ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

«ΕΞΕΤΑΣΗ ΜΕΙΓΜΑΤΟΣ ΕΙΔΙΚΩΝ ΜΕΘΟΔΩΝ ΓΙΑ ΤΗΝ ΤΑΞΙΝΟΜΗΣΗ ΓΟΝΙΔΙΩΝ»

«MIXTURE OF EXPERTS APPROACH IN GENE CLASSIFICATION»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΒΙΓΔΙΔΕΛΗ ΘΕΟΔΩΡΑ

ΕΠΙΤΡΟΠΗ ΔΙΠΛΩΜΑΤΙΚΗΣ

Επιβλέπων καθηγητής: Καθηγητής Ζερβάκης Μιχάλης

Εξεταστική επιτροπή: Καθηγητής Ζερβάκης Μιχάλης

Καθηγητής Λιάβας Αθανάσιος

Καθηγητής Πετράκης Ευριπίδης

Χανιά, Μάιος 2012

Θα ήθελα να ευχαριστήσω

Τον καθηγητή κύριο Ζερβάκη Μιχάλη, για την συνεχή βοήθεια και την πολύτιμη καθοδήγησή του, κατά τη διάρκεια της εκπόνησης αυτής της διπλωματικής εργασίας.

Την κυρία Μπέη Κατερίνα, τόσο για τις βιολογικές γνώσεις με τις οποίες εμπλούτισε αυτή την εργασία, αλλά περισσότερο για το κουράγιο και την υποστήριξη που μου προσέφερε στο διάστημα που συνεργαστήκαμε.

Τους καθηγητές κύριο Λιάβα Αθανάσιο και κύριο Πετράκη Ευριπίδη, για τη συνεισφορά τους ως μέλη της εξεταστικής επιτροπής.

Τους λίγους αλλά πολύτιμους φίλους μου, οι οποίοι με τη συμπαράστασή τους, αλλά κυρίως με το χιούμορ τους, με βοήθησαν να αντιμετωπίσω τις δυσκολίες που συνάντησα.

Την οικογένεια μου, που είναι πάντα δίπλα μου και που με την αφοσίωση και την αγάπη της με στηρίζει όλα αυτά τα χρόνια σε οτιδήποτε κάνω.

Οι βιολογικές διεργασίες μελετώνται με τη γονιδιακή ανάλυση, η οποία καταλήγει σε μεγάλο αριθμό γονιδίων, με αποτέλεσμα να γίνονται απαραίτητες οι αυτοματοποιημένες μέθοδοι. Σε αυτή την εργασία, εξετάζουμε την επιρροή των αλγοριθμικών παραμέτρων τόσο στη δύναμη της πρόβλεψης μίας γονιδιακής υπογραφής όσο και στη διαδικασία επιλογής της ίδιας της υπογραφής. Εστιάζουμε σε μία προσέγγιση επιλογής γονιδίων, η οποία εφαρμόζεται στο σύνολο δεδομένων του ζυμομύκητα *Saccharomyces cerevisiae*, με αρκετά διαφορετικές παραμέτρους, και στην εκτίμηση της επιρροής της επιλεγμένης γονιδιακής υπογραφής. Πιο συγκεκριμένα, προσαρμόζουμε μία διαδικασία αναδρομικής εξάλειψης των χαρακτηριστικών, όπου σε κάθε βήμα η προγνωστική δύναμη του συνόλου των υπολοίπων γονιδίων εκτιμάται από διαφορετικούς ταξινομητές καθώς και από μείγματα αυτών. Επιπρόσθετα, πολλαπλές επαναλήψεις αυτής της διαδικασίας μας εξασφαλίζουν την εκχώρηση στατιστικής σημαντικότητας στα γονίδια η οποία βασίζεται στις σχετικές συχνότητες εμφάνισης τους.

Επίσης, μελετάμε το σύστημα της συγχώνευσης ταξινομητών με σκοπό να μειώσουμε αβεβαιότητες που σχετίζονται με τον ταξινομητή που αξιολογεί τη σημαντικότητα των γονιδίων. Πιο ειδικά, ασχολούμαστε με ταξινομητές όπως ο logistic-sigmoid, ο kernel nearest centroid, ο kernel minimum squared error, ο kernel subspace, ο k-nearest neighbor, και το support vector machine, με διαφορετικές παραμέτρους και /ή συναρτήσεις πυρήνα, καθώς και τέσσερις μεθόδους συνδυασμού αυτών όπως, πρώτον, η επιλογή ταξινομητών με διαχωρισμό του χώρου των χαρακτηριστικών σε ομάδες μέσω του self organized map, και δεύτερον, η συγχώνευση ταξινομητών με συνδυασμό βασισμένο σε ασαφή βάρη, σε ένα gradient descent learning αλγόριθμο καθώς και στη μίξη των δύο τελευταίων. Η διαδικασία της επιλογής γονιδίων είναι ενσωματωμένη σε ένα cross-validation σύστημα, με σκοπό να ενισχυθεί η εμπιστοσύνη στα αποτελέσματα. Εξετάζουμε τις διαφορές των γονιδιακών υπογραφών βασιζόμενοι στην πιθανή επικάλυψή τους, και επίσης τη βιολογική περιγραφή των επιλεγμένων γονιδίων, η οποία διεξάγεται με το σύστημα ταξινόμησης MIPS FunCat.

Από τα αποτελέσματα γίνεται φανερό ότι, η σωστή επιλογή ενός αριθμού από σημαντικά γονίδια, προσφέρει καλή προγνωστική ακρίβεια στα μοντέλα, καθώς και ότι η ακρίβεια ταξινόμησης που επιτυγχάνεται από τους συνδυασμούς των ειδικών μπορεί να είναι αρκετά καλύτερη από εκείνη των ατομικών ταξινομητών. Τέλος, καταλήξαμε ότι διαφορετικά συστήματα επιλογής οδηγούν σε ποικίλα μεγέθη γονιδιακής υπογραφής, με διαφορές στα επιλεγμένα γονίδια. Ωστόσο, όταν περιγράφουμε τα γονίδια κάθε υπογραφής φαίνεται ότι επικαλούνται τις ίδιες βιολογικές διεργασίες, με μικρές πιθανές διαφορές στη σχετική συχνότητα συμμετοχής.

The biological processes are studied by the genome analysis, which leads to a large number of genes, thus become necessary, the automated methods. In this study, we examine the influence of algorithmic parameters in the prediction power of a gene signature and in the selection process of the signature. We focus on one gene selection approach applied on a dataset of the budding *Saccharomyces cerevisiae*, with quite different parameters and evaluate the influence on the selected signature. In particular, we adopt a recursive feature elimination process, where at each step the prognostic power of the set of remaining genes is evaluated by five different classifiers as well as by four fusions of these classifiers. Moreover, multiple runs of this process provide us the assignment of statistical significance in the genes, based on their relative frequencies of occurrence.

We also study the fusion scheme in order to reduce uncertainties related to the classifier that evaluates the significance of genes. More specifically, we consider logistic-sigmoid, kernel nearest centroid, kernel minimum squared error, kernel subspace, k-nearest neighbor, and support vector machines classifiers with different parameters and/or kernel functions, as well as four fusion methods, i.e. firstly, classifier selection with feature space partitioning using self organized map, and secondly, classifier fusion with combination based on fuzzy weights, on a gradient descent learning algorithm and on the mixing of the last two. The selection process is embedded into a cross validation scheme in order to enhance the confidence on the generalization of results. We consider the differences of signatures based on the overlap and also the biological annotation of selected genes, using the MIPS FunCat architecture.

We find out that a robustly identification of a number of highly differential genes can offer “good” predictive power to the models, and also the classification accuracy, achieved by the mixtures of experts, can be significantly better than the one of the individual classifiers. Last but not least, we found out that different selection schemes result in a diverse size of gene signature, with differences in the selected genes. Nevertheless, when we annotate the genes of each signature we find out that the same biological processes are invoked, with possibly small differences in the relative frequency of participation.

Περιεχόμενα

Λίστα Πινάκων

ΚΕΦΑΛΑΙΟ 1: Εισαγωγή

Σε αυτό το κεφάλαιο εισάγεται γενικά η έννοια της ταξινόμησης που αποτελεί το αντικείμενο της παρούσας εργασίας. Έπειτα, αναφέρονται κάποιες βασικές μέθοδοι ανάλυσης γονιδιακών δεδομένων στις οποίες απευθύνεται η παρούσα εργασία, καθώς και η προσφορά τους στην επιστήμη.

Δίνονται οι θεμελιώδεις ορισμοί για το γονίδιο και τις μικροσυστοιχίες DNA, ενώ παράλληλα περιγράφεται ο ζυμομύκητας *Saccharomyces cerevisiae*, στο γονιδίωμα του οποίου έχει βασιστεί η εργασία.

Προχωρώντας στις ενότητες του πρώτου κεφαλαίου, παρουσιάζεται το πρόβλημα που η εργασία αυτή καλείται να αντιμετωπίσει, η μεθοδολογία, εν συντομία, με την οποία επιλύει το πρόβλημα καθώς και οι στόχοι της. Επιπλέον, περιγράφεται το σύνολο δεδομένων του σακχαρομύκητα, πάνω στο οποίο πραγματοποιήθηκαν τα πειράματα αυτής της εργασίας.

Τέλος, δίνεται ο σκελετός της εργασίας, δηλαδή ο τρόπος με τον οποίο οργανώνονται τα κεφάλαια από το δεύτερο μέχρι και το έκτο.

1.1 Γενική Εισαγωγή

Οι τεχνολογίες των μικροσυστοιχιών μετρούν [1], μέσω μιας διαδικασίας υβριδοποίησης, τα επίπεδα έκφρασης των γονιδίων σε βιολογικά δείγματα. Η γνώση που κερδίζεται από αυτές τις μελέτες, θεωρείται ολοένα και περισσότερο σημαντική εξαιτίας των δυνατοτήτων που προσφέρουν στην κατανόηση θεμελιωδών ερωτήσεων στη βιολογία και στην κλινική ιατρική. Τα πειράματα μικροσυστοιχιών μπορούν είτε να παρακολουθούν κάθε γονίδιο αρκετές φορές κάτω από ποικίλες συνθήκες είτε να αναλύουν τα γονίδια σε ένα ενιαίο περιβάλλον αλλά από διαφορετικούς τύπους ιστών.

Για την ανάλυση των δεδομένων των μικροσυστοιχιών [2] έχουν αναπτυχθεί πολλές τεχνικές οι οποίες διακρίνονται σε επιβλεπόμενες και μη επιβλεπόμενες. Οι επιβλεπόμενες μέθοδοι απαιτούν τα γονίδια ή οι συνθήκες να σχετίζονται με ετικέτες που παρέχουν πληροφορία σχετική μία προϋπάρχουσα ταξινόμηση. Αυτή η πληροφορία προέρχεται από τον εξωτερικό χώρο του πειράματος μικροσυστοιχιών και μπορεί να περιλαμβάνει γνώση της γονιδιακής λειτουργίας ή ρύθμισης, της υποκατηγορίας μιας αρρώστιας ή της προέλευσης του ιστού ενός κυτταρικού τύπου. Αυτή η πληροφορία ταξινόμησης χρησιμοποιείται για να οδηγήσει την ανάλυση των πειραματικών αποτελεσμάτων των μικροσυστοιχιών και με αυτό τον τρόπο να «επιβλέπει» την ανάλυση.

Κάποιες φορές υπάρχει μία αναλυτική πρόκληση για την κατανόηση του τρόπου με τον οποίο οι νέες ομαδοποιήσεις των συνόλων δεδομένων σχετίζονται με τις ήδη γνωστές ομαδοποιήσεις. Η ταξινόμηση είναι μία επιβλεπόμενη διαδικασία, η οποία χρησιμοποιεί τις γνωστές ομαδοποιήσεις και δημιουργεί κανόνες για την αξιόπιστη εκχώρηση γονιδίων ή συνθηκών σε αυτές τις ομάδες. Απαιτεί κατά κανόνα ένα σύνολο παραδειγμάτων, για το οποίο τα δεδομένα έκφρασης σχετίζονται με κάποιο φαινότυπο ή κάποια κατηγοριοποίηση.

