ την χαμηλότερη απόδοση εμφάνισε το Majority Voting με $AUROC_{Majority} = 60{,}56\%$. Συγκρίνοντας τον συνδυασμό με την καλύτερη απόδοση του Πίνακα 6 (NBC-HSSVM-LSSVM) με αυτόν του Πίνακα 5 (RBNC-LSSVM), διαπιστώνεται πως ο συνδυασμός NBC-HSSVM-LSSVM με diversity κατά πολύ μικρότερο του RBNC-LSSVM, πετυχαίνει υψηλότερη απόδοση.

Classifier Fusion με όλους τους ταξινομητές	Μέθοδος Fusion	Ταξινομητές 1ου επιπέδου	Sensitivity	Specificity	FPR	FNR	AUROC	Diversity
	Majority	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	81,21%	51,00%	49,00%	18,79%	66,10%	0,448
	Minimum	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	78,79%	57,00%	43,00%	21,21%	67,90%	0,448
	Maximum	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	77,93%	57,33%	42,67%	22,07%	67,63%	0,448
	Average	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	78,62%	57,33%	42,67%	21,38%	67,98%	0,448
	Product	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	77,93%	58,00%	42,00%	22,07%	67,97%	0,448
	Decision Templates	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	73,45%	66,00%	34,00%	26,55%	69,72%	0,448

Πίνακας 7: Αποτελέσματα των fusion μεθόδων για τον συνδυασμό που περιέχει όλους τους ταξινομητές για το AML Long Term Analysis.

Στον παραπάνω πίνακα, παρουσιάζονται για το AML Long Term Analysis dataset, τα αποτελέσματα των fusion μεθόδων όταν στον συνδυασμό συμμετέχουν όλοι οι διαθέσιμοι ταξινομητές. Για άλλη μία φορά, τα Decision Templates εμφανίζουν υψηλότερη απόδοση σε σχέση με τις υπόλοιπες μεθόδους με $AUROC_{D.T.} = 69{,}72\%$ ενώ την χαμηλότερη απόδοση παρουσιάζει και πάλι το Majority Voting με $AUROC_{Majority} = 66{,}10\%$. Η υψηλή απόδοση των Decision Templates, παρά το γεγονός ότι εμφανίζει το χαμηλότερο Sensitivity, οφείλεται στο αυξημένο Specificity. Επιπλέον, θα πρέπει να αναφερθεί, πως ο συνδυασμός όλων των ταξινομητών, συγκρίνοντάς τον και με τον συνδυασμό RBNC-LSSVM που βασίζεται στο μέγιστο diversity, δεν πέτυχε κάποιο υψηλό ποσοστό AUROC.



Σχήμα 7: Confidence Intervals για το AUROC για κάθε περίπτωση αριθμού ταξινομητών που συμμετέχουν στον συνδυασμό για το AML Long Term Analysis.

Στο παραπάνω σχήμα, παρουσιάζονται για το AML Long Term Analysis dataset, τα Confidence Intervals (μπλε ορθογώνια) του AUROC για όλες τις περιπτώσεις, ανάλογα με τον αριθμό των ταξινομητών που συμμετέχουν στον συνδυασμό. Με κόκκινη γραμμή δηλώνεται η median τιμή του AUROC, με αστερίσκο ο μέσος όρος των τιμών ενώ με το σύμβολο × η τυπική απόκλιση από τον μέσο όρο. Όπως φαίνεται και από το σχήμα, όταν οι ταξινομητές λειτουργούν μεμονωμένα το AUROC παίρνει τιμές σε ένα αρκετά μεγαλό διάστημα τιμών που κυμαίνεται από 0,61% έως 0,69%. Στην περίπτωση που οι ταξινομητές συνδυάζονται ανά δύο, το AUROC κυμαίνεται από 0,65% έως 0,68%. Καθώς ο αριθμός των ταξινομητών αυξάνεται, το κάτω όριο του confidence interval αυξάνεται, με μόνη εξαίρεση την περίπτωση των οκτώ ταξινομητών που το κάτω όριο πέφτει ελάχιστα. Το άνω όριο του confidence interval δεν παρουσιάζει σταθερή πορεία, καθώς από 0,69% που είναι στους μεμονωμένους ταξινομητές στη συνέχεια πέφτει στο 0,68% στους συνδυασμούς ανά δύο, έπειτα αυξάνεται ελάχιστα και τελικά αρχίζει πάλι να μειώνεται. Η μέση τιμή των AUROCs αυξάνεται συνεχώς με την μέγιστη αύξηση να παρατηρείται στο

πέρασμα από τους μεμονωμένους ταξινομητές στους συνδυασμούς ανά δύο. Γενικά, όσο αυξάνεται ο αριθμός των ταξινομητών που συμμετέχουν στον συνδυασμό, τόσο ελαττώνεται το confidence interval του AUROC γεγονός που καθιστά το σύστημα ολοένα και πιο σταθερό.

4.2.2 Αποτελέσματα για το AML Short Term Analysis dataset

Στην ενότητα αυτή, παρουσίαζονται τα αποτελέσματα για το AML Short Term Analysis dataset. Στο dataset αυτό, όπως και στο AML Long Term Analysis dataset, οι ασθενείς πάσχουν από οξεία μυελοειδή λευχαιμία και αφού έχουν υποβληθεί σε θεραπεία, την 30^{η} μέρα από την έναρξη της θεραπείας χωρίζονται σε δύο κατηγορίες. Ο διαχωρισμός γίνεται ανάλογα με το εάν απεβίωσε ο ασθενής στη διάρκεια των 30 ημερών (κλάση 1) ή όχι (κλάση 2). Πριν την παράθεση των αποτελεσμάτων, αξίζει να σημειωθεί πως το συγκεκριμένο dataset, παρόλο που περιέχει την ίδια πληροφορία με το AML Long Term Analysis dataset με μόνη διαφορά τον καθορισμό των κλάσεων, αποτελεί δύσκολο πρόβλημα ταξινόμησης καθώς η διαφορά στις κατανομές των κλάσεων είναι μεγάλη.

Classifier Fusion με βάση το καλύτερο AUROC	Μέθοδος Fusion	Ταξινομητές 1ου επιπέδου	Sensitivity	Specificity	FPR	FNR	AUROC	Diversity
	Majority	LDC-NBC-LSSVM	76,67%	60,00%	40,00%	23,33%	68,33%	0,459
	Minimum	LDC-LSSVM	72,69%	62,00%	38,00%	27,31%	67,35%	0,263
	Maximum	LDC-LSSVM	72,69%	62,00%	38,00%	27,31%	67,35%	0,263
	Average	LDC-LSSVM	72,69%	62,00%	38,00%	27,31%	67,35%	0,263
	Product	LDC-LSSVM	72,69%	62,00%	38,00%	27,31%	67,35%	0,263
	Decision Templates	LDC-RBNC-LSSVM	74,62%	60,00%	40,00%	25,39%	67,31%	0,624

Πίνακας 8: Αποτελέσματα της καλύτερης επίδοσης (με βάση το AUROC) όλων των fusion μεθόδων για το AML Short Term Analysis.

Στον παραπάνω πίνακα, παρουσιάζονται για το AML Short Term Analysis dataset, οι συνδυασμοί που πέτυχαν το μέγιστο AUROC για κάθε fusion μέθοδο. Από όλες τις μεθόδους, την καλύτερη απόδοση εμφανίζει (όπως και στο AML Long Term Analysis) το Majority Voting με $AUROC_{majority} = 68,33\%$ ενώ την χαμηλότερη τα Decision Templates με $AUROC_{D.T.} = 67,31\%$. Αξίζει να σημειωθεί, ότι ο συνδυασμός LDC-NBC-LSSVM που εμφανίζεται στο Majority Voting έχει μικρότερο diversity από τον συνδυασμό LDC-RBNC-LSSVM που εμφανίζεται στα Decision Templates αλλά παρόλα αυτά πετυχαίνει την μεγαλύτερη απόδοση. Ακόμη, ο συνδυασμός LDC-LSSVM έχει μικρότερο diversity από τον LDC-RBNC-LSSVM αλλά εμφανίζει μεγαλύτερη απόδοση. Έτσι, δεν παρατηρείται, ούτε και εδώ, κάποια αντιστοιχία μεταξύ του diversity και της απόδοσης. Επιπλέον, οι ταξινομητές που εμφανίζονται σε όλους τους συνδυασμούς είναι ο LSSVM και ο LDC. Αυτό που έχει ιδιαίτερη σημασία, είναι ότι όπως και στο AML Long Term Analysis έτσι και σε αυτό το dataset, η μέθοδος του Majority Voting παρόλο που είναι η απλούστερη πετυχαίνει την μεγαλύτερη απόδοση.