Στις μεθόδους της επιβλεπόμενης μάθησης, ο υπολογιστής μαθαίνει από τα γνωστά πρότυπα και έπειτα μπορεί να τα προσδιορίσει σε νέα σύνολα δεδομένων. Πολλές τεχνικές ταξινόμησης προτύπων έχουν εφαρμοστεί για την ανάλυση των δεδομένων των μικροσυστοιχιών, όπως ένα weighted voting σχήμα, τα support vector machines, οι nearest neighbor ταξινομητές, η discriminant analysis, τα neural networks, και τα probabilistic neural networks. Ένα πλεονέκτημα των τεχνικών επιβλεπόμενης μάθησης είναι το γεγονός ότι τοποθετούν διαφορετικά βάρη στα χαρακτηριστικά, με κριτήριο τη χρησιμότητά τους στη διάκριση ανάμεσα στις κατηγορίες. Έτσι, η επιτυχία των επιβλεπόμενων μεθόδων εξαρτάται από το αν παρέχονται υψηλής ποιότητας σύνολα που διαθέτουν ετικέτες. Στην περίπτωση των δεδομένων των μικροσυστοιχιών, η ταξινόμηση θα αποτύχει αν τα δεδομένα έκφρασης δεν εξηγούν τον εν λόγω φαινότυπο.

Η επιβλεπόμενη μείωση των διαστάσεων ονομάζεται επιλογή χαρακτηριστικών και έχει δύο στόχους. Την επιλογή σχετικών χαρακτηριστικών και τη μείωση του αριθμού των χαρακτηριστικών που είναι απαραίτητα για την εκτέλεση της σωστής ταξινόμησης. Αρκετές στατιστικές και υπολογιστικές προσεγγίσεις έχουν επινοηθεί, παρόλο που η εφαρμογή στα δεδομένα των μικροσυστοιχιών είναι περιορισμένη. Παρ’ όλα αυτά, οι εν λόγω τεχνικές είναι χρήσιμες για τον εντοπισμό των ιδιαίτερων χαρακτηριστικών και για την πραγματοποίηση των αναλύσεων με, πιο γνωστικά, ευαγώγους τρόπους.

1.2 Μέθοδοι Ανάλυσης Γονιδιακών Δεδομένων

Ένα από τα βασικά προβλήματα [3] στην ανάλυση των γονιδιακών εκφράσεων είναι ο καθορισμός των γονιδίων τα οποία εκφράζονται διαφορετικά είτε σε διαφορετικούς ιστούς, είτε στον ίδιο ιστό σε δύο φαινοτυπικά διαφορετικές συνθήκες. Η εύρεση των γονιδίων των οποίων τα επίπεδα έκφρασης σχετίζονται με μία συγκεκριμένη ασθένεια είναι σημαντική για την επιλογή της καταλληλότερης θεραπείας και για την πρόβλεψη της επανεμφάνισής της από τους επιστήμονες. Επιπλέον, κάτι τέτοιο θα διευκόλυνε και τους βιολόγους, αφού επιτρέπει το σχεδιασμό πιο οικονομικών μικροσυστοιχιών DNA, προσαρμοσμένα σε μία συγκεκριμένη ασθένεια, καταγράφοντας τα επίπεδα έκφρασης ορισμένων δεκάδων μόνο γονιδίων. Επιπρόσθετα, η επιλογή ενός υποσυνόλου γονιδίων μειώνει τον κίνδυνο του over fitting, ο οποίος προκύπτει όταν ο αριθμός των χαρακτηριστικών (επίπεδα γονιδιακής έκφρασης) είναι εξαιρετικά μεγαλύτερος από τον αριθμό των δειγμάτων.

Ένα άλλο πρόβλημα, το οποίο σχετίζεται σε μεγάλο βαθμό με το παραπάνω, είναι η ακρίβεια που σημειώνουν τα μοντέλα πρόβλεψης κατά την ταξινόμηση. Μία λάθος πρόβλεψη στο πεδίο μιας συγκεκριμένης ασθένειας μπορεί να προκαλέσει δύο εξίσου σημαντικά προβλήματα: Να χαρακτηριστεί ένα υγιές άτομο ως άτομο που φέρει την εν λόγω ασθένεια, όπως επίσης και να μην αναγνωριστεί ένα άτομο που νοσεί θεωρώντας ότι είναι υγιές, και κατά συνέπεια, να μην του χορηγηθεί η απαραίτητη θεραπεία. Ως εκ τούτου, είναι εξαιρετικά σημαντικό για τους επιστήμονες, η πληροφορία που λαμβάνουν από τα μοντέλα πρόβλεψης, να είναι όσο το δυνατό πιο αξιόπιστη, ώστε οι ίδιοι να είναι ακριβείς στη διάγνωση, πρόβλεψη και θεραπεία των ασθενειών.

Η ανάγκη έγκαιρης και έγκυρης διάγνωσης νέων ασθενών στον τομέα της ιατρικής, οδηγεί στην ανάγκη βελτίωσης της απόδοσης των ατομικών μοντέλων που εκτελούν ταξινόμηση. Για την απόκτηση μίας υψηλής ακρίβειας ταξινόμησης, βοηθά η εύρεση ενός υποσυνόλου σχετικών γονιδίων το οποίο θα αποτελείται από ένα μικρό αριθμό δειγμάτων και ακόμα πιο σημαντικό, από ένα μικρό αριθμό γονιδίων. Γονίδια τα οποία είναι σχετικά με μία δεδομένη ασθένεια, θεωρούνται εκείνα με τα οποία μπορεί να προβλεφθεί σωστά η εμφάνισή της σε νέους ασθενείς, οπότε αξίζουν ιδιαίτερης προσοχής και μελέτης από τους επιστήμονες.

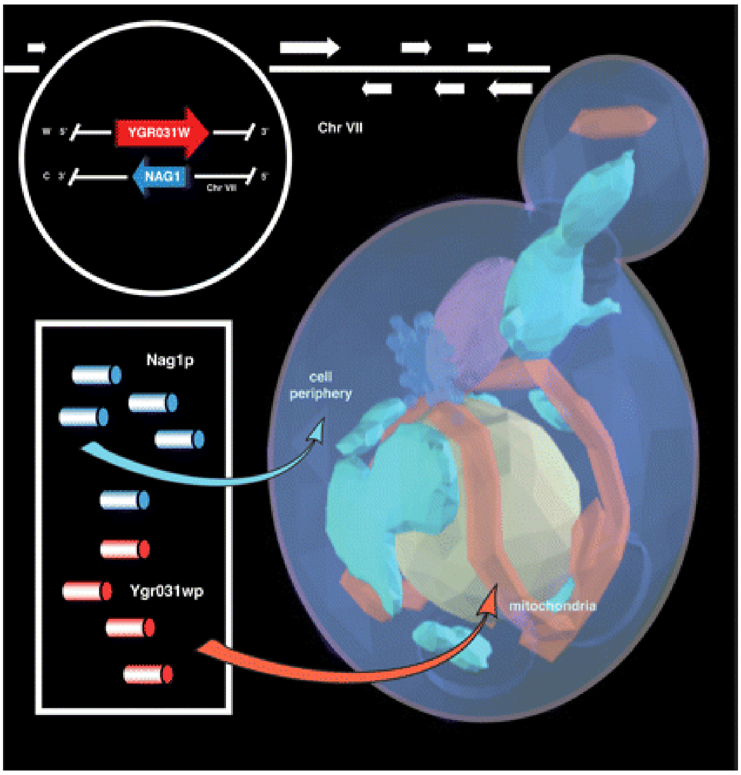
Οι προσδοκίες των ερευνητών και βιολόγων τοποθετούν στο κέντρο του ενδιαφέροντός τους τη βελτίωση της ακρίβειας ταξινόμησης με τη χρήση αποδοτικότερων μοντέλων. Προς αυτή την κατεύθυνση, σημαντικός αρωγός αποτελεί ένας αριθμός από μεθόδους συγχώνευσης ταξινομητών [4] που έχουν πρόσφατα αναπτυχθεί, προτείνοντας μία εναλλακτική προσέγγιση που οδηγεί σε μία πιθανή βελτίωση της απόδοσης ταξινόμησης. Το αντικείμενο όλων των συστημάτων υποστήριξης αποφάσεων είναι η δημιουργία ενός μοντέλου, το οποίο αν του δοθεί μία μικρή ποσότητα από δεδομένα/πληροφορία, είναι ικανό να παράγει σωστές αποφάσεις. Αρκετά συχνά, ειδικά στην ασφάλεια κρίσιμων συστημάτων, η ορθότητα των αποφάσεων που λαμβάνονται είναι ζωτικής σημασίας.

Σύμφωνα με μία προσέγγιση, η πρόοδος των συστημάτων υποστήριξης αποφάσεων πρέπει να βασίζεται στη συνεχόμενη ανάπτυξη των ήδη υπαρχουσών μεθόδων καθώς και στην ανακάλυψη καινούριων. Μία άλλη προσέγγιση προτείνει ότι, αφού τα όρια των υπαρχουσών ατομικών μεθόδων έχουν προσεγγιστεί και είναι δύσκολο να αναπτυχθεί κάποια καλύτερη μέθοδος, η λύση του προβλήματος μπορεί να είναι απλά ο συνδυασμός των υπαρχουσών μεθόδων που εκτελούνται καλά, ελπίζοντας ότι, με αυτό τον τρόπο, θα επιτευχθούν καλύτερα αποτελέσματα.

Αυτή η συγχώνευση πληροφορίας φαίνεται να αξίζει να εφαρμοστεί, όσον αφορά στη μείωση της αβεβαιότητας πρόβλεψης, ενός παράγοντα, δηλαδή, που είναι αρκετά ανασταλτικός για την πρόοδο της επιστήμης αφού αποτελεί τροχοπέδη στην κατανόηση των μοριακών μηχανισμών των ασθενειών. Κάθε ατομική μέθοδος παράγει κάποια λάθη. Ωστόσο, διαφορετικές μέθοδοι που εκτελούνται σε διαφορετικά δεδομένα μπορεί να παράγουν διαφορετικά σφάλματα, και υποθέτοντας ότι όλες οι ατομικές μέθοδοι αποδίδουν καλά, ο συνδυασμός αυτών των πολλαπλών ειδικών μπορεί να μειώσει το συνολικό σφάλμα ταξινόμησης και κατά συνέπεια να δίνει σωστές εξόδους. Οι τεχνικές συγχώνευσης της πληροφορίας έχουν εντατικά ερευνηθεί τα τελευταία χρόνια και η εφαρμοσιμότητά τους στον τομέα της ταξινόμησης έχει δοκιμαστεί ευρέως.

1.3 Γονίδιο και Μικροσυστοιχίες DNA

Τα γονίδια είναι τμήματα του DNA [5] που βρίσκονται στα χρωμοσώματα και υπάρχουν σε εναλλακτικές μορφές οι οποίες ονομάζονται αλληλόμορφα. Τα αλληλόμορφα γονίδια βρίσκονται στην ίδια γονιδιακή θέση στα ομόλογα χρωμοσώματα και ελέγχουν με διαφορετικό τρόπο τις ίδιες ιδιότητες, οι οποίες μεταδίδονται από τους γονείς στους απογόνους. Η διαδικασία με την οποία τα γονίδια μεταδίδονται ανακαλύφθηκε από τον Gregor Mendel και διατυπώθηκε σε αυτό που είναι σήμερα γνωστό ως νόμος διαχωρισμού του Mendel.