Classifier Fusion με βάση το καλύτερο diversity	Μέθοδος Fusion	Ταξινομητές 1ου επιπέδου	Sensitivity	Specificity	FPR	FNR	AUROC	Diversity
	Majority	LDC-RBNC	85,90%	35,00%	65,00%	14,10%	60,45%	0,822
	Minimum	LDC-RBNC	78,08%	51,00%	49,00%	21,92%	64,54%	0,822
	Maximum	LDC-RBNC	78,08%	51,00%	49,00%	21,92%	64,54%	0,822
	Average	LDC-RBNC	78,08%	51,00%	49,00%	21,92%	64,54%	0,822
	Product	LDC-RBNC	78,08%	51,00%	49,00%	21,92%	64,54%	0,822
	Decision Templates	LDC-RBNC	74,36%	59,00%	41,00%	25,64%	66,68%	0,822

Πίνακας 9: Αποτελέσματα των fusion μεθόδων για τον συνδυασμό των ταξινομητών που εμφανίζει το μέγιστο diversity για το AML Short Term Analysis.

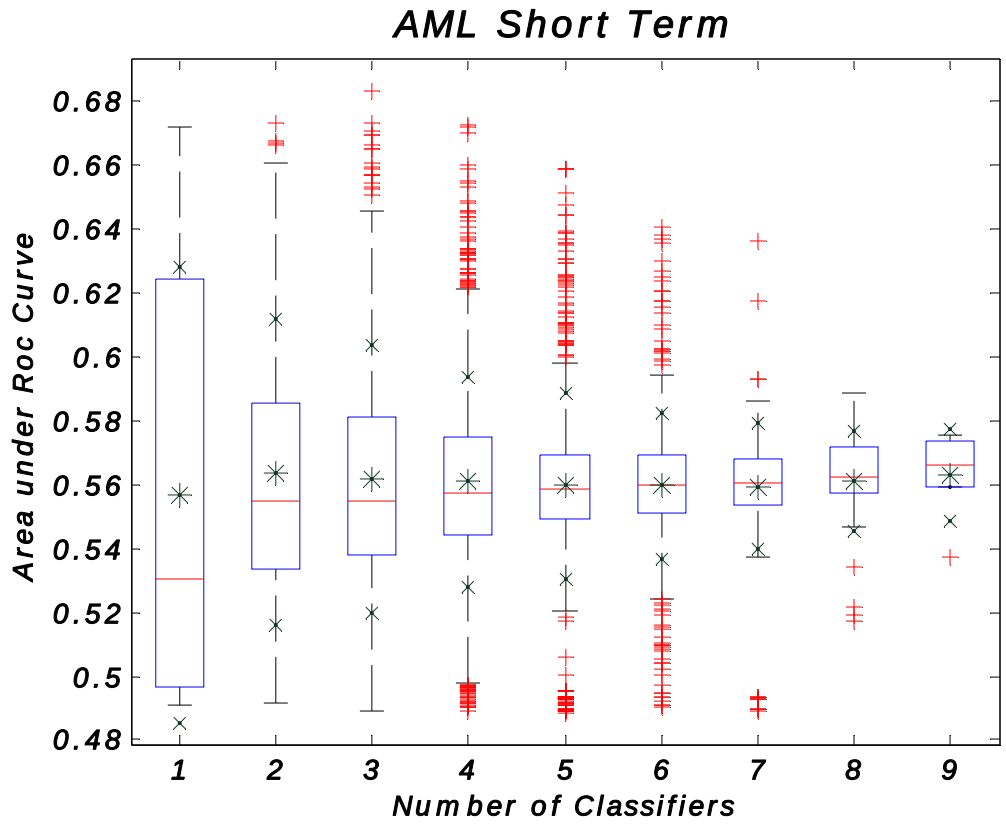
Στον παραπάνω πίνακα, παρουσιάζονται για το AML Short Term Analysis dataset, τα αποτελέσματα του συνδυασμού με το μεγαλύτερο diversity για όλες τις fusion μεθόδους. Τον συνδυασμό αποτελούν οι ταξινομητές LDC και RBNC και το diversity τους είναι 0,822. Από όλες τις μεθόδους, τα Decision Templates πέτυχαν την

υψηλότερη απόδοση με $AUROC_{D.T.} = 66{,}68\%$ ενώ το Majority Voting παρουσίασε την χαμηλότερη απόδοση με $AUROC_{Majority} = 60{,}45\%$. Συγκρίνοντας τον συνδυασμό με την καλύτερη απόδοση του Πίνακα 9 (LDC-RBNC) με αυτόν του Πίνακα 8 (LDC-NBC-LSSVM), διαπιστώνεται πως ο συνδυασμός LDC-NBC-LSSVM με diversity κατά πολύ μικρότερο του LDC-RBNC, πετυχαίνει υψηλότερη απόδοση.

Classifier Fusion με όλους τους ταξινομητές	Μέθοδος Fusion	Ταξινομητές 1ου επιπέδου	Sensitivity	Specificity	FPR	FNR	AUROC	Diversity
	Majority	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	95,51%	12,00%	88,00%	4,49%	53,76%	0,503
	Minimum	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	90,13%	22,00%	78,00%	9,87%	56,06%	0,503
	Maximum	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	89,87%	22,00%	78,00%	10,13%	55,94%	0,503
	Average	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	92,69%	22,00%	78,00%	7,31%	57,35%	0,503
	Product	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	92,05%	23,00%	77,00%	7,95%	57,53%	0,503
	Decision Templates	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	80,39%	34,00%	66,00%	19,62%	57,19%	0,503

Πίνακας 10: Αποτελέσματα των fusion μεθόδων για τον συνδυασμό που περιέχει όλους τους ταξινομητές για το AML Short Term Analysis.

Στον παραπάνω πίνακα παρουσιάζονται για το AML Short Term Analysis dataset, τα αποτελέσματα των fusion μεθόδων όταν στον συνδυασμό συμμετέχουν όλοι οι διαθέσιμοι ταξινομητές. Την μεγαλύτερη απόδοση πετυχαίνει η μέθοδος του Product Rule με $AUROC_{\Pr oduct} = 57{,}53\%$ ενώ την χαμηλότερη απόδοση παρουσιάζει το Majority Voting με $AUROC_{Majority} = 53{,}75\%$. Στον συγκεκριμένο συνδυασμό παρατηρείται έντονα το φαινόμενο του υψηλού Sensitivity σε σχέση με το Specificity. Αυτό πρακτικά σημαίνει πως ο συνδυασμός καταφέρνει να ταξινομήσει σωστά το μεγαλύτερο ποσοστό των δειγμάτων της μεγάλης κλάσης αλλά αποτυγχάνει στα δείγματα της μικρής κλάσης. Αυτό βεβαίως οφείλεται στην αρχική κατανομή των κλάσεων στο συγκεκριμένο dataset.



Σχήμα 8: Confidence Intervals για κάθε περίπτωση αριθμού ταξινομητών που συμμετέχουν στον συνδυασμό για το AML Short Term Analysis.

Στο παραπάνω σχήμα, παρουσιάζονται για το AML Short Term Analysis dataset, τα Confidence Intervals του AUROC για όλες τις περιπτώσεις, ανάλογα με τον αριθμό των ταξινομητών που συμμετέχουν στον συνδυασμό και ακολουθούνται οι ίδιοι συμβολισμοί που χρησιμοποιήθηκαν και στο AML Long Term Analysis dataset. Όπως φαίνεται και από το σχήμα, όταν οι ταξινομητές λειτουργούν μεμονωμένα το AUROC παίρνει τιμές σε ένα αρκετά μεγαλό διάστημα τιμών που κυμαίνεται από 0,5% έως 0,62%. Στην περίπτωση που οι ταξινομητές συνδυάζονται ανά δύο, το AUROC κυμαίνεται από 0,53% έως 0,58%. Καθώς ο αριθμός των ταξινομητών αυξάνεται, το κάτω όριο του confidence interval αυξάνεται, ενώ το άνω όριο αρχικά μειώνεται και στη συνέχεια παρουσιάζει μία μικρή αύξηση. Η μέση τιμή των AUROCs δεν παρουσιάζει σταθερή πορεία αν και παραμένει για όλες τις περιπτώσεις στα ίδια επίπεδα. Πάντως και στο συγκεκριμένο dataset, η αύξηση του αριθμού των ταξινομητών που συμμετέχουν στον συνδυασμό προκαλεί μείωση του διαστήματος του άνω και του κάτω ορίου του confidence interval του AUROC καθιστώντας το σύστημα πιο σταθερό. Συγκριτικά με τις AUROCs που έχουν επιτευχθεί στο AML

Long Term Analysis dataset, εδώ οι AUROCs είναι πιο χαμηλές και αυτό οφείλεται στην κατανομή των κλάσεων στα δύο datasets. Η αναλογία των κλάσεων στην long term analysis είναι 31,83% - 68,17% ενώ η αναλογία αυτή στην short term analysis γίνεται 13,16% - 86,84%. Αυτό σημαίνει πως είναι πιο δύσκολο στο short term analysis να εκπαιδευτούν οι ταξινομητές και να πετυχούν επομένως υψηλές αποδόσεις. Έτσι, η όποια έλλειψη στην εκπαίδευση μεταφέρεται στις αποδόσεις των μεμονωμένων ταξινομητών και αυτές με τη σειρά τους επηρεάζουν τις αποδόσεις των fusion μεθόδων. Προφανώς, είναι ακατόρθωτο, όταν οι αποδόσεις του level 1 κυμαίνονται από 0,5% έως 0,62%, οι αποδόσεις του συνδυασμού να ξεπεράσουν αυτά τα όρια.

4.2.3 Αποτελέσματα για το Breast Cancer Recursion dataset

Στην ενότητα αυτή, παρουσίαζονται τα αποτελέσματα για το Breast Cancer Recursion dataset. Το dataset αυτό, αφορά ασθενείς που είχαν προσβληθεί στο παρελθόν από καρκίνο στο στήθος και είχαν υποβληθεί σε θεραπεία. Οι ασθενείς εξετάζονται πλέον με βάση την επανεμφάνιση (κλάση 1) ή όχι (κλάση 2) του όγκου.

Classifier Fusion με βάση το καλύτερο AUROC	Μέθοδος Fusion	Ταξινομητές 1ου επιπέδου	Sensitivity	Specificity	FPR	FNR	AUROC	Diversity
	Majority	LDC-NBC-LSSVM	78,18%	52,86%	47,14%	21,82%	65,52%	0,417
	Minimum	LDC-NBC	73,64%	59,64%	40,36%	26,36%	66,64%	0,307
	Maximum	LDC-NBC	73,64%	59,64%	40,36%	26,36%	66,64%	0,307
	Average	LDC-NBC	73,64%	59,64%	40,36%	26,36%	66,64%	0,307
	Product	LDC-NBC	73,64%	59,64%	40,36%	26,36%	66,64%	0,307
	Decision Templates	NBC-RBNC	72,73%	62,50%	37,50%	27,27%	67,61%	0,740

Πίνακας 11: Αποτελέσματα της καλύτερης επίδοσης (με βάση το AUROC) όλων των fusion μεθόδων για το Breast Cancer Recursion.

Στον παραπάνω πίνακα, παρουσιάζονται οι συνδυασμοί που πέτυχαν το μέγιστο AUROC για κάθε fusion μέθοδο στο Breast Cancer Recursion dataset. Από

όλες τις μεθόδους, την καλύτερη απόδοση εμφανίζουν τα Decision Template με $AUROC_{D.T.} = 67{,}61\%$ ενώ την χειρότερη εμφανίζει το Majority Voting με $AUROC_{Majority} = 65{,}52\%$. Επιπλέον, ο συνδυασμός NBC-RBNC που εμφανίζεται στα Decision Templates έχει το μεγαλύτερο diversity σε σχέση με όλους τους υπόλοιπους, αλλά ο LDC-NBC-LSSVM, που παρουσιάζει την μικρότερη απόδοση, δεν έχει και το μικρότερο diversity. Έτσι, δεν παρατηρείται, ούτε και εδώ, κάποια αντιστοιχία μεταξύ του diversity και της απόδοσης. Ακόμη, οι ταξινομητές που εμφανίζονται στους περισσότερους συνδυασμούς είναι ο NBC και ο LDC.

Classifier Fusion με βάση το καλύτερο diversity	Μέθοδος Fusion	Ταξινομητές 1ου επιπέδου	Sensitivity	Specificity	FPR	FNR	AUROC	Diversity
	Majority	LDC-RBNC	83,94%	31,79%	68,21%	16,06%	57,86%	0,799
	Minimum	LDC-RBNC	75,30%	54,64%	45,36%	24,70%	64,97%	0,799
	Maximum	LDC-RBNC	75,46%	54,29%	45,71%	24,55%	64,87%	0,799
	Average	LDC-RBNC	75,30%	54,29%	45,71%	24,70%	64,79%	0,799
	Product	LDC-RBNC	75,46%	54,29%	45,71%	24,55%	64,87%	0,799
	Decision Templates	LDC-RBNC	70,30%	58,93%	41,07%	29,70%	64,62%	0,799

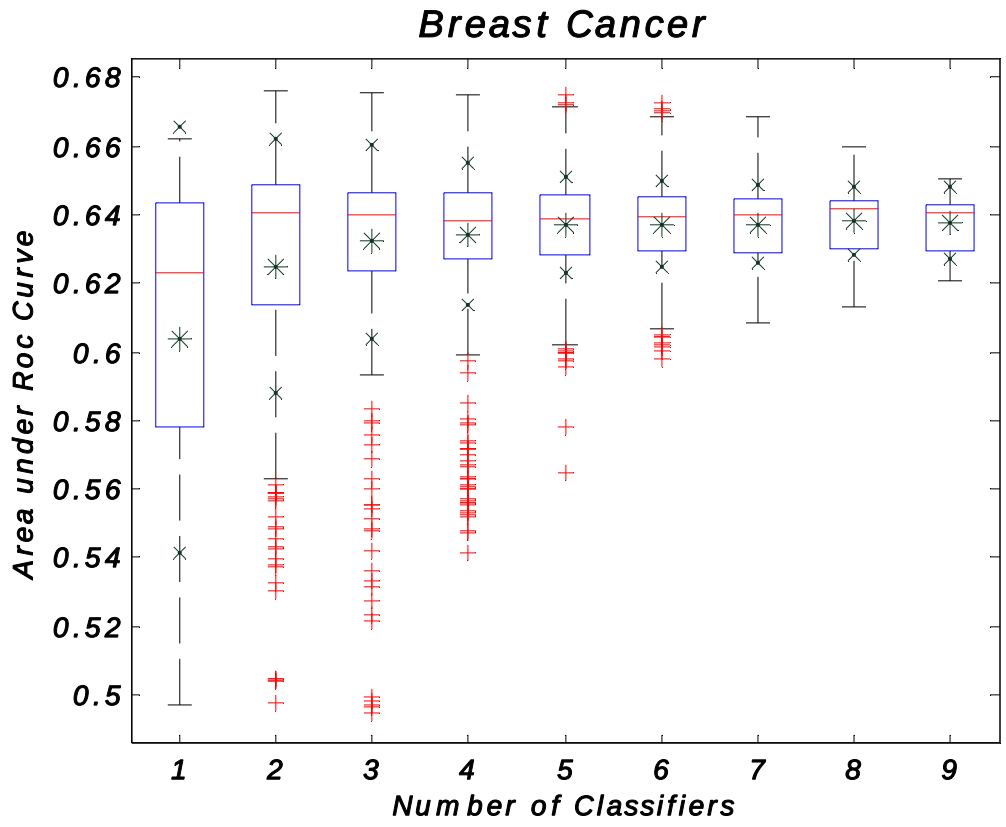
Πίνακας 12: Αποτελέσματα των fusion μεθόδων για τον συνδυασμό των ταξινομητών που εμφανίζει το μέγιστο diversity για το Breast Cancer Recursion.

Στον παραπάνω πίνακα, παρουσιάζονται για το Breast Cancer Recursion dataset τα αποτελέσματα του συνδυασμού με το μεγαλύτερο diversity για όλες τις fusion μεθόδους. Τον συνδυασμό αποτελούν οι ταξινομητές LDC και RBNC και το diversity τους είναι 0,799. Από όλες τις μεθόδους ο Minimum Rule πέτυχε την υψηλότερη απόδοση με $AUROC_{Minimum} = 64{,}97\%$ ενώ το Majority Voting παρουσίασε την χαμηλότερη απόδοση με $AUROC_{Majority} = 57{,}86\%$. Συγκρίνοντας το καλύτερο AUROC που δίνει ο συνδυασμός LDC-RBNC με το καλύτερο AUROC του Πίνακα 11 που το δίνει ο συνδυασμός NBC-RBNC, διαπιστώνεται πως ο συνδυασμός NBC-RBNC με μικρότερο diversity από τον LDC-RBNC, πετυχαίνει αρκετά μεγαλύτερη απόδοση.

Classifier Fusion με όλους τους ταξινομητές	Μέθοδος Fusion	Ταξινομητές 1ου επιπέδου	Sensitivity	Specificity	FPR	FNR	AUROC	Diversity
	Majority	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	90,61%	33,57%	66,43%	9,39%	62,09%	0,445
	Minimum	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	78,94%	49,29%	50,71%	21,06%	64,11%	0,445
	Maximum	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	78,64%	50,00%	50,00%	21,36%	64,32%	0,445
	Average	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	83,79%	42,14%	57,86%	16,21%	62,97%	0,445
	Product	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	83,33%	44,64%	55,36%	16,67%	63,99%	0,445
	Decision Templates	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	73,33%	56,79%	43,21%	26,67%	65,06%	0,445

Πίνακας 13: Αποτελέσματα των fusion μεθόδων για τον συνδυασμό που περιέχει όλους τους ταξινομητές για το Breast Cancer Recursion.