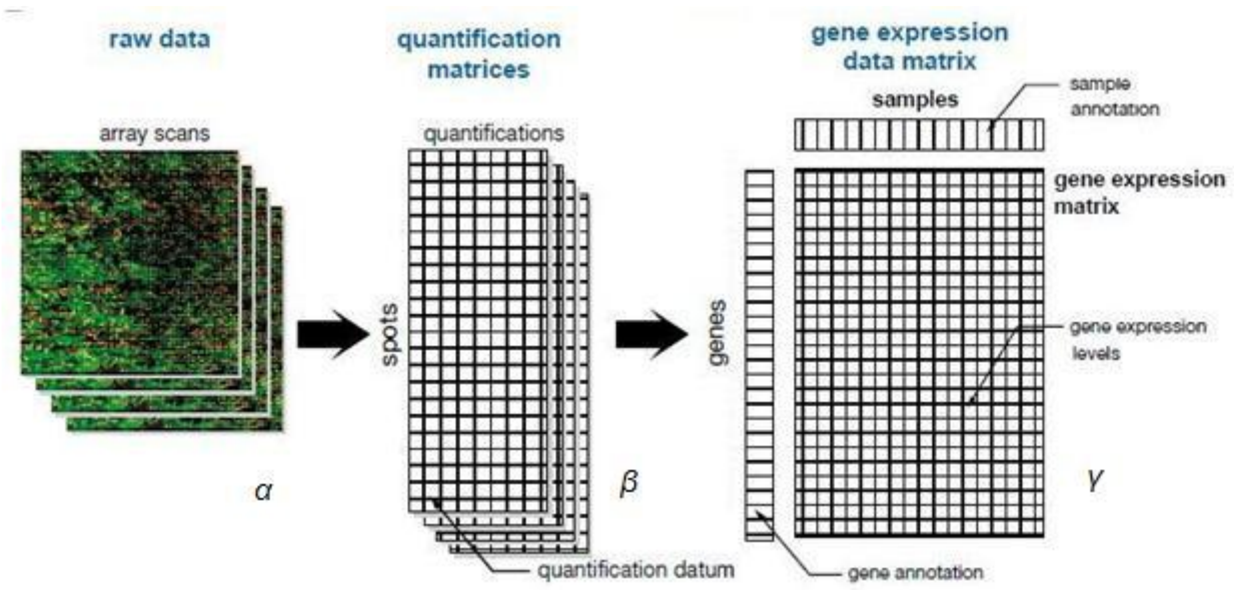
Εικόνα 1: Γονίδια στον ασκομύκητα *Saccharomyces cerevisiae*. Το γονίδιο *NAG1* κωδικοποιεί μια πρωτεΐνη που εντοπίζεται στην περιφέρεια των κυττάρων του σακχαρομύκητα και το γονίδιο YGR031W κωδικοποιεί μια μιτοχονδριακή πρωτεΐνη [6].

Όπως απεικονίζεται στην εικόνα 1, τα γονίδια περιέχουν τους κωδικούς για την παραγωγή συγκεκριμένων πρωτεϊνών. Η πληροφορία που περιέχεται στο DNA δε μετατρέπεται άμεσα σε πρωτεΐνες, αλλά πρέπει πρώτα να μεταγραφεί μέσω μιας διαδικασίας που ονομάζεται μεταγραφή του DNA. Αυτή η διαδικασία λαμβάνει χώρα μέσα στον πυρήνα των κυττάρων μας. Η πραγματική παραγωγή πρωτεϊνών λαμβάνει χώρα στο κυτταρόπλασμα των κυττάρων μας μέσω μιας διαδικασίας που ονομάζεται μετάφραση.

Η ταχεία πρόοδος στην ανάλυση του προσδιορισμού της αλληλουχίας [7] του γονιδιώματος του ανθρώπου, αλλά και άλλων πρότυπων οργανισμών (*Saccharomyces cerevisiae*) έχει οδηγήσει στην ανάπτυξη των μεθόδων για την αξιοποίηση αυτής της πληροφορίας, μέσω του χαρακτηρισμού βιολογικών διεργασιών με νέους τρόπους. Η γνώση των ακολουθιών κωδικοποίησης σχεδόν κάθε γονιδίου σε έναν οργανισμό, για παράδειγμα, δημιουργεί την ανάγκη για μελέτη της ανάπτυξης όλων αυτών μονομιάς, διότι η μελέτη της γονιδιακής έκφρασης, με τα γονίδια να εξετάζονται ένα προς ένα, έχει ήδη παράσχει έναν πλούτο βιολογικών γνώσεων. Για το σκοπό αυτό, μία ποικιλία από τεχνικές έχει εξελιχθεί για να παρακολουθεί, γρήγορα και αποτελεσματικά, την πληθώρα των αντιγράφων για το σύνολο των γονιδίων ενός οργανισμού. Μέσα στη μάζα των αριθμών που παράγονται από αυτές τις τεχνικές, οι οποίες ανέρχονται σε εκατοντάδες δειγμάτων για χιλιάδες ή δεκάδες χιλιάδων γονιδίων, υπάρχει μία τεράστια ποσότητα βιολογικής πληροφορίας.

Τα δεδομένα της γονιδιακής έκφρασης [8] είναι αρκετά πολύπλοκα με την έννοια ότι έχουν νόημα μόνο στο πλαίσιο μιας αναλυτικής περιγραφής των συνθηκών κάτω από τις οποίες παράγονται, συμπεριλαμβανομένης και της ιδιαίτερης κατάστασης του συστήματος διαβίωσης που είναι υπό μελέτη και των διαταραχών στις οποίες έχει υποβληθεί. Σε αντίθεση με το γονιδίωμα ενός οργανισμού, υπάρχουν τόσα πολλά μετάγραφα, όσα και οι κυτταρικοί τύποι που υπάρχουν αλλά και οι εκάστοτε περιβαλλοντικές συνθήκες στις οποίες αναπτύσσονται. Επιπλέον, η σύγκριση των δεδομένων γονιδιακής έκφρασης είναι σημαντικά δυσκολότερη, διότι επί του παρόντος, οι μικροσυστοιχίες δεν μετρούν επίπεδα γονιδιακής έκφρασης σε οποιεσδήποτε αντικειμενικές μονάδες. Στην πραγματικότητα, οι περισσότερες μετρήσεις αναφέρουν μόνο σχετικές αλλαγές στη γονιδιακή έκφραση, χρησιμοποιώντας μία (τιμή) αναφορά(ς) που είναι σπανίως τυποποιημένη. Τέλος, οι διαφορετικές πλατφόρμες μικροσυστοιχιών καθώς και τα πειραματικά σχέδια, παράγουν δεδομένα σε διαφορετικές μορφές και μονάδες και κανονικοποιούνται με διαφορετικούς τρόπους.

Τα ανεπεξέργαστα δεδομένα, από τα πειράματα μικροσυστοιχιών, είναι εικόνες από σαρώσεις μικροσυστοιχιών υβριδοποίησης, που πρέπει να αναλυθούν για να προσδιοριστεί και να ποσοτικοποιηθεί κάθε χαρακτηριστικό (spot) στην εικόνα. Μία ακολουθία DNA μπορεί να εντοπιστεί σε μία μικροσυστοιχία αρκετές φορές. Επιπλέον, μπορεί να εντοπιστούν αρκετές διακριτές ακολουθίες DNA που απεικονίζουν το ίδιο γονίδιο. Για να αποδοθεί μία μοναδική τιμή γι' αυτά, οι αντίστοιχες μετρήσεις πρέπει να συνδυαστούν. Επιπροσθέτως, η ίδια βιολογική συνθήκη μπορεί να μετρηθεί σε πολλαπλά αντίγραφα υβριδοποίησης, και η πληροφορία από όλα τα αντίγραφα πρέπει να συνοψιστεί για να εξαχθεί ένας πίνακας δεδομένων μοναδικών γονιδιακών εκφράσεων. Τελικά, για να συγκριθούν τα δεδομένα μικροσυστοιχιών από αρκετά δείγματα, τα δεδομένα πρέπει να κανονικοποιηθούν κατάλληλα.



Εικόνα 2: Εννοιολογική άποψη των τριών επιπέδων της επεξεργασίας των δεδομένων της μικροσυστοιχίας γονιδιακής έκφρασης. α. Ανεπεξέργαστα δεδομένα (row data), β. Ποσοτικοποίηση πινάκων μικροσυστοιχιών (quantification matrices), γ. Πίνακες δεδομένων γονιδιακής έκφρασης (gene expression data matrix) [8].

Όπως φαίνεται στην εικόνα 2, υπάρχουν, τουλάχιστον, τρία επίπεδα δεδομένων σχετικών με ένα πείραμα μικροσυστοιχιών: α) οι σαρωμένες εικόνες (ανεπεξέργαστα δεδομένα, εικόνα 2α), β) οι ποσοτικές έξοδοι από τη διαδικασία ανάλυσης της εικόνας (ποσοτικοποίηση πινάκων μικροσυστοιχιών, εικόνα 2β) και γ) οι εξαγόμενες μετρήσεις (πίνακες δεδομένων γονιδιακής έκφρασης, εικόνα 2γ). Το μοντέλο αυτού του τρίτου επιπέδου διαθέτει τρία τμήματα: i) περιγραφή των γονιδίων, που μπορεί να διατίθεται σαν σύνδεσμος στις βάσεις δεδομένων της γονιδιακής ακολουθίας, ii) περιγραφή των δειγμάτων, για τα οποία δεν υπάρχουν τρέχουσες δημόσιες εξωτερικές βάσεις δεδομένων (εκτός από την ταξινόμηση των ειδών) και iii) τον πίνακα της γονιδιακής έκφρασης, στον οποίο κάθε θέση περιλαμβάνει πληροφορία που χαρακτηρίζει την έκφραση ενός συγκεκριμένου γονιδίου σε ένα συγκεκριμένο δείγμα (εικόνα 2γ).

1.4 Ο *Saccharomyces cerevisiae* ως Μοντέλο

1.4.1 Γιατί η μαγιά ως πρότυπο σύστημα μελέτης;

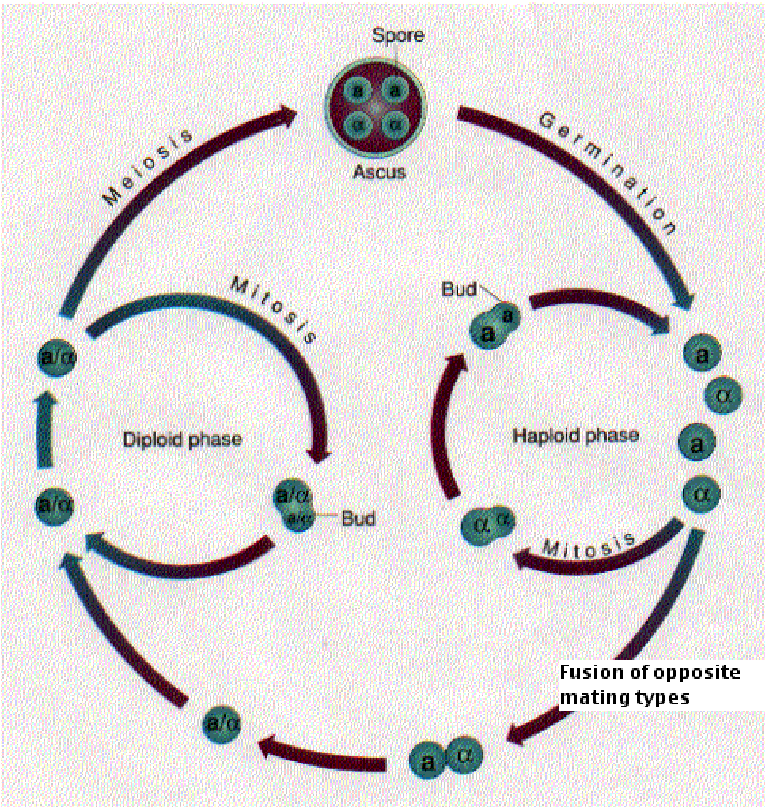
Ο *Saccharomyces cerevisiae* [9] είναι ένας ασκομύκητας ζυμομύκητας που πολλαπλασιάζεται με εκβλάστηση και παράγει ψευδοϋφές. Στην προσπάθεια για μια πληρέστερη κατανόηση του τρόπου με τον οποίο τα κύτταρα ενσωματώνουν διαφορετικές εισόδους σήματος και προσαρμόζονται στις συνεχώς μεταβαλλόμενες περιβαλλοντικές συνθήκες, αρκετός όγκος εργασιών έχει επικεντρωθεί στο μονοκύτταρο ευκαρυωτικό μοντέλο του ζυμομύκητα (*Saccharomyces cerevisiae*). Η αποκωδικοποίηση της πλήρους αλληλουχίας του γονιδιώματός του, το καλά σχολιασμένο γονιδίωμά του, καθώς επίσης και η αξιοσημείωτη «γενετική του ευπείθεια», οδήγησαν σε μία πληθώρα πειραματικών δεδομένων υψηλής ποιότητας, που καθιστούν τη μαγιά ιδανικό υποψήφιο για ποσοτικές προσπάθειες μοντελοποίησης.

1.4.2 Κύκλος ζωής του σακχαρομύκητα

Ο βιολογικός κύκλος του σακχαρομύκητα [10], όπως παρουσιάζεται στην εικόνα 3, αποτελείται από δύο φάσεις: την απλοειδή και τη διπλοειδή φάση. Ο σακχαρομύκητας μπορεί να επιβιώσει και να αναπτυχθεί είτε με τη μορφή διπλοειδών κυττάρων είτε με τη μορφή απλοειδών κυττάρων. Σε κάθε περίπτωση, αναπαράγεται ξανά σχηματίζοντας ασκούς (εξού και το όνομα) με μίτωση.