Στον παραπάνω πίνακα παρουσιάζονται για το Breast Cancer Recursion dataset, τα αποτελέσματα των fusion μεθόδων όταν στον συνδυασμό συμμετέχουν όλοι οι διαθέσιμοι ταξινομητές. Την μεγαλύτερη απόδοση πετυχαίνουν για ακόμη μία φορά τα Decision Templates με $AUROC_{D.T.} = 65{,}06\%$ ενώ η χαμηλότερη συναντάται και πάλι στη μέθοδο του Majority Voting με $AUROC_{Majority} = 62{,}09\%$. Τα Decision Templates εδώ, ξεπερνούν την απόδοση των Decision Templates με τον συνδυασμό LDC-RBNC όπου το diversity είναι μέγιστο. Επιπλέον, όπως και στο AML Short Term Analysis dataset, έτσι και εδώ, παρατηρείται έντονα το φαινόμενο του υψηλού Sensitivity σε σχέση με το Specificity.



Σχήμα 9: Confidence Intervals για κάθε περίπτωση αριθμού ταξινομητών που συμμετέχουν στον συνδυασμό για το Breast Cancer Recursion.

Στο παραπάνω σχήμα, παρουσιάζονται για το Breast Cancer Recursion dataset, τα Confidence Intervals του AUROC για όλες τις περιπτώσεις, ανάλογα με τον αριθμό των ταξινομητών που συμμετέχουν στον συνδυασμό και ακολουθούνται οι ίδιοι συμβολισμοί που χρησιμοποιήθηκαν και στα προηγούμενα datasets. Όπως φαίνεται και από το σχήμα, όταν οι ταξινομητές λειτουργούν μεμονωμένα το AUROC παίρνει τιμές σε ένα αρκετά μεγάλο διάστημα τιμών που κυμαίνεται από 0,58% έως 0,64%. Στην περίπτωση που οι ταξινομητές συνδυάζονται ανά δύο, το AUROC κυμαίνεται από 0,61% έως 0,65%. Καθώς ο αριθμός των ταξινομητών αυξάνεται, το όρια του confidence interval αυξάνονται αρχικά και στη συνέχεια διατηρούνται στα ίδια περίπου επίπεδα με μηδαμινές αυξομειώσεις. Από την άλλη, η μέση τιμή των AUROCs παρουσιάζει μία συνεχή αύξηση. Σε αντίθεση με τα δύο προηγούμενα dataset που μελετήθηκαν, εδώ, το fusion λειτουργεί καλύτερα από τους μεμονωμένους ταξινομητές. Καταρχήν, όπως φαίνεται και από το παραπάνω σχήμα, είναι σημαντική η αύξηση της απόδοσης του μοντέλου ταξινόμησης όταν οι ταξινομητές συνδυάζονται. Εκτός από αυτό, παρατηρείται όπως και στα προηγούμενα datasets ελάττωση των

ορίων του confidence interval του AUROC στις περιπτώσεις που οι ταξινομητές συνδυάζονται σε σχέση με την περίπτωση που λειτουργούν μεμονωμένα και αυτό όπως αναφέρθηκε και στα προηγούμενα συνεπάγεται πιο σταθερό σύστημα. Γενικά, και σε αυτό το dataset, παρόλο που το fusion των ταξινομητών βελτίωσε έστω και λίγο τα αποτελέσματα, η απόδοση παραμένει μικρή. Στην συγκεκριμένη περίπτωση, εκτός της κατανομής των κλάσεων που είναι δυσανάλογη, σημαντικός ανασταλτικός παράγοντας είναι ο μικρός αριθμός των συνολικών αντικειμένων του dataset.

4.2.4 Αποτελέσματα για το Breast Cancer Diagnosis dataset

Στην ενότητα αυτή, παρουσίαζονται τα αποτελέσματα για το Breast Cancer Diagnosis dataset. Το dataset αυτό, αφορά ασθενείς που εμφάνισαν όγκο στο στήθος και το πρόβλημα έγκειται στον διαχωρισμό τους με βάση την κακοήθεια (κλάση 1) ή καλοήθεια (κλάση 2) του όγκου.

Classifier Fusion με βάση το καλύτερο AUROC	Μέθοδος Fusion	Ταξινομητές 1ου επιπέδου	Sensitivity	Specificity	FPR	FNR	AUROC	Diversity
	Majority	KNNC-LDC-QDC-RNNC-FISHER-HSSVM-LSSVM	98,25%	94,44%	5,56%	1,75%	96,35%	0,349
	Minimum	LDC-HSSVM-LSSVM	97,90%	94,88%	5,12%	2,11%	96,39%	0,479
		LDC-RBNC-HSSVM-LSSVM	97,90%	94,88%	5,12%	2,11%	96,39%	0,492
	Maximum	LDC-HSSVM-LSSVM	97,90%	94,88%	5,12%	2,11%	96,39%	0,479
		LDC-RBNC-HSSVM-LSSVM	97,90%	94,88%	5,12%	2,11%	96,39%	0,492
	Average	KNNC-QDC-RNNC-Fisher-LSSVM	97,72%	95,61%	4,39%	2,28%	96,67%	0,306
	Product	LDC-HSSVM-LSSVM	97,98%	94,74%	5,26%	2,02%	96,36%	0,479
	Decision Templates	KNNC-QDC-RNNC-Fisher-HSSVM-LSSVM	98,07%	95,18%	4,82%	1,93%	96,62%	0,384
		KNNC-QDC-RNNC-RBNC-Fisher-HSSVM-LSSVM	98,07%	95,18%	4,82%	1,93%	96,62%	0,408

Πίνακας 14: Αποτελέσματα της καλύτερης επίδοσης (με βάση το AUROC) όλων των fusion μεθόδων για το Breast Cancer Diagnosis.

Στον Πίνακα 14, παρουσιάζονται για το Breast Cancer Diagnosis dataset, οι συνδυασμοί που πέτυχαν το μέγιστο AUROC για κάθε fusion μέθοδο. Από όλες τις μεθόδους, την καλύτερη απόδοση πέτυχε ο Average Rule με $AUROC_{Average} = 96{,}67\%$ ενώ την χειρότερη το Majority Voting με $AUROC_{Majority} = 96{,}35\%$. Καταρχήν, είναι προφανές πως σε αυτό το dataset έχουν επιτευχθεί πολύ καλύτερες αποδόσεις σε σχέση με τα προηγούμενα και κατά δεύτερον όλες οι fusion μέθοδοι που εφαρμόστηκαν πέτυχαν την ίδια σχεδόν μέγιστη AUROC (η διαφόρα έγκειται στα δεκαδικά ψηφία). Αυτό οφείλεται πρωτίστως στο γεγονός ότι το συγκεκριμένο dataset έχει περισσότερα αντικείμενα από τα προηγούμενα datasets (και επομένως οι ταξινομητές του πρώτου επιπέδου έχουν εκπαιδευτεί καλύτερα) και επιπλέον, το dataset αυτό δεν παρουσιάζει τόσο μεγάλη διαφορά στην κατανομή των κλάσεων. Ακόμη, όπως φαίνεται και από τον παραπάνω πίνακα, οι μέγιστες τιμές του AUROC για κάθε μέθοδο, προήλθαν από συνδυασμό ταξινομητών με μικρό diversity.

Classifier Fusion με βάση το καλύτερο diversity	Μέθοδος Fusion	Ταξινομητές 1ου επιπέδου	Sensitivity	Specificity	FPR	FNR	AUROC	Diversity
	Majority	RBNC-HSSVM	49,21%	42,84%	57,16%	50,79%	46,02%	0,982
	Minimum	RBNC-HSSVM	0,00%	100,00%	0,00%	100,00%	50,00%	0,982
	Maximum	RBNC-HSSVM	0,00%	100,00%	0,00%	100,00%	50,00%	0,982
	Average	RBNC-HSSVM	0,00%	100,00%	0,00%	100,00%	50,00%	0,982
	Product	RBNC-HSSVM	0,00%	100,00%	0,00%	100,00%	50,00%	0,982
	Decision Templates	RBNC-HSSVM	100,00%	0,00%	100,00%	0,00%	50,00%	0,982

Πίνακας 15: Αποτελέσματα των fusion μεθόδων για τον συνδυασμό των ταξινομητών που εμφανίζει το μέγιστο diversity για το Breast Cancer Diagnosis.

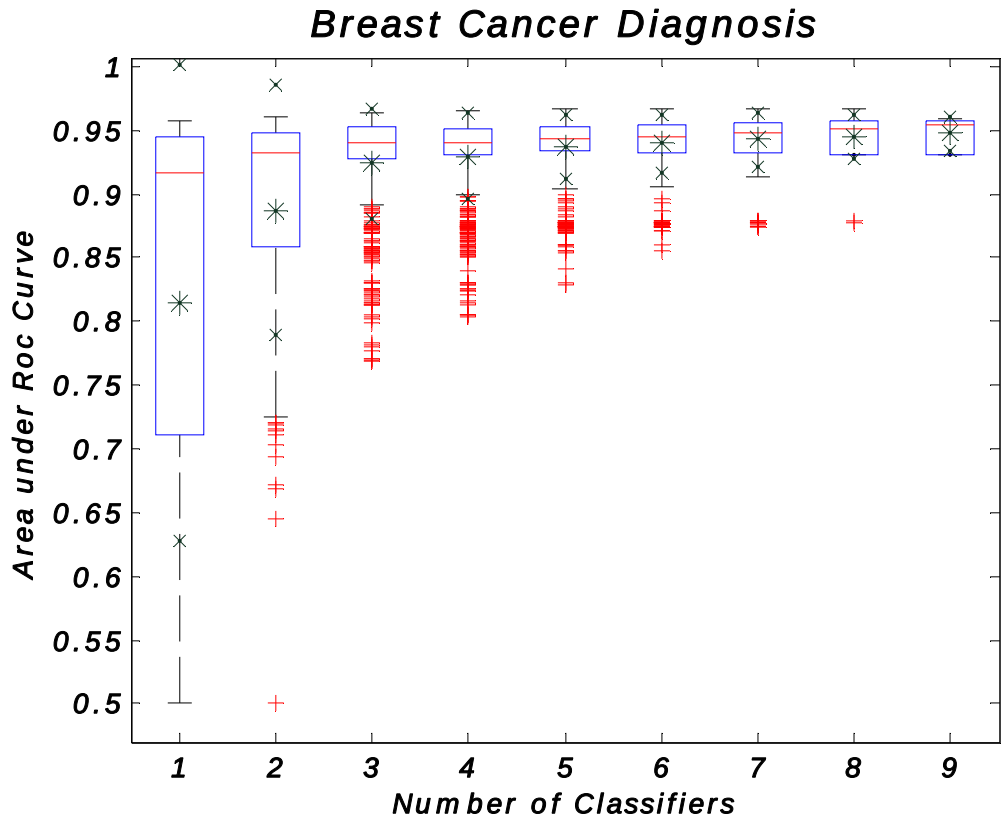
Στον Πίνακα 15, παρουσιάζονται για το Breast Cancer Diagnosis dataset, τα αποτελέσματα του συνδυασμού με το μεγαλύτερο diversity για όλες τις fusion μεθόδους. Τον συνδυασμό αποτελούν οι ταξινομητές RBNC και HSSVM και το diversity τους είναι 0,982. Από όλες τις μεθόδους, η Majority Voting εμφάνισε την χειρότερη απόδοση με $AUROC_{Majority} = 46{,}02\%$ ενώ όλες οι υπόλοιπες πέτυχαν την μεγαλύτερη, για τον συγκεκριμένο συνδυασμό, απόδοση με $AUROC = 50\%$. Ο

συνδυασμός αυτός, παρόλο που εμφανίζει το μέγιστο diversity, φαίνεται πως δεν λειτουργεί καθόλου καλά μιας και φαίνεται να ταξινομεί τα αντικείμενα με τυχαίο τρόπο ($AUROC = 50\%$). Επίσης, η τόσο χαμηλή απόδοση αυτού του συνδυασμού, οφείλεται και στο γεγονός ότι σε όλες τις μεθόδους σχεδόν, όλα τα αντικείμενα αντιστοιχίζονται σε μία και μόνο κλάση με αποτελέσματα οι διαφορές στο sensitivity και στο specificity να είναι τεράστιες.

Classifier Fusion με όλους τους ταξινομητές	Μέθοδος Fusion	Ταξινομητές 1ου επιπέδου	Sensitivity	Specificity	FPR	FNR	AUROC	Diversity
	Majority	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	98,68%	92,54%	7,46%	1,32%	95,61%	0,344
	Minimum	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	87,90%	98,39%	1,61%	12,11%	93,14%	0,344
	Maximum	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	87,81%	98,39%	1,61%	12,19%	93,10%	0,344
	Average	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	97,72%	93,86%	6,14%	2,28%	95,79%	0,344
	Product	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	92,90%	97,52%	2,49%	7,11%	95,21%	0,344
	Decision Templates	ΟΛΟΙ ΟΙ ΤΑΞΙΝΟΜΗΤΕΣ	97,81%	94,15%	5,85%	2,19%	95,98%	0,344

Πίνακας 16: Αποτελέσματα των fusion μεθόδων για τον συνδυασμό που περιέχει όλους τους ταξινομητές για το Breast Cancer Diagnosis.

Στον Πίνακα 16, παρουσιάζονται για το Breast Cancer Diagnosis dataset, τα αποτελέσματα των fusion μεθόδων όταν στον συνδυασμό συμμετέχουν όλοι οι διαθέσιμοι ταξινομητές. Για άλλη μία φορά, τα Decision Templates εμφανίζουν υψηλότερη απόδοση σε σχέση με τις υπόλοιπες μεθόδους με $AUROC_{D.T.} = 95{,}98\%$ ενώ την χαμηλότερη απόδοση παρουσιάζει το Maximum rule με $AUROC_{Maximum} = 93{,}10\%$. Επιπλέον, ο συνδυασμός όλων των ταξινομητών, παρόλο το μικρό diversity, λειτουργεί σαφώς καλύτερα από τον συνδυασμό RBNC-HSSVM, που παρουσιάζει τη μέγιστη διαφορετικότητα, σε όλες τις fusion μεθόδους.



Σχήμα 10: Confidence Intervals για κάθε περίπτωση αριθμού ταξινομητών που συμμετέχουν στον συνδυασμό για το Breast Cancer Diagnosis.

Στο Σχήμα 10, παρουσιάζονται για το Breast Cancer Diagnosis dataset, τα Confidence Intervals του AUROC για όλες τις περιπτώσεις, ανάλογα με τον αριθμό των ταξινομητών που συμμετέχουν στον συνδυασμό και ακολουθούνται οι ίδιοι συμβολισμοί που χρησιμοποιήθηκαν και στα προηγούμενα datasets. Όπως φαίνεται και από το σχήμα, όταν οι ταξινομητές λειτουργούν μεμονωμένα το AUROC παίρνει τιμές σε ένα αρκετά μεγάλο διάστημα τιμών που κυμαίνεται από 0,71% έως 0,94%. Στην περίπτωση που οι ταξινομητές συνδυάζονται ανά δύο, το AUROC κυμαίνεται από 0,86% έως 0,95%. Καθώς ο αριθμός των ταξινομητών αυξάνεται, το όρια του confidence interval αυξάνονται αρχικά και στη συνέχεια διατηρούνται στα ίδια περίπου επίπεδα με μηδαμινές αυξομειώσεις. Από την άλλη, η μέση τιμή των AUROCs παρουσιάζει μία συνεχή αύξηση. Όπως και στην περίπτωση του Breast Cancer dataset, και εδώ, το fusion λειτουργεί καλύτερα από τους μεμονωμένους ταξινομητές. Έτσι πετυχαίνεται σημαντική αύξηση της απόδοσης του μοντέλου όταν οι ταξινομητές συνδυάζονται και ταυτόχρονα το σύστημα γίνεται πιο σταθερό από την άποψη της απόδοσης που πετυχαίνει. Όπως σχολιάστηκε και στον Πίνακα 14, στο

συγκεκριμένο dataset, η απόδοση είναι υψηλή και αυτό οφείλεται ταυτόχρονα στην κατανομή των κλάσεων και στον μεγάλο αριθμό των αντικειμένων του dataset.

ΚΕΦΑΛΑΙΟ 5

ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΕΡΑΙΤΕΡΩ ΕΡΕΥΝΑ

5.1 Συμπεράσματα

Στην διπλωματική αυτή εργασία μελετήθηκαν και υλοποιήθηκαν διάφορες μέθοδοι συνδυασμού (fusion) των ταξινομητών, οι οποίες εφαρμόστηκαν σε ιατρικά προβλήματα ταξινόμησης. Για να υπάρχει μία πιο σφαιρική εικόνα για το κατά πόσο ο συνδυασμός των ταξινομητών υπερτερεί σε απόδοση σε σχέση με το να εφαρμοστούν οι ταξινομητές μεμονωμένα, χρησιμοποιήθηκαν τέσσερα datasets. Τα datasets επιλέχθηκαν έτσι ώστε να μην έχουν ίδιες κατανομές κλάσεων όπως και ίδιο αριθμό συνολικών αντικειμένων.