Τα απλοειδή κύτταρα εμφανίζονται σε δύο διαφορετικούς συζευκτικούς τύπους, τους a ή α (εικόνα 3). Ο τύπος καθορίζεται από την έκφραση ενός γονιδίου σε ένα ενεργό τόπο (active locus) του συζευκτικού τύπου. Τα απλοειδή κύτταρα μπορούν να ζουν επ' αόριστον στην απλοειδή κατάσταση. Ωστόσο, εάν δύο κύτταρα αντίθετων συζευκτικών τύπων «συναντηθούν», μπορούν να συγχωνευτούν και να εισέλθουν στη διπλοειδή φάση του κυτταρικού κύκλου.

Τα κύτταρα στη διπλοειδή φάση είναι πιο ανθεκτικά στις σκληρές περιβαλλοντικές συνθήκες. Όταν αρχίζουν να παρουσιάζονται συνθήκες έλλειψης θρεπτικών στοιχείων (ασιτία), τα διπλοειδή κύτταρα υποβάλλονται σε μείωση σχηματίζοντας τέσσερα απλοειδή σπόρια σε έναν ασκό (ο *Saccharomyces cerevisiae* ανήκει στους ασκομύκητες). Όταν επικρατήσουν ευνοϊκές συνθήκες, τα σπόρια εκβλαστάνουν, παράγοντας τέσσερα απλοειδή κύτταρα ζύμης: δύο τύπου a και δύο τύπου α (εικόνα 3).



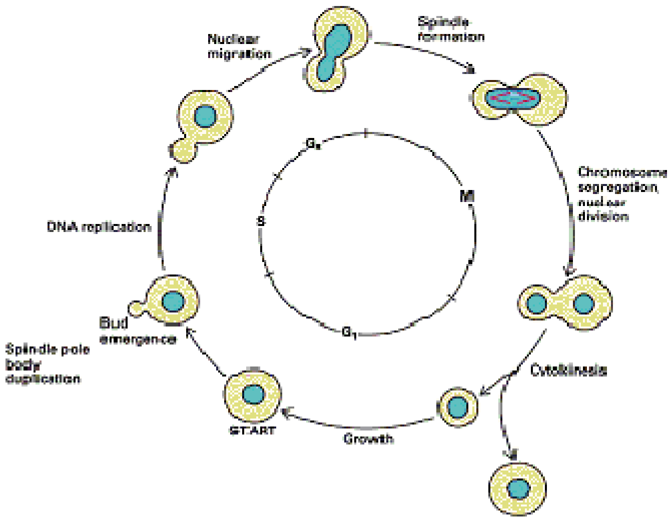
Εικόνα 3: Ο βιολογικός κύκλος του σακχαρομύκητα [10].

1.4.3 Κυτταρικός κύκλος του σακχαρομύκητα *Saccharomyces cerevisiae*

Ο κυτταρικός κύκλος [11], όπως φαίνεται στην εικόνα 4, είναι η διαδοχή των γεγονότων βάσει της οποίας, ένα κύτταρο αναπτύσσεται και χωρίζεται σε δύο θυγατρικά κύτταρα, κάθε ένα από τα οποία περιέχει την πληροφορία και το μηχανισμό που είναι απαραίτητα για να επαναληφθεί η διαδικασία.

Μεταξύ μιας κυτταρικής διαίρεσης και της επόμενης, όλα τα βασικά συστατικά του κυττάρου πρέπει να αντιγραφούν. Το πιο σημαντικό συστατικό είναι το γενετικό υλικό (τα μόρια DNA που υπάρχουν στα χρωμοσώματα), το οποίο πρέπει να αντιγραφεί με ακρίβεια και τα δύο αντίγραφα να διαχωριστούν προσεκτικά στα δύο θυγατρικά κύτταρα.

Οι διαδικασίες της αντιγραφής του DNA και του διαχωρισμού των αδελφών χρωματίδων συμβαίνουν σε χρονικά διακριτές φάσεις του ευκαρυωτικού κυτταρικού κύκλου. Αυτές είναι γνωστές ως φάση S (σύνθεση του DNA) και φάση Μ (μίτωση). Σε γενικές γραμμές, οι S και Μ φάσεις διαχωρίζονται από τις φάσεις G1 και G2. Ο μονοκύτταρος ασκομύκητας *Saccharomyces cerevisiae*, είναι ένα πρότυπο σύστημα για τη μελέτη της ρύθμισης του κυτταρικού κύκλου.



Εικόνα 4: Ο κυτταρικός κύκλος του *Saccharomyces cerevisiae* [11].

1.5 Ορισμός του Προβλήματος και Στόχοι Εργασίας

Από τις προηγούμενες ενότητες καθίσταται σαφές το ολοένα και αυξανόμενο ενδιαφέρον των ερευνητών, στα προφίλ γονιδιακής έκφρασης των συνόλων δεδομένων, με στόχο τη διευκόλυνση της διάγνωσης, πρόγνωσης και θεραπείας των ασθενειών. Η έρευνα κατευθύνεται προς την κατανόηση των μοριακών μηχανισμών των ασθενειών, την ολοκληρωμένη αντίληψη των γονιδιωματικών αλληλεπιδράσεων και την εξαγωγή συγκεκριμένων γονιδιακών/πρωτεϊνικών υπογραφών.

Αρκετές είναι οι μελέτες που έχουν διεξαχθεί στο πεδίο της επιλογής υποσυνόλου γονιδίων για ταξινομητές που βασίζονται σε πυρήνες. Κάποιες από αυτές χρησιμοποιούν μία filter προσέγγιση πάνω σε σύνολα δεδομένων για λευχαιμία [12], ή εφαρμόζουν προσέγγιση επιλογής χαρακτηριστικών που βασίζεται σε ένα βαθμό εξάρτησης a σε σύνολα δεδομένων για καρκίνο εγκεφάλου, προστάτη και μαστού [13]. Τα αποτελέσματα αυτών των ερευνών αναδεικνύουν την επιλογή των χαρακτηριστικών και την ανίχνευση των γονιδίων δεικτών σε παράγοντες καταλυτικούς για την καλύτερη απόδοση των ταξινομητών που βασίζονται σε πυρήνες και επιβεβαιώνουν τη βιολογική σχετικότητα των γονιδίων δεικτών. Ποικίλες μέθοδοι συνδυασμού ατομικών ταξινομητών για την εξαγωγή μίας ακριβέστερης και αποδοτικότερης ταξινόμησης, έχουν εντατικά ερευνηθεί τα τελευταία χρόνια από αρκετές μελέτες [14, 15] καθιστώντας τη χρήση διαφόρων μειγμάτων ειδικών, καρποφόρα προσέγγιση για τη μηχανική μάθηση.

Σε αυτή την εργασία χρησιμοποιείται ο ζυμομύκητας *Saccharomyces cerevisiae* ως πρότυπο σύστημα μελέτης, και εκτελείται ταξινόμηση στο σύνολο δεδομένων του, για την εξαγωγή πειραμάτων υψηλής ποιότητας, αφού το γονιδίωμά του παρουσιάζει πολλές ομοιότητες με το ανθρώπινο. Ο *Saccharomyces cerevisiae* είναι ένας από τους περισσότερο διαδεδομένους οργανισμούς ευκαρυωτικού μοντέλου. Έχει χρησιμοποιηθεί ως μοντέλο για τη μελέτη της ρύθμισης της γονιδιακής έκφρασης, της μεταγωγής σήματος, του κυτταρικού κύκλου, του μεταβολισμού, της απόπτωσης και πολλών άλλων βιολογικών διεργασιών.

Πραγματοποιείται σύγκριση των διαφορετικών μεθόδων ταξινόμησης και μελετάται η προσέγγιση των μειγμάτων ειδικών στη γονιδιακή ταξινόμηση. Επίσης, εξετάζεται η επιρροή των αλγοριθμικών παραμέτρων τόσο στη δύναμη της πρόβλεψης μίας γονιδιακής υπογραφής όσο και στη διαδικασία επιλογής της ίδιας της υπογραφής.

Προτείνουμε μία νέα μεθοδολογία επιλογής χαρακτηριστικών για την επιλογή των πιο σημαντικών γονιδίων, η οποία εφαρμόζεται στο σύνολο δεδομένων του ζυμομύκητα *Saccharomyces cerevisiae,* που παρουσιάζεται αναλυτικά στην επόμενη ενότητα. Στη βάση αυτής της μεθοδολογίας τοποθετείται 1) ο αλγόριθμος RFE-LNW για την επίτευξη της αναδρομικής εξάλειψης των χαρακτηριστικών και 2) οι αλγόριθμοι της εκτίμησης της προγνωστικής δύναμης του συνόλου των υπολοίπων γονιδίων από διαφορετικούς ταξινομητές καθώς και από μείγματα αυτών. Παράλληλα, πολλαπλές επαναλήψεις αυτής της διαδικασίας μας εξασφαλίζουν την εκχώρηση στατιστικής σημαντικότητας στα γονίδια, η οποία βασίζεται στις σχετικές συχνότητες εμφάνισης τους. Η διαδικασία της επιλογής γονιδίων είναι ενσωματωμένη σε ένα five fold cross-validation σύστημα, με σκοπό να ενισχυθεί η εμπιστοσύνη στα αποτελέσματα.

Ως μοντέλα μάθησης και πρόβλεψης χρησιμοποιούνται, τόσο τα support vector machines που τοποθετούνται στο υψηλότερο επίπεδο της ανάπτυξης των μεθόδων ταξινόμησης, όσο και άλλοι ταξινομητές που βασίζονται σε πυρήνα. Πιο συγκεκριμένα, εφαρμόζονται τα svm με πυρήνα γραμμικό και τετραγωνικό, και οι ταξινομητές kernel nearest centroid με πολυωνυμικό πυρήνα δευτέρου βαθμού, kernel minimum squared error machine με Gaussian πυρήνα και kernel subspace με πολυωνυμικό πυρήνα τρίτου βαθμού.

Επιπλέον, αυτή η εργασία παρουσιάζει τέσσερις μεθόδους συνδυασμού των παραπάνω, αλλά και άλλων, ταξινομητών. Η πρώτη μέθοδος μείγματος ειδικών εφαρμόζει συγχώνευση των πέντε παραπάνω ατομικών ταξινομητών σε ένα συνδυασμό βασισμένο σε ασαφή βάρη. Η δεύτερη μέθοδος μείγματος ειδικών εφαρμόζει συγχώνευση δύο logistic-sigmoid ταξινομητών και βασίζεται σε ένα gradient descent learning αλγόριθμο. Η τρίτη μέθοδος μείγματος ειδικών αποτελεί μίξη των δύο πρώτων. Τέλος, η τέταρτη μέθοδος μείγματος ειδικών εφαρμόζει επιλογή δύο ταξινομητών οι οποίο βασίζονται σε αποστάσεις για να εκτελέσουν ταξινόμηση, και πραγματοποιείται με το διαχωρισμό του χώρου των χαρακτηριστικών σε ομάδες μέσω του self organized map.

Εξετάζουμε τις διαφορές των γονιδιακών υπογραφών όλων των προαναφερθέντων μοντέλων, βασιζόμενοι στην πιθανή επικάλυψή τους, και επίσης τη βιολογική περιγραφή των επιλεγμένων γονιδίων, η οποία διεξάγεται με το σύστημα ταξινόμησης MIPS FunCat.

Από τα αποτελέσματα γίνεται φανερό ότι, η σωστή επιλογή ενός αριθμού από σημαντικά γονίδια, προσφέρει καλή προγνωστική ακρίβεια στα μοντέλα, καθώς και ότι η ακρίβεια ταξινόμησης που επιτυγχάνεται από τους συνδυασμούς των ειδικών μπορεί να είναι αρκετά καλύτερη από εκείνη των ατομικών ταξινομητών. Τέλος, καταλήγουμε ότι διαφορετικά συστήματα επιλογής οδηγούν σε ποικίλα μεγέθη γονιδιακής υπογραφής, με διαφορές στα επιλεγμένα γονίδια. Ωστόσο, όταν περιγράφουμε τα γονίδια κάθε υπογραφής φαίνεται ότι επικαλούνται τις ίδιες βιολογικές διεργασίες, με μικρές πιθανές διαφορές στη σχετική συχνότητα συμμετοχής.