Οι fusion μέθοδοι που μελετήθηκαν ήταν οι: Majority Voting, Maximum/Minimum/Average/Product Rule και τα Decision Templates. Από τις 6 αυτές μεθόδους η πιο απλή είναι το Majority Voting μιας και κάνει χρήση των crisp αποτελεσμάτων των level 1 ταξινομητών ενώ η πιο σύνθετη είναι η μέθοδος των Decision Templates, που βασίζεται σε πιο πολύπλοκους υπολογισμούς. Ο Πίνακας 17 περιέχει συνοπτικά τα αποτελέσματα που παρουσιάστηκαν στο Κεφάλαιο 4. Στον πίνακα αυτό, δεν παρουσιάζονται αριθμητικά αποτελέσματα παρά μόνο τα ονόματα των μεθόδων που εμφάνισαν την καλύτερη και την χειρότερη απόδοση για το κάθε dataset. Όπως φαίνεται και στον παρακάτω πίνακα, η μέθοδος του Majority Voting ήταν αυτή που τις περισσότερες φορές παρουσίασε την μικρότερη απόδοση, ενώ τα Decision Templates πέτυχαν αρκετές φορές την μέγιστη απόδοση.

Γενικά, και στα τέσσερα datasets, ο συνδυασμός των ταξινομητών δεν προκάλεσε θεαματική αύξηση της απόδοσης του μοντέλου ταξινόμησης. Αυτό που επιτεύχθηκε και στις τέσσερις περιπτώσεις ήταν πιο σταθερά συστήματα από την άποψη του διαστήματος των τιμών της απόδοσης. Πιο συγκεκριμένα, και στα δύο datasets που αφορούν την οξεία μυελοειδή λευχαιμία, η απόδοση του συστήματος, παρόλο που παρουσίασε μία μικρή πτώση, κατάφερε με τον συνδυασμό των

ταξινομητών να γίνει πιο σταθερή. Στην περίπτωση του Breast Cancer Recursion και του Breast Cancer Diagnosis dataset, η απόδοση αυξήθηκε και επιπλέον και εδώ έγινε πιο σταθερή. Η κατανομή των κλάσεων στα AML datasets λειτούργησε ανασταλτικά αρχικά στην απόδοση των μεμονωμένων ταξινομητών στο level 1, αφού οι ταξινομητές εμφάνισαν δυσκολία στην ταξινομήση των αντικειμένων της μικρής κλάσης. Στην περίπτωση του Breast Cancer Recursion, εκτός της κατανομής των κλάσεων το πρόβλημα έγκειται στον μικρό αριθμό αντικειμένων του dataset. Έτσι, ήταν αναμενόμενο, στα τρία αυτά datasets να μην παρατηρηθεί αρχικά υψηλή απόδοση στους ταξινομητές του level 1 και κατόπιν και στο στάδιο του fusion (level 2).

		AML Long Term	**AML Short Term**	**Breast Cancer**	**Breast Cancer Diagnosis**
Καλύτερο AUROC	**Μέγιστη απόδοση**	Majority	Majority	D.T.	Average
	Ελάχιστη απόδοση	Average Product	D.T.	Majority	Majority
Μέγιστο Diversity	**Μέγιστη απόδοση**	D.T.	D.T.	Minimum	Όλες εκτός Majority
	Ελάχιστη απόδοση	Majority	Majority	Majority	Majority
Όλοι οι ταξινομητές	**Μέγιστη απόδοση**	D.T.	Product	D.T.	D.T.
	Ελάχιστη απόδοση	Majority	Majority	Majority	Maximum

Πίνακας 17: Συνοπτικός πίνακας και για τα τέσσερα datasets στον όποιο φαίνονται τα ονόματα των μεθόδων που παρουσίασαν την μέγιστη και την ελάχιστη απόδοση για κάθε περίπτωση με βάση την οποία παρουσιάστηκαν τα αποτελέσματα στο Κεφάλαιο 4.

Αντιθέτως, στο Breast Cancer Diagnosis dataset, η κατανομή των κλάσεων ήταν περισσότερο ομοιόμορφη και επιπλέον το dataset αυτό περιείχε τα περισσότερα δείγματα από τα τρία προηγούμενα. Το γεγονός αυτό, έγινε αντιληπτό στην απόδοση του level 1, η οποία με την εφαρμογή του συνδυασμού των ταξινομητών αυξήθηκε και σταθεροποιήθηκε.

Ένα ακόμη στοιχείο το οποίο παρατηρηθήκε και στα τέσσερα datasets είναι ο μη συσχετισμός του diversity των ταξινομητών με την απόδοση της fusion μεθόδου

στην οποία συμμετέχουν οι ταξινομητές. Έτσι, παρατηρήθηκαν συνδυασμοί με υψηλό diversity και πολύ χαμηλή απόδοση αλλά επίσης παρατηρήθηκαν και συνδυασμοί με πολύ χαμηλό diversity και υψηλή απόδοση. Βεβαίως, είναι γνωστό ότι το diversity από μόνο του δεν αποτελεί κριτήριο επιλογής κάποιου συνδυασμού έναντι των υπολοιπών, αλλά σημαντικός παράγοντας είναι και η απόδοση των μεμονωμένων ταξινομητών που συμμετέχουν στον συνδυασμό. Επομένως, χρειάζεται να έχουμε στη διάθεση μας ταξινομητές που να εμφανίζουν υψηλή απόδοση και ταυτόχρονα να έχουν μεγάλο diversity μεταξύ τους ώστε να αποδόσει τα μέγιστα η διαδικασία του fusion.

Επιπλεόν, σε όλα τα datasets ο συνδυασμός με το μεγαλύτερο diversity και ο συνδυασμός που περιέχει όλους τους ταξινομητές αποδίδουν λιγότερο από άλλους συνδυασμούς. Αυτό είναι λογικό και για τις δύο περιπτώσεις. Στην περίπτωση του συνδυασμού με το μεγαλύτερο diversity συνήθως επιλέγονται, στα συγκεκριμένα datasets και λόγω της κατανομής των κλάσεων, συνδυασμοί που οι ταξινομητές αντιστοιχίζουν σωστά δείγματα από τις αντίθετες κλάσεις και έτσι υπάρχει διαφωνία στις ταξινομήσεις τους με αποτέλεσμα το diversity να είναι υψηλό.Στην περίπτωση του συνδυασμού με όλους τους διαθέσιμους ταξινομητές, η απόδοση παραμένει και πάλι σε χαμηλά επίπεδα, λόγω του ότι οι περισσότεροι ταξινομητές αδυνατούν να αντιστοιχίσουν σωστά τα δείγματα της μικρής κλάσης και έτσι το σφάλμα από την ταξινόμηση όταν λειτουργούν μεμονωμένα μεταφέρεται και στον συνδυασμό.

5.2 Περαιτέρω έρευνα

Επανασχεδιασμός του πρώτου επιπέδου

Όπως αναφέρθηκε και στο Κεφάλαιο 1, η διαδικασία της ταξινόμησης περιλαμβάνει πολλά στάδια και ακόμη δεν είναι εκ των προτέρων γνωστό ποια μπορεί να είναι η βέλτιστη απόδοση που μπορεί να επιτευχθεί για ένα συγκεκριμένο πρόβλημα. Έτσι ο σχεδιαστής του μοντέλου ταξινόμησης μπορεί να ακολουθήσει πολλές εναλλακτικές λύσεις στην προσπάθεια του να αυξήσει την απόδοση του συστήματος. Αυτό μπορεί να σημαίνει είτε την συλλογή περισσότερων δεδομένων για το πρόβλημα, είτε την επιλογή διαφορετικών ή περισσότερων ή και λιγότερων χαρακτηριστικών που περιγράφουν τα αντικείμενα. Μπορούν, ακόμη, να χρησιμοποιηθούν και άλλες

μέθοδοι διαχωρισμού των datasets σε training και test set. Επιπλέον, μπορεί να κριθεί αναγκαίο ακόμη και η χρησιμοποίηση νέων ταξινομητών.

Χρήση περισσότερων fusion μεθόδων

Στην εργασία αυτή χρησιμοποιήθηκαν έξι μέθοδοι συνδυασμού των ταξινομητών. Για την περαιτέρω ανάπτυξη αυτής της εργασίας θα μπορούσαν να χρησιμοποιηθούν και νέες μέθοδοι ή και παραλλαγές αυτών που υλοποιήθηκαν. Για παράδειγμα, στην μέθοδο των Decision Templates θα είχε ενδιαφέρον να δοκιμαστούν και άλλα κριτήρια μέτρησης απόστασης εκτός της Ευκλείδιας απόστασης που χρησιμοποιήθηκε, όπως για παράδειγμα η συμμετρική απόσταση.

ΠΑΡΑΡΤΗΜΑ

ΠΕΡΙΓΡΑΦΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΤΩΝ DATASETS

A.1 BREAST CANCER RECURSION DATASET

Το dataset για τον καρκίνο του στήθους προήλθε από το University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia και συγκεκριμένα από τους M. Zwitter και M. Soklic που συνέλλεξαν τα δεδομένα.

Το dataset αποτελείται από 286 ασθενείς, οι οποίες στο παρελθόν είχαν εμφανίσει όγκο στο στήθος και ακολούθησαν κάποια θεραπεία. Ανάλογα με την επανεμφάνιση ή μη του όγκου στο στήθος, οι ασθενείς κατατάσσονται σε δύο κλάσεις: η πρώτη κλάση, περιέχει 85 ασθενείς στις οποίες υπήρξε επανεμφάνιση του όγκου, ενώ στη δεύτερη κλάση ανήκουν 201 ασθενείς που δεν παρουσίασαν επανεμφάνιση του όγκου (Πίνακας 18).

Κλάσεις	# ασθενών	Ποσοστό επί του συνόλου
Επανεμφάνιση όγκου	85	29.72%
Μη επανεμφάνιση όγκου	201	70.28%

Πίνακας 18: Κατανομή ασθενών στις δύο κλάσεις για το Breast Cancer dataset.

Κάθε ασθενής στο dataset περιγράφεται από 9 χαρακτηριστικά, τα οποία περιγράφονται στη συνέχεια.

- **Age**: *10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99*

Το χαρακτηριστικό αυτό, αναφέρεται στην ηλικία της ασθενούς σε χρόνια και παίρνει ως τιμή ένα από τα παραπάνω διαστήματα.

- **menopause**: *lt40, ge40, premeno*

Το χαρακτηριστικό αυτό, αναφέρεται στο αν η ασθενής δεν έχει περάσει από το στάδιο της εμμηνόπαυσης (premeno), αν πέρασε την εμμηνόπαυση πριν την ηλικία των 40 (lt40) ή μετά (ge40).

- **tumor-size**: *0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59*

Το χαρακτηριστικό αυτό, αναφέρεται στο μέγεθος του όγκου σε χιλιοστά και παίρνει ως τιμή ένα από τα παραπάνω διαστήματα.

- **inv-nodes**: *0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39*

Το χαρακτηριστικό αυτό, αναφέρεται στο πλήθος των όζων που εξαπλώνονται και παίρνει ως τιμή ένα από τα παραπάνω διαστήματα.

- **node-caps***: yes, no*

Αναφέρεται στο εάν ο όγκος αναπτύσσεται μέσα σε περίβλημα (yes) ή όχι (no).

- **deg-malign**: *1, 2, 3*

Αναφέρεται στη διαβάθμιση της κακοήθειας (μικρή: 1, μεσαία: 2, μεγάλη: 3).

- **breast**: *left, right*

Το χαρακτηριστικό αυτό, δηλώνει σε ποιο στήθος εμφανίστηκε ο όγκος (αριστερό: left, δεξί: right).

- **breast-quad**: *left-up, left-low, right-up, right-low, central*

Το χαρακτηριστικό αυτό, αναφέρεται σε ποιο σημείο του στήθους εμφανίστηκε ο όγκος (πάνω αριστερά: left-up, κάτω αριστερά: left-low, πάνω δεξιά: right-up, κάτω δεξιά: right-low, κεντρικά: central).

- **irradiat**: *yes, no*

Αναφέρεται στο εάν η ασθενής είχε υποβληθεί σε θεραπεία με ακτινοβολία (yes) ή όχι (no).

A.2 AML DATASET

Το dataset για την οξεία μυελοειδή λευχαιμία (Acute Myeloid Leukemia), ανακτήθηκε σύμφωνα με το πρωτόκολλο AML99 και προήλθε από την ομάδα της Gimema, μία ιταλική ομάδα που έχει ασχοληθεί με την έρευνα σχετικά με αιματολογικές ασθένειες [13]. Το dataset αποτελείται από 509 ασθενείς. Σύμφωνα με το AML99 πρωτόκολλο, οι ασθενείς υποβλήθηκαν σε ιατρικές εξετάσεις την $30^{η}$ και την $90^{η}$ μέρα από την έναρξη της θεραπείας, με σκοπό να διαπιστωθεί η εξέλιξη της ασθένειας από την επίδραση της θεραπείας στο διάστημα των 30 και των 90 ημερών, αντίστοιχα. Ο μεταπτυχιακός φοιτητής Γιώργος Μανίκης σε συνεργασία με ομάδα του πανεπιστημίου της Πίζας (university of Pisa) και του Ινστιτούτου Καρκίνου του Μιλάνου (Cancer Institute of Milan) κατέληξαν στην μελέτη του συγκεκριμένου dataset με δύο διαφορετικούς τρόπους, την Short Term Analysis και την Long Term Analysis. Σύμφωνα με την Short Term Analysis, την $30^{η}$ μέρα από την έναρξη της θεραπείας, οι ασθενείς κατατάχθηκαν σε δύο κλάσεις: η πρώτη κλάση αφορά αυτούς που πέθαναν κατά τη διάρκεια των 30 ημερών (induction death) και η δεύτερη κλάση αφορά όλους τους υπόλοιπους ασθενείς (αυτούς που θεραπεύτηκαν είτε πλήρως (complete remission) είτε μερικώς (partial remission) και αυτούς που δεν παρουσίασαν καμία βελτίωση (resistant)). Στην Long Term Analysis, την $90^{η}$ μέρα οι ασθενείς διαχωρίστηκαν σε δύο κλάσεις : στην πρώτη κλάση κατατάχθηκαν οι ασθενείς που θεραπεύτηκαν πλήρως (complete remission) και στην δεύτερη κλάση όλοι οι

υπόλοιποι (αυτοί που θεραπεύτηκαν μερικώς (partial remission), που δεν παρουσίασαν καμία βελτίωση (resistant) και αυτοί που πέθαναν (induction death)). Οι παρακάτω πίνακες δείχνουν την κατανομή των ασθενών στις κλάσεις στην Short (Πίνακας 19) και στην Long Term Analysis (Πίνακας 20).

Κλάσεις	#ασθενών	Ποσοστό επί του συνόλου
Θάνατος στην διάρκεια της θεραπείας	67	13.16%
Όλοι οι υπόλοιποι	442	86.84%

Πίνακας 19: Κατανομή ασθενών στις κλάσεις για την Short Term Analysis.

Κλάσεις	#ασθενών	Ποσοστό επί του συνόλου
Πλήρης υποχώρηση της ασθένειας	347	68.17%
Όλοι οι υπόλοιποι	162	31.83%

Πίνακας 20: Κατανομή ασθενών στις κλάσεις για την Long Term Analysis.

Κάθε ασθενής, στην Short και στην Long Term Analysis, περιγράφεται από τα παρακάτω χαρακτηριστικά:

- **Sex**: *male, female*

Αναφέρεται στο φύλο του ασθενούς (άνδρας: male, γυναίκα: female).

- **WBC_dia**: *0.4 – 400.0*

Αναφέρεται στον αριθμό των λευκών αιμοσφαιρίων του αίματος στο στάδιο της διάγνωσης (υπολογισμένο σε 10^9 / *lt*).

- **PS_dia**: *1, 2, 3, 4*

Το χαρακτηριστικό αυτό, αναφέρεται στην κατάσταση του ασθενή όσον αφορά την οξεία μυελοειδή λευχαιμία, όπως έχει κατηγοριοποιηθεί από την παγκόσμια οργάνωση

υγείας (World Health Organization). Ο Πίνακας 21 περιέχει τις κατηγορίες της κατάστασης του ασθενή σύμφωνα με την W.H.O.

PS_dia	
1	Ασθενείς με χαρακτηριστικές γενετικές ανωμαλίες, π.χ. εναλλαγή κάποιων χρωμοσωμάτων
2	Ασθενείς με πολλαπλή δυσπλασία
3	Ασθενείς που είχαν υποβληθεί στο παρελθόν σε χημειοθεραπεία και/ή ακτινοβολία και στη συνέχεια εμφάνισαν οξεία μυελοειδή λευχαιμία
4	Ασθενείς που δεν υπάγονται σε κάποια από τις παραπάνω κατηγορίες

Πίνακας 21: Κατηγορίες κατάστασης του ασθενή για την οξεία μυελοειδή λευχαιμία, σύμφωνα με την παγκόσμια οργάνωση υγείας (World Health Organization).

.

- **Bl_bm_dia**: *30.0 – 99.0*

Αναφέρεται στο ποσοστό των βλαστοκυττάρων στον μυελό των οστών κατά τη διάγνωση.

- **Hb_on**: *3.9 – 15.9*

Αναφέρεται στα επίπεδα αιμοσφαιρίνης στο αίμα κατά τη διάρκεια της θεραπείας και είναι υπολογισμένο σε gm/dl .