Ο στόχος αυτής της εργασίας έχει τρεις πτυχές. Ο πρώτος μεγάλος στόχος είναι η εύρεση των πιο σημαντικών και άρα πληροφοριακών γονιδίων από το σύνολο δεδομένων που οδηγεί στην επιλογή των κατάλληλων βιοδεικτών (biomarkers), γονιδίων δηλαδή που είναι σχετικά με ένα συγκεκριμένο θέμα ταξινόμησης και που αποτελούν δείκτες μιας συγκεκριμένης βιολογικής κατάστασης. Σκοπός μας είναι να γίνει σαφές ποια γονίδια είναι σημαντικά και ποια μπορούν να παραλειφθούν χωρίς να μειωθεί η απόδοση ταξινόμησης. Για το λόγο αυτό καταλήγουμε σε 9 γονιδιακές υπογραφές, που αντιστοιχούν στους ατομικούς ταξινομητές και τα μείγματα ειδικών που αναφέρθηκαν παραπάνω.

Δεύτερον, εστιάζουμε στην αντιμετώπιση του λεγόμενου προβλήματος «curse of dimensionality» το οποίο εκδηλώνεται στο σύνολο δεδομένων μας, αφού ο αριθμός των γονιδίων είναι πολύ μεγαλύτερος από τον αριθμό των, προς ανάλυση, δειγμάτων και το οποίο εισάγει αστάθεια στην απόδοση των αλγορίθμων πρόβλεψης. Με την προτεινόμενη μεθοδολογία που περιγράφεται παραπάνω, επιχειρούμε μείωση των διαστάσεων του συνόλου δεδομένων με την επιλογή ενός μικρού υποσυνόλου σχετικών γονιδίων για το κάθε μοντέλο, μέσω του οποίου επιτυγχάνεται υψηλότερη ακρίβεια ταξινόμησης σε σχέση με αυτή που σημειώνεται όταν η ταξινόμηση βασίζεται στο αρχικό μεγάλο σύνολο των γονιδίων. Με άλλα λόγια, αν μία γονιδιακή υπογραφή [16], από πολύ διαφορετικά εκφρασμένα γονίδια, μπορεί ευρέως να προσδιοριστεί, τότε με μεγάλη πιθανότητα, μέσω αυτού του συνόλου γονιδίων, θα επιτυγχάνεται καλή προγνωστική ακρίβεια.

Τρίτον, στοχεύουμε στην παροχή εύρωστων και παράλληλα αξιόπιστων μεθόδων αναγνώρισης προτύπων, που συνδυάζουν τους ατομικούς ταξινομητές με συνδυαστικές μεθόδους (μείγματα) για την ακριβέστερη εκτέλεση της ταξινόμησης, αφού δεν έχει βρεθεί ταξινομητής που να αποδίδει καλά σε όλα τα σύνολα δεδομένων.

Παράλληλα με όλα τα παραπάνω, η διαλεύκανση σημαντικών βιολογικών λειτουργιών και διαδικασιών μέσα σε μεγάλα σύνολα δεδομένων είναι ένας δύσκολος και σημαντικός στόχος. Σε αυτή την εργασία, στοχεύουμε στην ανάδειξη της σημαντικής βιολογικής σχέσης ανάμεσα στα γονίδια διαφορετικών γονιδιακών υπογραφών, προσδιορίζοντας τις κοινές βιολογικές διεργασίες που εμπλέκονται σε αυτές. Η βιολογική εξερεύνηση των αποτελεσμάτων της εργασίας που σχετίζονται με το γενότυπο και η βιολογική υποστήριξη των μέτρων ομοιότητας των συναρτήσεων που χρησιμοποιούνται στις στατιστικές προσεγγίσεις, μειώνουν την αβεβαιότητα στη λήψη αποφάσεων που περιλαμβάνουν γονιδιακά δεδομένα.

1.6 Περιγραφή του Συνόλου Δεδομένων

Τα δεδομένα που αναλύονται [7] στο σύνολο δεδομένων αυτής της εργασίας, συλλέγονται από διάστικτες μικροσυστοιχίες του DNA (*DNA microarrays*). Η γονιδιακή έκφραση του ασκομύκητα *Saccharomyces cerevisiae* μελετήθηκε κατά τη διάρκεια της μετατόπισης «δύο φάσεων ανάπτυξης», του κύκλου της μιτωτικής κυτταρικής διαίρεσης, της σπορ(ι)ογένεσης ή σπορίωσης και των θερμοκρασιακών και αναγωγικών σοκ, χρησιμοποιώντας μικροσυστοιχίες, οι οποίες ουσιαστικά περιέχουν κάθε ανοικτό πλαίσιο ανάγνωσης (open reading frame, ORF) από αυτόν τον πλήρως αποκωδικοποιημένο οργανισμό. Σήμερα γνωρίζουμε, ότι η πλήρης αλληλουχία του γονιδιώματος του σακχαρομύκητα περιλαμβάνει περίπου 6000 ORFs, δηλαδή τμήματα γονιδίων που κωδικεύουν πρωτεΐνες. Το γονιδίωμα του *Saccharomyces cerevisiae* (12,8 Mb (mega base))[1] είναι περίπου 200 φορές μικρότερο από το ανθρώπινο γονιδίωμα και εντοπίζεται σε 16 χρωμοσώματα.

Σε αυτό το σύνολο δεδομένων, οι γραμμές αναπαριστούν όλα τα γονίδια από τα οποία έχουν συγκεντρωθεί τα δεδομένα, οι στήλες αντιπροσωπεύουν ατομικές συστοιχίες πειραμάτων (για παράδειγμα μεμονωμένες χρονικές στιγμές ή συνθήκες) και κάθε κελί αποτελεί τη μετρούμενη ένταση φθορισμού του λόγου Cy5/Cy3 (log(Cy5/Cy3)) στο αντίστοιχο στοιχείο-στόχο στον κατάλληλο πίνακα. Σε αυτό το σημείο θα πρέπει να αναφερθεί ότι οι Cy5 και οι Cy3 είναι φθορίζουσες ουσίες που αντιστοιχούν σε κόκκινο και πράσινο χρώμα και οι οποίες χρησιμοποιούνται για την ιχνηθέτηση των μορίων DNA (κυκλικού DNA, cDNA) του δείγματος ελέγχου και του δείγματος αναφοράς, αντίστοιχα, πριν την ανάμειξη και την υβριδοποίηση[2] των δύο δειγμάτων με το DNA στόχο του πλακιδίου της μικροσυστοιχίας [17]. Όλες οι τιμές των λόγων έχουν μετασχηματιστεί από το λογάριθμο με βάση το 2 (για ευκολία), για την αντιμετώπιση επαγωγών ή καταστολών πανομοιότυπου μεγέθους ως αριθμητικά ίσα αλλά με αντίθετο πρόσημο.

Τα δεδομένα, από ξεχωριστές χρονικές στιγμές της γονιδιακής έκφρασης του ζυμομύκητα, συνδυάζονται και ομαδοποιούνται. Τα δεδομένα αντιστοιχούν σε

[1] Mb= mega base pairs= 1.000.000 bp= unit of length of DNA fragments

[2] Υβριδοποίηση καλείται η διαδικασία, κατά την οποία σχηματίζεται ένα μακρομοριακό υβρίδιο από τον τεχνητό ανασυνδυασμό των υπομονάδων, όπως των πολυνουκλεοτιδικών αλυσίδων DNA ή/και RNA, σχηματίζοντας ζεύγη βάσεων Watson-Crick.

συνολικά 79 χρονικές στιγμές κατά τη διάρκεια των 8 ακόλουθων διεργασιών: τον κύκλο της κυτταρικής διαίρεσης μετά από συγχρονισμό με διακοπή (σταμάτημα) του παράγοντα άλφα (Alpha Factor Arrest, ALPH με 18 χρονικές στιγμές) και με το φυγόκεντρο διαχωρισμό επίπλευσης (Centrifugal Elutriation, ELU με 14 χρονικές στιγμές) και με το θερμοευαίσθητο μετάλλαγμα cdc15 (Temperature Sensitive cdc15 mutant, CDC15 με 15 χρονικές στιγμές), τη σπορ(ι)ογένεση ή σπορίωση (Sporulation, SPO με 7 χρονικές στιγμές και 4 επιπρόσθετα δείγματα), τα σοκ από υψηλή θερμοκρασία (Shock by High Temperature, HT με 6 χρονικές στιγμές), από αναγωγικά μέσα (Reducing Agents, D με 4 χρονικές στιγμές) και από χαμηλή θερμοκρασία (Low Temperature, C με 4 χρονικές στιγμές) και τη μετατόπιση «δύο φάσεων ανάπτυξης» (Diauxic Shift, DX με 7 χρονικές στιγμές). Για όλα τα 2467 γονίδια που μελετήθηκαν στην παρούσα εργασία είναι διαθέσιμη η λειτουργική περιγραφή τους στη γονιδιακή βάση του σακχαρομύκητα, η οποία ονομάζεται *Saccharomyces* Genome Database (SGD) [18].

Από αυτές τις 8 διεργασίες που προαναφέρθηκαν, απομονώθηκαν και χρησιμοποιήθηκαν, στα πλαίσια αυτής της εργασίας, όπως αναλυτικά περιγράφεται στην ενότητα 4.1, τα θερμικά και αναγωγικά σοκ, καθώς και ο κύκλος της κυτταρικής διαίρεσης μετά το συγχρονισμό με το φυγόκεντρο διαχωρισμό επίπλευσης. Η τελευταία αποτελεί διεργασία στην οποία [19] τα κύτταρα σε μία καλλιέργεια χωρίζονται με βάση το μέγεθος, τη μάζα και το σχήμα.

Οι κυτταρικοί οργανισμοί [20] απαιτούν συγκεκριμένες εσωτερικές συνθήκες για τη βέλτιστη ανάπτυξη και λειτουργία τους. Ενώ οι πολυκύτταροι οργανισμοί μπορούν να χρησιμοποιούν εξειδικευμένα όργανα και ιστούς για να παρέχουν ένα σχετικά σταθερό και ομογενοποιημένο περιβάλλον, οι μονοκύτταροι οργανισμοί, όπως ο ζυμομύκητας *Saccharomyces cerevisiae*, έχουν αναπτύξει αυτόνομους μηχανισμούς για να προσαρμόζονται στις δραστικές περιβαλλοντικές αλλαγές. Οι ζυμομύκητες συχνά αντέχουν σε διακυμάνσεις στον τύπο της θερμοκρασίας και στη μεταβλητή παρουσία βλαβερών παραγόντων όπως οι τοξικές χημικές ουσίες. Έτσι, όταν οι περιβαλλοντικές συνθήκες αλλάξουν απότομα, το κύτταρο πρέπει γρήγορα να ρυθμίσει το πρόγραμμα της γονιδιακής του έκφρασης για να προσαρμοστεί στις νέες συνθήκες. Γονίδια των οποίων τα μετάγραφα ανταποκρίνονται σε μία ποικιλία από πιέσεις εμπλέκονται σε μία γενική στρεσογόνα απόκριση για το ζυμομύκητα. Άλλες αποκρίσεις γονιδιακής έκφρασης εμφανίζονται να είναι συγκεκριμένες σε πιο ιδιαίτερες περιβαλλοντικές συνθήκες.

Η διαίρεση του κυτταρικού κύκλου [21] είναι ένα σύνθετο αυτορυθμιζόμενο πρόγραμμα, έτσι ώστε πολλά γονίδια που εμπλέκονται στα διάφορα στάδια του κυτταρικού κύκλου να ελέγχονται επίσης από αυτόν. Ειδικότερα, πολλά γονίδια, που διαθέτουν κύτταρα ρυθμιζόμενα από τον κυτταρικό κύκλο, εμπλέκονται σε διεργασίες που συμβαίνουν μόνο μία φορά στον κυτταρικό κύκλο. Τέτοιες διεργασίες περιλαμβάνουν τη σύνθεση του DNA, την εκβλάστηση (budding) και την κυτταροκίνηση. Επιπρόσθετα, πολλά από αυτά τα γονίδια εμπλέκονται στον έλεγχο του κυτταρικού κύκλου. Ο συγχρονισμός του κυτταρικού κύκλου [19] με το φυγόκεντρο διαχωρισμό επίπλευσης βασίζεται στη συλλογή κυττάρων με παρόμοιο μέγεθος, μάζα και όγκο.

1.7 Δομή της Εργασίας

Για την εκτέλεση γονιδιακής ταξινόμησης στο σύνολο δεδομένων του ζυμομύκητα *Saccharomyces cerevisiae,* υλοποιήσαμε μία σειρά από ατομικούς ταξινομητές βασισμένους σε πυρήνα, καθώς και μείγματα αυτών των ειδικών. Επιπλέον, πραγματοποιήσαμε επιλογή υποσυνόλου γονιδίων, εξάγοντας γονιδιακές υπογραφές για κάθε αλγόριθμο ταξινόμησης. Κατά συνέπεια, τα κεφάλαια που ακολουθούν στην παρούσα εργασία οργανώνονται ως εξής:

Στο κεφάλαιο 2 παρουσιάζεται το θεωρητικό υπόβαθρο της εργασίας, αναλύοντας τη βασική θεωρία γύρω από την αναγνώριση προτύπων, το cross validation, την ανάλυση της ταξινόμησης, τη χρησιμότητα των συναρτήσεων πυρήνα και την επιλογή χαρακτηριστικών, όροι οι οποίοι αποτελούν την ευρύτερη «οικογένεια» εννοιών στις οποίες εντάσσεται η παρούσα εργασία.

Στο κεφάλαιο 3 γίνεται λεπτομερής αναφορά στους αλγορίθμους με τους οποίους υλοποιούνται τα μοντέλα μάθησης και πρόβλεψης, που χρησιμοποιούνται για την εκτέλεση της ταξινόμησης. Παρουσιάζονται τόσο οι τέσσερις ατομικοί ταξινομητές, όσο και οι τέσσερις συνδυασμοί ειδικών μεθόδων καθώς και τα μέτρα για την αξιολόγηση της απόδοσής τους. Επιπρόσθετα, εξηγείται το μέτρο για την εκτίμηση της ευθυγράμμισης ενός πυρήνα με το σύνολο δεδομένων που διαθέτουμε και η μέθοδος για την αναδρομική εξάλειψη χαρακτηριστικών που συμβάλλει στην επιλογή γονιδίων.

Στο κεφάλαιο 4 περιλαμβάνεται η προτεινόμενη μεθοδολογία επιλογής χαρακτηριστικών για την επιλογή των πιο σημαντικών γονιδίων, καθώς και το σύνολο των αποτελεσμάτων της εργασίας. Παρουσιάζονται αναλυτικά όλα τα βήματα που ακολουθήθηκαν για την εφαρμογή του αρχικού επιπέδου επεξεργασίας των γονιδίων, καθώς και του πρώτου και δεύτερου επιπέδου επιλογής τους. Στο τέλος αυτού του κεφαλαίου, αναφέρεται η στατιστική και βιολογική ερμηνεία όλων των αποτελεσμάτων και γίνεται συγκριτική μελέτη πάνω σε αυτά.

Στο κεφάλαιο 5 συνοψίζονται, εν συντομία, τα ευρήματα του κεφαλαίου 4. Επίσης αναφέρονται κάποιες καινούριες ιδέες για επιπλέον έρευνα.

ΚΕΦΑΛΑΙΟ 2: Θεωρητικό Υπόβαθρο

Σε αυτό το κεφάλαιο παρουσιάζεται ο ορισμός της αναγνώρισης προτύπων, καθώς και οι κατηγορίες στις οποίες διακρίνεται. Εξηγούνται συνοπτικά οι κυριότερες μέθοδοι που αφορούν σε αυτήν, δίνοντας μεγαλύτερη έμφαση στην ταξινόμηση και στη διάκρισή της σε γραμμικούς και μη γραμμικούς ταξινομητές.

Στη συνέχεια, αναλύεται η έννοια του cross validation, μίας τεχνικής που διαχωρίζει ένα σύνολο δεδομένων σε σύνολο εκπαίδευσης, το οποίο χρησιμοποιείται κατά τη φάση του σχεδιασμού του ταξινομητή και σε σύνολο ελέγχου, το οποίο χρησιμοποιείται για την αξιολόγηση της απόδοσης του.

Επίσης, εισάγεται η έννοια του πυρήνα για να εξηγηθεί η απεικόνιση των δεδομένων εισόδου σε ένα χώρο χαρακτηριστικών υψηλότερων διαστάσεων, καθώς και το kernel trick.

Τέλος, παρατίθεται γιατί είναι σημαντική η διαδικασία της επιλογής χαρακτηριστικών καθώς και οι μέθοδοι filter, wrapper και embedded με τις οποίες αυτή εφαρμόζεται.

2.1 Αναγνώριση Προτύπων

Στη μηχανική μάθηση (*machine learning*), η αναγνώριση προτύπων (*pattern recognition*) είναι η εκχώρηση μιας ετικέτας (*label*) σε μια δεδομένη τιμή εισόδου. Γενικά, κατηγοριοποιείται σύμφωνα με τον τύπο της διαδικασίας μάθησης που χρησιμοποιείται για να παραχθεί η τιμή εξόδου [22].

- ✓ <u>Η μάθηση με επίβλεψη (*supervised learning*)</u> υποθέτει ότι παρέχεται ένα σύνολο δεδομένων εκπαίδευσης (*training set*), το οποίο αποτελείται από ένα σύνολο περιπτώσεων (*instances*) ή αλλιώς δειγμάτων (*samples*) που φέρουν κατάλληλη ετικέτα με τη σωστή έξοδο. Η διαδικασία μάθησης δημιουργεί ένα μοντέλο που προσπαθεί να ικανοποιήσει δύο πολλές φορές αντικρουόμενους στόχους: Να εκπαιδευτεί όσο καλύτερα γίνεται πάνω στα δεδομένα εκπαίδευσης και να μπορέσει να γενικεύσει μετά όσο καλύτερα γίνεται στα νέα δεδομένα.

Μία μεγάλη οικογένεια από θεωρίες στατιστικής μάθησης έχουν μελετηθεί για να επιτύχουν την ελαχιστοποίηση του ρίσκου και τη μεγιστοποίηση της γενίκευσης στα θέματα της μάθησης. Αυτές οι θεωρίες έχουν οδηγήσει στη δημιουργία πολλών τύπων αλγορίθμων, για διάφορες εφαρμογές της επιβλεπόμενης μάθησης. Ανάμεσά τους, το supervised kernel-machine learning είναι η ανώτερη μεθοδολογία στις εφαρμογές του πραγματικού κόσμου. Πολλοί αλγόριθμοι όπως τα Support Vector Machines (SVM) έχουν επιδείξει εξαιρετική απόδοση σε ένα εύρος εφαρμογών, ειδικά σε θέματα ταξινόμησης.

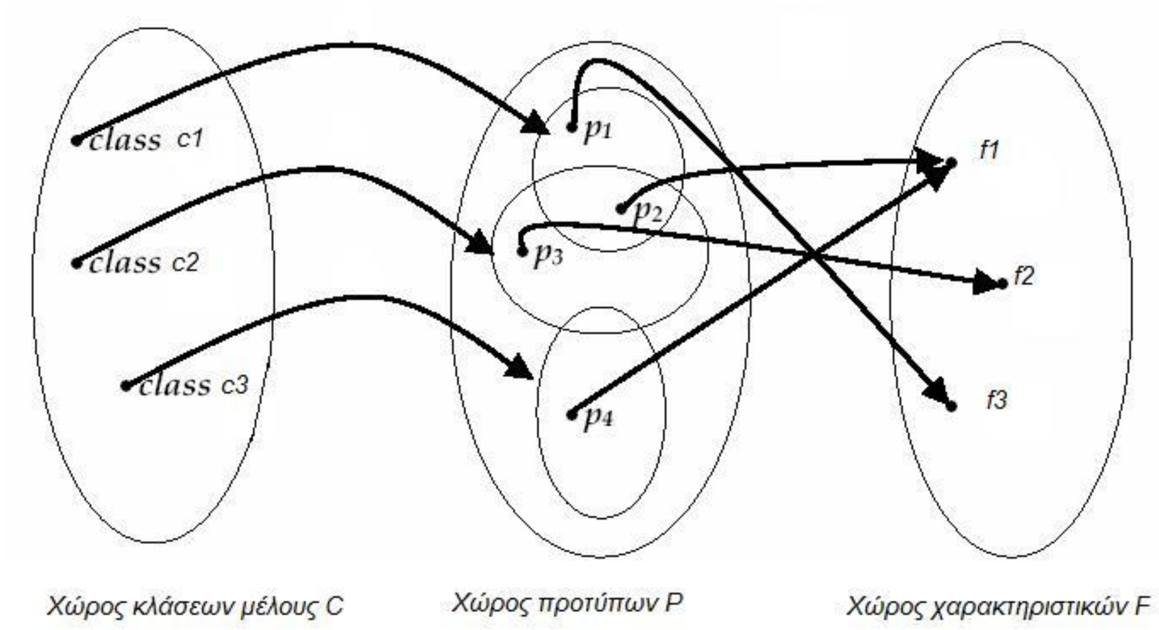
- ✓ <u>Η μάθηση χωρίς επίβλεψη (*unsupervised learning*)</u> από την άλλη πλευρά, λαμβάνει δεδομένα εκπαίδευσης τα οποία δεν διαθέτουν ετικέτες, και προσπαθεί να ανακαλύψει εγγενή πρότυπα ανάμεσα στα δεδομένα, τα οποία να μπορούν μετά να χρησιμοποιηθούν για να καθορίσουν τη σωστή τιμή εξόδου για τα νέα δείγματα δεδομένων. Σε αυτή την περίπτωση μάθησης, μπορεί να μην υπάρχουν καθόλου δεδομένα εκπαίδευσης, δηλαδή τα δεδομένα που πρόκειται να κατηγοριοποιηθούν να αποτελούν το σύνολο εκπαίδευσης.

  Σε πολλές εφαρμογές του πραγματικού κόσμου, όπου η εκχώρηση ετικετών στα δεδομένα είναι ακριβή διαδικασία, οι τεχνικές της μη επιβλεπόμενης μάθησης χρησιμοποιούνται συχνά για να ανακαλύψουν άγνωστη γνώση από μία μεγάλη ποσότητα δεδομένων χωρίς ετικέτες. Μία από τις πιο δημοφιλείς μεθοδολογίες, ανάμεσα στις διάφορες τεχνικές για τη μάθηση χωρίς επίβλεψη, είναι η ομαδοποίηση των δεδομένων (*data clustering*), η οποία έχει ευρέως χρησιμοποιηθεί στις εφαρμογές για την εξόρυξη δεδομένων (*data mining*).

- ✓ <u>Η μάθηση με ημί-επίβλεψη (*semi-supervised learning*)</u> συνδυάζει τις δύο παραπάνω περιπτώσεις μάθησης. Ένα πρόβλημα στην επιβλεπόμενη μάθηση είναι η περίπτωση της μάθησης όταν υπάρχουν ανεπαρκή δείγματα εκπαίδευσης. Για να επωφεληθεί και από τα δεδομένα με ετικέτα (*labeled data*) αλλά και από αυτά που δε διαθέτουν (*unlabeled data*), η μάθηση με ημί-επίβλεψη έχει πρόσφατα προταθεί για να αντιμετωπίσει την πρόκληση της μάθησης χρησιμοποιώντας ένα μικρό αριθμό δειγμάτων εκπαίδευσης με ετικέτα και ένα μεγάλο αριθμό δειγμάτων χωρίς ετικέτα. Έχει φανεί ότι είναι μία πολλά υποσχόμενη προσέγγιση, παρέχοντας βελτιωμένη απόδοση σε σύγκριση με τις παραδοσιακές προσεγγίσεις της επιβλεπόμενης μάθησης, όταν ο αριθμός των δειγμάτων εκπαίδευσης που προσφέρεται είναι περιορισμένος.

Συνοψίζοντας κάποιες βασικές έννοιες [23], προκύπτουν οι εξής ορισμοί: Οι κλάσεις (*classes*) είναι καταστάσεις της «φύσης» ή κατηγορίες αντικειμένων που σχετίζονται με έννοιες ή με πρωτότυπα. Τα πρότυπα (*patterns)* είναι οι «φυσικές» αναπαραστάσεις των αντικειμένων και συχνά γίνεται αναφορά σε αυτά με τους όρους αντικείμενα, περιπτώσεις ή δείγματα. Πιο συγκεκριμένα, ένα πρότυπο είναι ένα σύνολο από μετρήσεις ή παρατηρήσεις που χαρακτηρίζουν μία κατάσταση και αναπαρίσταται συνήθως με ένα διάνυσμα ή με έναν πίνακα. Τα χαρακτηριστικά (*features*) είναι μετρήσεις ή ιδιότητες που προέρχονται από τα πρότυπα και που είναι χρήσιμα για τον χαρακτηρισμό τους.