- **PLTS_on**: *3.0 – 870.0*

Αναφέρεται στα επίπεδα αιμοπεταλίων στο αίμα κατά τη διάρκεια της θεραπείας και είναι υπολογισμένο σε $10^9/lt$.

- **Citomol**: *normal kariotype, inv(16), t(8;21), +8, t(11)(q23), t(6;9), t(9;22), inv(3), iperdiploid, complex kariotype, other*

Αναφέρεται στην σύνθεση της πληροφορίας που προέρχεται α) από το είδος της αλλοίωσης που επιφέρει η λευχαιμία στο χρωμόσωμα του κακοήθη κυττάρου του ασθενούς (cytogenetic abnormality) και β) από την βιολογική πληροφορία που λαμβάνεται από το DNA του ασθενή (molecular biology). Πρόκειται δηλαδή, για ένα σύνθετο χαρακτηριστικό που δηλώνει αν υπάρχει κάποια ανωμαλία στο χρωμόσωμα του κακοήθη κυττάρου και τι είδους ανωμαλία είναι αυτή. Θα πρέπει να αναφερθεί πως ο ιατρικός όρος «καρυότυπος» αναφέρεται στην ταξινόμηση των χρωμοσωμάτων ενός ατόμου σύμφωνα με το μέγεθος και το σχήμα. Στον Πίνακα 22, υπάρχει η ερμηνεία των δυνατών τιμών της μεταβλητής citomol.

Citomol	
normal kariotype	Κανενός είδους ανωμαλία στον καρυότυπο
inv(16)	Εναλλαγή γονιδίων στο χρωμόσωμα 16
t(8;21)	Ανταλλαγή γενετικού υλικού μεταξύ των χρωμοσωμάτων 8 και 21
+8	Μετατροπή χρωμοσώματος 8 σε τρίσωμο χρωμόσωμα
t(11)(q23)	Ανωμαλία στο χρωμόσωμα 11 και συγκεκριμένα στο γονίδιο 23
t(6;9)	Ανταλλαγή γενετικού υλικού μεταξύ των χρωμοσωμάτων 6 και 9
t(9;22)	Ανταλλαγή γενετικού υλικού μεταξύ των χρωμοσωμάτων 9 και 22
inv(3)	Εναλλαγή γονιδίων στο χρωμόσωμα 3
iperdiploid	Καρυότυπος με διπλό αριθμό χρωμοσωμάτων
complex kariotype	Σύνθετη ανωμαλία στον καρυότυπο
other	Άλλου είδους ανωμαλία

Πίνακας 22: Περιγραφή ανωμαλιών στο χρωμόσωμα που σχετίζονται με την οξεία μυελοειδή λευχαιμία και οι οποίες συνθέτουν το χαρακτηριστικό Citomol.

- **Exm_on**: *no, lymphnodes, cutaneous, both, other*

Αναφέρεται στον τύπο των μυελοβλαστικών κυττάρων που εμφανίστηκαν κατά τη διάρκεια της θεραπείας (κανένα: no, λεμφαδένας: lymphnodes, δερματικό: cutaneous, και τα δύο: both, άλλος τύπος: other).

- **Itd**: *negative, positive*
- **Npm2**: *negative, positive*
- **D835**: *negative, positive*

Τα itd, npm2 και d835 αναφέρονται στην ύπαρξη ή μη διαφορετικού τύπου γενετικής μεταβολής του γονιδίου FLT3 (ύπαρξη μεταβολής: positive, μη ύπαρξη μεταβολής: negative).

A.3 BREAST CANCER DIAGNOSIS DATASET

Το δεύτερο dataset για τον καρκίνο του στήθους προήλθε από τους Dr. William H. Wolberg, W. Nick Street και Olvi L. Mangasarian και είναι διαθέσιμο και στο internet [14].

Το dataset αποτελείται από 569 ασθενείς, οι οποίοι έχουν εμφανίσει όγκο στο στήθος. Οι ασθενείς χωρίζονται σε δύο κατηγορίες ανάλογα με το αν πρόκειται για καλοήθη (benign) ή κακοήθη (malignant) όγκο. Από τους 569 ασθενείς οι 357 εμφάνισαν καλοήθη όγκο ενώ οι 212 κακοήθη, όπως φαίνεται και στον παρακάτω πίνακα (Πίνακας 23).

Κλάσεις	#ασθενών	Ποσοστό επί του συνόλου
Κακοήθης όγκος	212	37.26%
Καλοήθης όγκος	357	62.74%

Πίνακας 23: Κατανομή των ασθενών στις δύο κλάσεις για το Breast Cancer Diagnosis dataset.

Κάθε ασθενής στο dataset περιγράφεται από 10 διαφορετικά χαρακτηριστικά που προκύπτουν από μετρήσεις στην εικόνα που απεικονίζονται οι πυρήνες των καρκινικών κυττάρων [15]. Τα χαρακτηριστικά αυτά, είναι τα ακόλουθα:

- **Radius**

Είναι η ακτίνα του κυττάρου μετρημένη ως ο μέσος όρος της απόστασης σημείων που βρίσκονται στην περίμετρο από το κέντρο.

- **Texture**

Αναφέρεται στην υφή του πυρήνα και είναι μετρημένη ως η τυπική απόκλιση από τις gray-scale τιμές.

- **Perimeter**

Είναι η περίμετρος του πυρήνα του κυττάρου.

- **Area**

Εκφράζει το εμβαδό της επιφάνειας του πυρήνα και είναι μετρημένο σε pixels.

- **Smoothness**

Εκφράζει την ομαλότητα της επιφάνειας και είναι μετρημένη σε σχέση με τις τοπικές αποκλίσεις στα μήκη της ακτίνας.

- **Compactness**

Αναφέρεται στην πυκνότητα της μάζας και είναι μετρημένη ως $perimeter^2 / area$.

- **Concavity**

Εκφράζει την ένταση των κοίλων μερών του περιγράμματος.

- **Concave points**

Αναφέρεται στον αριθμό των κοίλων μερών του περιγράμματος.

- **Symmetry**

Αναφέρεται στην συμμετρία του πυρήνα ως προς τη μεγαλύτερη χορδή που περνάει από το κέντρο.

- **Fractal dimension**

Αναφέρεται στην διάσταση του fractal του πυρήνα. Η αύξηση της διάστασης αντιστοιχεί σε ολοένα και πιο μη φυσιολογικό περίγραμμα που σχετίζεται με την κακοήθεια του όγκου.

Θα πρέπει να τονιστεί ότι κάθε χαρακτηριστικό παίρνει πραγματικές τιμές. Για κάθε χαρακτηριστικό έχουν υπολογιστεί ο μέσος όρος, το standard error και ο μέσος όρος των τριών μεγαλύτερων τιμών και έτσι κάθε ασθενής περιγράφεται τελικά από 30 χαρακτηριστικά.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. **Combining Pattern Classifiers, Methods and Algorithms. Ludmila I. Kuncheva. John Wiley & Sons, 2004**
2. **Data Mining: Practical Machine Learning Tools and Techiniques with Java Implementations. Written, Ian H., Eibe Frank. San Diego, CA: Morgan Kaufmann, 2000**
3. **Machine Learning. Mitchell, Tom. M. New York: McGraw-Hill, 1997**
4. **ROC Graphs, Notes, and Practical Considerations for Researchers. T. Fawcett. HP Labs Tech Report, 2004**
5. **Data Mining, Classification: Basic Concepts, Decision Trees and Model Evaluation. Tan, Steinbach, Kumar. Lecture notes on chapter 4. 2004**
6. **A Tutorial on Support Vector Machines for Pattern Recognition. Christopher J. C. Burges. Data Mining and Knowledge Discovery 2:121 - 167, 1998**
7. **Classifier selection for Majority Voting. Dymitr Ruta, Bogdan Gabryst. Information Fusion, Elsevier, 2005**
8. **Is independence good for combining classifiers? L.I. Kuncheva, C.J. Whitaker, C.A. Shipp. Pattern Recognition, 2000**
9. **Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Ludmilla I. Kuncheva, Christopher J. Whitaker. Machine Learning, vol.51, pages 181-207, 2003**
10. **Relationships between combination methods and measures of diversity in combining classifiers. Catherine A. Shipp, Ludmila I. Kuncheva. Information Fusion, Elsevier, 2002**
11. **Decision Templates for Multiple Classifier Fusion: An Experimental Comparison. L. I. Kuncheva, J. C. Bezdek, R. P. W. Duin. Pattern Recognition, vol. 34, no. 2, 2001**
12. **http://prtools.org/**
13. **www.gimema.org**
14. **www.ics.uci.edu/~mlearn/MLSummary.html**
15. **Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. William H. Wolberg,**

W. Nick Street, O. L. Mangasarian. Cancer Letters vol. 77, pages 163-171, 1994