Η αναγνώριση προτύπων μπορεί να χαρακτηριστεί ως μία διαδικασία αναγωγής πληροφορίας (*information reduction*), ή αποτύπωσης πληροφορίας (*information mapping*), ή χαρακτηρισμού πληροφορίας (*information labeling*), όπως φαίνεται στην εικόνα 5. Κάθε μία από τις κλάσεις, στο χώρο των μελών κλάσεων C, απεικονίζεται σε ένα υποσύνολο προτύπων στο χώρο των προτύπων P. Αυτά τα υποσύνολα μπορεί ωστόσο να επικαλύπτονται, επιτρέποντας σε πρότυπα από διαφορετικές κλάσεις να μοιράζονται χαρακτηριστικά. Επιπλέον, ο χώρος προτύπων P, απεικονίζεται στο χώρο χαρακτηριστικών F, δηλαδή, κάθε διάνυσμα προτύπου περιλαμβάνει χαρακτηριστικά από το χώρο F. Όπως φαίνεται και από την εικόνα 5, το ίδιο χαρακτηριστικό μπορεί να ανήκει σε διαφορετικά υποσύνολα προτύπων, τα οποία μπορούν επίσης να αντιστοιχούν σε διαφορετικές κλάσεις.



Εικόνα 5: Η διαδικασία της αναγνώρισης προτύπων [23].

2.1.1 Εφαρμογές της αναγνώρισης προτύπων

Στο πεδίο της αναγνώρισης προτύπων, ο στόχος είναι να μάθει ένας υπολογιστής, από παραδείγματα, να αναγνωρίζει πρότυπα σε σύνολα δεδομένων. Η διαδικασία της αναγνώρισης μπορεί να εκτελεστεί με ταξινόμηση (*classification*) ή με ομαδοποίηση (*clustering*) των δεδομένων.

Στη μηχανική μάθηση [24] και στην αναγνώριση προτύπων, η ταξινόμηση αναφέρεται σε μία επιβλεπόμενη αλγοριθμική διαδικασία για την εκχώρηση ενός δοθέντος τμήματος δεδομένων εισόδου σε μία κατηγορία από ένα σύνολο δεδομένων κατηγοριών. Ο αλγόριθμος που υλοποιεί την ταξινόμηση ονομάζεται ταξινομητής (*classifier*). Το τμήμα των δεδομένων εισόδου ονομάζεται δείγμα και οι κατηγορίες ονομάζονται κλάσεις. Κάθε δείγμα περιγράφεται από ένα διάνυσμα (*vector*) χαρακτηριστικών, τα οποία όλα μαζί αποτελούν μία περιγραφή όλων των γνωστών χαρακτηριστικών του δείγματος.

Πιο συγκεκριμένα, ένας ταξινομητής εκπαιδεύεται πάνω στα δεδομένα εκπαίδευσης, προσαρμόζοντας τις παραμέτρους του σε αυτά, και τελικά μαθαίνει να αναγνωρίζει συγκεκριμένα πρότυπα. Το αποτέλεσμα της διαδικασίας της ταξινόμησης βασίζεται στο πιο σημαντικό χαρακτηριστικό ενός ταξινομητή που είναι η ικανότητα γενίκευσης. Αυτό αναφέρεται στην παραγωγή μίας λογικής απόφασης για τα νέα δεδομένα τα οποία είναι αθέατα κατά τη διάρκεια της σχεδίασης του μοντέλου του ταξινομητή (εκπαίδευση). Αν ο ταξινομητής σχεδιαστεί αποτελεσματικά, θα είναι ικανός να συνδυάσει σωστά τα χαρακτηριστικά ενός δείγματος για να καθορίσει σε ποια κλάση αυτό ανήκει. Ο πιο εύκολος τρόπος για να μετρήσουμε αυτή την ικανότητα γενίκευσης ενός πλήρους εκπαιδευμένου ταξινομητή είναι να χρησιμοποιήσουμε ένα σύνολο δεδομένων ελέγχου (*test set*) που περιλαμβάνει δεδομένα τα οποία δεν ανήκουν στο σύνολο εκπαίδευσης.

Το αντικείμενο της ομαδοποίησης δεδομένων [25] είναι η οργάνωση των δεδομένων (πρότυπα) σε ουσιαστικές ή χρήσιμες ομάδες (*clusters*), χρησιμοποιώντας κάποιο τύπο μέτρου ομοιότητας, έτσι ώστε τα αντικείμενα μέσα στην ίδια ομάδα να μοιάζουν περισσότερο μεταξύ τους παρά με εκείνα που βρίσκονται σε άλλες ομάδες. Η ομαδοποίηση δεδομένων δε χρησιμοποιεί προγενέστερη πληροφορία σχετικά με την κλάση των δεδομένων, αφού αποτελεί μία μέθοδο μη επιβλεπόμενης ταξινόμησης, και μπορεί να υλοποιηθεί με πολλούς αλγορίθμους οι οποίοι διαφέρουν σημαντικά ως προς την αντίληψη τους για το τι συνιστά μία ομάδα, και με ποιο τρόπο να τις βρουν. Δημοφιλείς έννοιες τέτοιων ομάδων περιλαμβάνουν ομάδες με μικρές αποστάσεις ανάμεσα στα μέλη μιας ομάδας, περιοχές πυκνότητας στο χώρο δεδομένων, διαστήματα ή ειδικότερα στατιστικές κατανομές. Μία από τις τεχνικές ομαδοποίησης είναι το self-organizing map.

2.2 Ανάγκες και Στόχοι της Ταξινόμησης

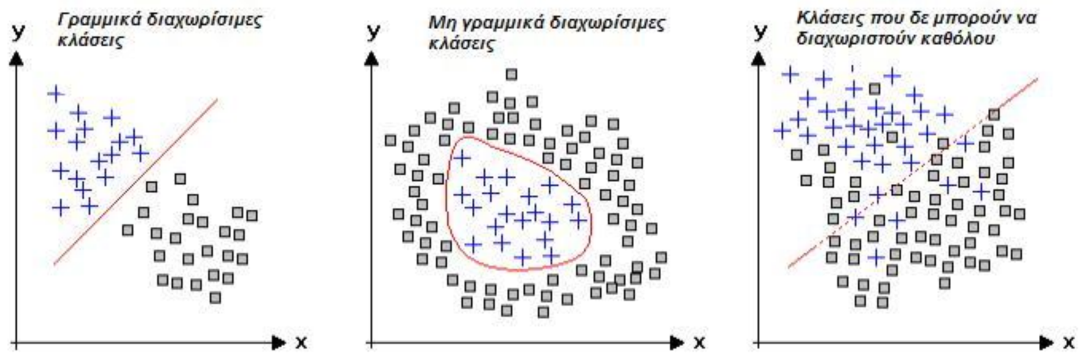
Η ανάλυση της ταξινόμησης είναι ένα από τα πρωταρχικά θέματα στη μηχανική μάθηση και την υπολογιστική βιολογία. Το ζητούμενο της ταξινόμησης [26] είναι να βρεθεί ένας κανόνας ο οποίος, βασισμένος σε εξωτερικές παρατηρήσεις, να μπορεί να εκχωρεί ένα δείγμα σε μία από τις δεδομένες κλάσεις. Η δυαδική ταξινόμηση (*binary classification*) είναι η πιο απλή περίπτωση ταξινόμησης σε δύο κλάσεις, που κατηγοριοποιεί ένα νέο δείγμα σε μία από τις δύο γνωστές κλάσεις ενός προβλήματος, ανάλογα με το αν αυτό διαθέτει μία ιδιότητα ή όχι. Τα χαρακτηριστικά των δειγμάτων έχουν κάποιες τιμές, σύμφωνα με τις οποίες, τα δείγματα βρίσκονται κάπου στο χώρο. Η σωστή ταξινόμηση βασίζεται στην προγενέστερη γνώση που έχει λάβει ο ταξινομητής στη φάση της εκπαίδευσης.

Τα χαρακτηριστικά των αλγορίθμων ταξινόμησης [27] εξαρτώνται από τον τρόπο με τον οποίο καθορίζεται η διαχωριστική γραμμή ανάμεσα στις κλάσεις ενός συνόλου δεδομένων. Επομένως, οι αλγόριθμοι ταξινόμησης επιχειρούν να προσδιορίσουν την καλύτερη διαχωριστική γραμμή ανάμεσα στις κλάσεις. Ένας τρόπος για να οριστεί το υπερεπίπεδο που διαχωρίζει τις δύο κλάσεις είναι μία συνάρτηση απόφασης (*discriminant function*) η οποία είναι 0 πάνω στο επίπεδο, θετική όταν τα δείγματα ανήκουν στην κλάση 1 και αρνητική όταν ανήκουν στην κλάση 2.

2.2.1 Γραμμικοί και μη γραμμικοί ταξινομητές

Η κύρια διάκριση των ταξινομητών [28] συνίσταται σε γραμμικούς (*linear*) και μη γραμμικούς (*non linear*). Ένας γραμμικός ταξινομητής μπορεί να διαχωρίσει δύο κλάσεις μόνο όταν αυτές είναι γραμμικά διαχωρίσιμες (*linearly separable*), δηλαδή υπάρχει ένα υπερεπίπεδο (στο δισδιάστατο χώρο είναι απλά μία γραμμή, στον τρισδιάστατο ένα επίπεδο, και σε περισσότερες διαστάσεις ένα υπερεπίπεδο) που διαχωρίζει τα δείγματα στις δύο κλάσεις ενδιαφέροντος. Τότε ο γραμμικός ταξινομητής βασιζόμενος στην τιμή ενός γραμμικού συνδυασμού των χαρακτηριστικών κάθε δείγματος, αποφασίζει σε ποια από τις δύο κλάσεις θα το εκχωρήσει.

Η αντίθετη περίπτωση συμβαίνει όταν οι κλάσεις είναι μη γραμμικά διαχωρίσιμες (*linearly unseparable*). Σε αυτή την περίπτωση είναι πιθανό να είναι λίγα τα δείγματα που βρίσκονται στη λάθος πλευρά του υπερεπιπέδου, και έτσι το σφάλμα στο να χρησιμοποιήσουμε ένα γραμμικό όριο να είναι μικρό. Ανάλογα με το βαθμό του σφάλματος, ένας γραμμικός ταξινομητής μπορεί ακόμα να είναι προτιμητέος, γιατί το μοντέλο που προκύπτει είναι πιο απλό και έτσι λιγότερο ευαίσθητο στο over fitting (χαμηλή ικανότητα γενίκευσης σε νέα δείγματα). Ωστόσο, κάποιες κλάσεις μπορούν να διαχωριστούν μόνο από ένα μη γραμμικό όριο. Γενικά, στην περίπτωση των μη γραμμικών δεδομένων, ένας μη γραμμικός ταξινομητής εμφανίζει πολύ καλύτερη απόδοση ταξινόμησης από έναν γραμμικό. Στην εικόνα 6 [29], φαίνονται οι τρεις περιπτώσεις διαχωρισμού των κλάσεων.

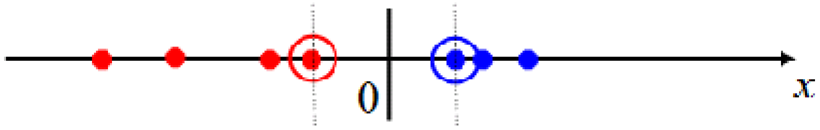


Εικόνα 6: Περιπτώσεις γραμμικά διαχωρίσιμων δεδομένων, μη γραμμικά διαχωρίσιμων δεδομένων και δεδομένων που δε διαχωρίζονται καθόλου [29].

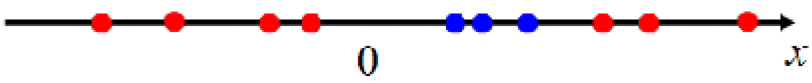
2.3 Μέθοδοι Πυρήνα

Με όσα έχουν παρουσιαστεί ως εδώ, συμπεραίνουμε ότι τα σύνολα δεδομένων που είναι γραμμικά διαχωρίσιμα μπορούμε να τα χειριστούμε καλά. Το πρόβλημα παρουσιάζεται όταν το σύνολο δεδομένων δεν επιτρέπει την ταξινόμηση από έναν γραμμικό ταξινομητή.

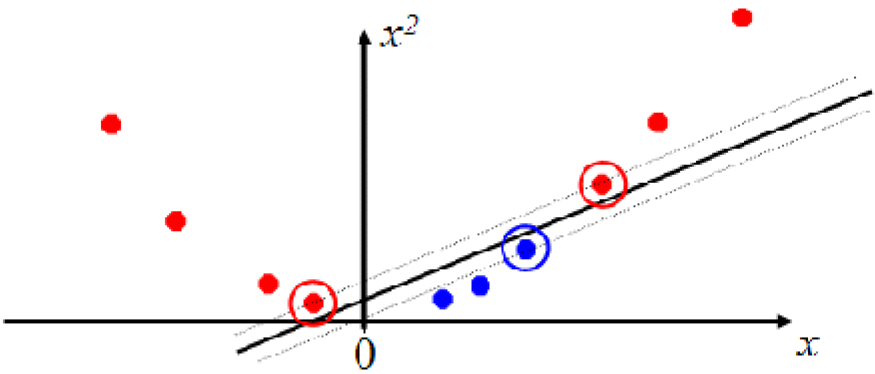
Ας εξετάσουμε την περίπτωση της μίας διάστασης [30]. Το σύνολο δεδομένων της εικόνας 7 μπορεί εύκολα να ταξινομηθεί από έναν γραμμικό ταξινομητή, δε συμβαίνει όμως το ίδιο και με το σύνολο δεδομένων της εικόνας 8. Ωστόσο, πρέπει και σε αυτή την περίπτωση να βρεθεί ένας διαχωρισμός.



Εικόνα 7: Γραμμικά διαχωρίσιμα δεδομένα μίας διάστασης [30].



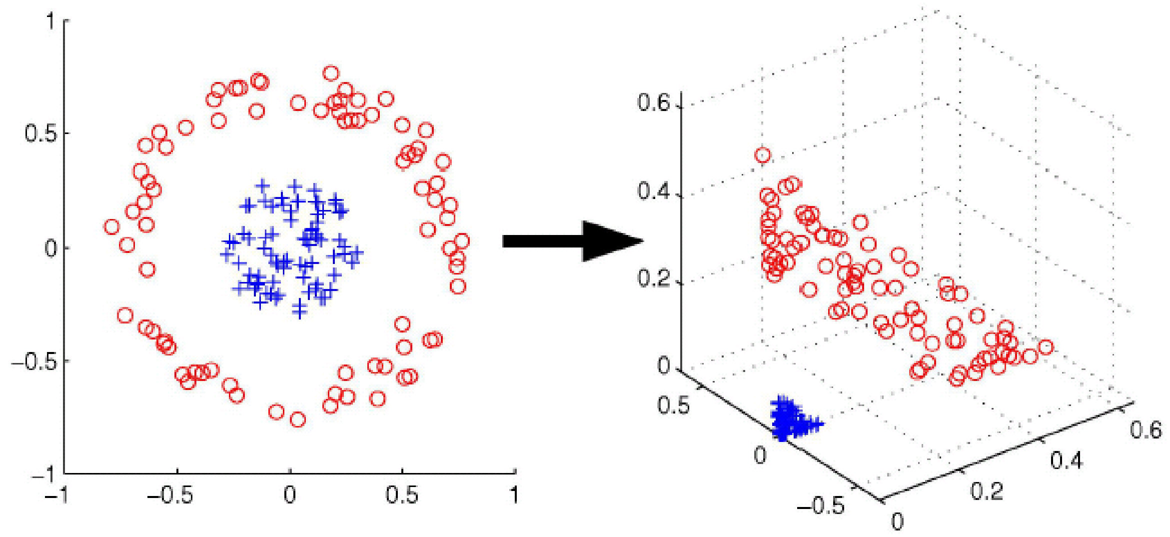
Εικόνα 8: Μη γραμμικά διαχωρίσιμα δεδομένα μίας διάστασης [30].



Εικόνα 9: Προβάλλοντας τα μη γραμμικά διαχωρίσιμα δεδομένα σε χώρο υψηλότερων διαστάσεων μπορούν να μετατραπούν σε γραμμικά διαχωρίσιμα [30].

Ένας τρόπος για να λυθεί αυτό το πρόβλημα είναι να γίνει απεικόνιση (*mapping*) των δεδομένων σε ένα χώρο υψηλότερων διαστάσεων και μετά να χρησιμοποιηθεί ένας γραμμικός ταξινομητής σε αυτό το χώρο των υψηλότερων διαστάσεων. Κάτι τέτοιο φαίνεται στην εικόνα 9 όπου ένας γραμμικός διαχωριστής μπορεί εύκολα να ταξινομήσει τα δεδομένα αν χρησιμοποιηθεί μία τετραγωνική (*quadratic*) συνάρτηση που θα απεικονίσει τα δεδομένα σε δύο διαστάσεις.

Συνοψίζοντας, η γενική ιδέα είναι η απεικόνιση του αρχικού χώρου των χαρακτηριστικών σε έναν άλλο χώρο χαρακτηριστικών με περισσότερες διαστάσεις όπου ένας γραμμικός διαχωρισμός μπορεί να εκτελεστεί όπως φαίνεται στην εικόνα 10 [31]. Φυσικά, κάτι τέτοιο πρέπει να γίνει με τρόπο τέτοιο ώστε να διατηρηθούν οι σχετικές διαστάσεις της συγγένειας ανάμεσα στα δεδομένα για να μπορεί ο ταξινομητής που θα προκύψει να γενικεύσει σωστά στα καινούρια δεδομένα. Έτσι, το γραμμικό πρόβλημα μπορεί να επεκταθεί σε μία μη γραμμική περίπτωση χρησιμοποιώντας τις μεθόδους πυρήνα στις οποίες βασίζονται πολλοί ταξινομητές.



Εικόνα 10: Μετασχηματίζοντας τα δεδομένα, σε ένα χώρο υψηλότερων διαστάσεων, γίνονται γραμμικά διαχωρίσιμα [31].

Μία μέθοδος πυρήνα (*kernel method*) [32], είναι ένας αλγόριθμος που εξαρτάται από τα δεδομένα μόνο μέσω των εσωτερικών γινομένων (*dot-products*). Όταν αυτό συμβαίνει, το εσωτερικό γινόμενο μπορεί να αντικατασταθεί από μία συνάρτηση πυρήνα η οποία υπολογίζει κάποιο εσωτερικό γινόμενο σε κάποιο πιθανό χώρο των χαρακτηριστικών, υψηλότερων διαστάσεων. Αυτό έχει δύο πλεονεκτήματα: Πρώτον, την ικανότητα δημιουργίας μη γραμμικών ορίων απόφασης με τη χρήση μεθόδων που είναι σχεδιασμένες για γραμμικούς ταξινομητές. Δεύτερον, η χρήση των συναρτήσεων πυρήνα επιτρέπει την εφαρμογή ενός ταξινομητή στα δεδομένα που δεν έχει προφανή αναπαράσταση σε διανυσματικό χώρο σταθερών διαστάσεων. Μία από τις πιο διαδεδομένες μεθόδους βασισμένες σε πυρήνα είναι τα support vector machines, που αναλύονται στην ενότητα 3.1 του κεφαλαίου 3.

2.3.1 Το Kernel trick : από μία μη γραμμική σε μία γραμμική λύση

Σε πολλές εφαρμογές [33], ένας μη γραμμικός ταξινομητής παρέχει καλύτερη απόδοση. Ωστόσο, οι γραμμικοί ταξινομητές έχουν πλεονεκτήματα, με ένα από αυτά να είναι ότι συχνά διαθέτουν απλούς αλγορίθμους εκπαίδευσης που ταιριάζουν καλά στον αριθμό των δειγμάτων. Έτσι, προκύπτει η ακόλουθη ερώτηση [32]: Μπορούν οι μηχανισμοί των γραμμικών ταξινομητών να επεκταθούν ώστε να δημιουργούν μη γραμμικά όρια απόφασης; Επιπλέον, μπορούμε να χειριστούμε τομείς όπως οι πρωτεϊνικές ακολουθίες ή δομές όπου η αναπαράσταση σε ένα διανυσματικό χώρο σταθερών διαστάσεων δεν είναι διαθέσιμη;

Ο πιο απλοϊκός τρόπος [34] δημιουργίας ενός μη γραμμικού ταξινομητή από έναν γραμμικό είναι η απεικόνιση των δεδομένων από το χώρο εισόδου X (*input space*) σε ένα χώρο χαρακτηριστικών F (*feature space*) χρησιμοποιώντας μία μη γραμμική συνάρτηση:

$$\varphi : X \to F$$

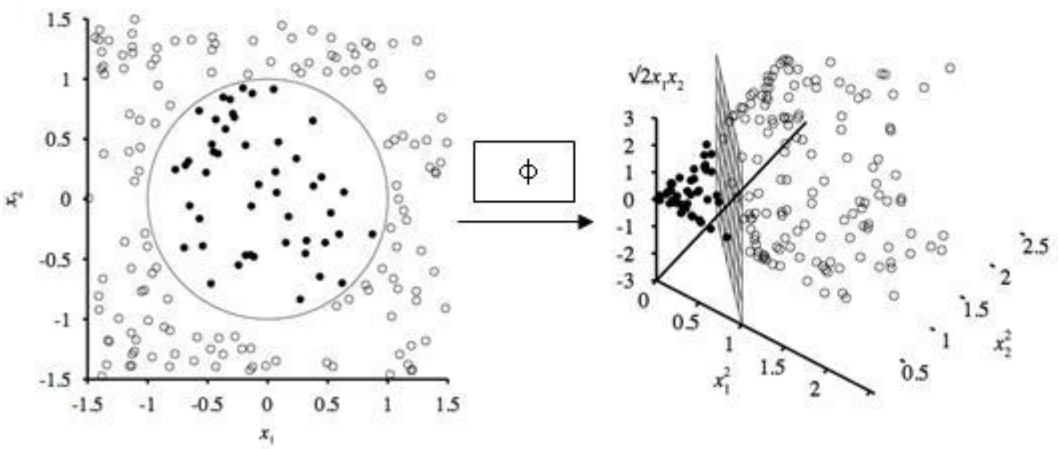
$$x \to \varphi(x)$$

Στην περίπτωση ενός χώρου εισόδου δύο διαστάσεων η απεικόνιση γίνεται:

$$\varphi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) \to \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1\, x_2 \end{bmatrix}$$

Η προσέγγιση του ρητού υπολογισμού [32] των μη γραμμικών χαρακτηριστικών δεν ταιριάζει καλά με τον αριθμό των χαρακτηριστικών του χώρου εισόδου: όταν εφαρμόσουμε την παραπάνω απεικόνιση, ο αριθμός των διαστάσεων του χώρου χαρακτηριστικών F είναι τετραγωνικός αυτού στον χώρου εισόδου, όπως δείχνει η εικόνα 11. Αυτό έχει ως αποτέλεσμα μια τετραγωνική αύξηση της χρήσης μνήμης για την αποθήκευση των χαρακτηριστικών και μια τετραγωνική αύξηση του χρόνου που απαιτείται για τον υπολογισμό της συνάρτησης απόφασης του ταξινομητή. Αυτή η τετραγωνική πολυπλοκότητα είναι εφικτή για δεδομένα λίγων διαστάσεων. Αλλά στην περίπτωση που χειριζόμαστε δεδομένα γονιδιακής έκφρασης που έχουν χιλιάδες διαστάσεις, η τετραγωνική πολυπλοκότητα στον αριθμό των διαστάσεων δεν είναι αποδεκτή. Οι μέθοδοι πυρήνα λύνουν αυτό το πρόβλημα αποφεύγοντας το βήμα της ρητής απεικόνισης των δεδομένων σε ένα χώρο υψηλότερων διαστάσεων.

Οι συναρτήσεις πυρήνα [34] δεν απεικονίζουν στην πραγματικότητα τα δεδομένα, αλλά εκτελούν μία έμμεση απεικόνιση – το kernel trick – των μεταβλητών στο χώρο



Εικόνα 11: Παράδειγμα ταξινόμησης δύο διαστάσεων. Στο χώρο εισόδου (αριστερό σχήμα) μόνο ένα μη γραμμικό κυκλικό όριο μπορεί να εφαρμοστεί. Στο χώρο των χαρακτηριστικών (δεξί σχήμα), χρησιμοποιώντας μονώνυμα, μπορεί να βρεθεί ένα γραμμικό διαχωριστικό υπερεπίπεδο [35].

των χαρακτηριστικών μέσω του εσωτερικού γινομένου ανάμεσα σε κάθε ζευγάρι σημείων. Θεωρείται ότι τα απεικονισμένα δεδομένα στο χώρο των χαρακτηριστικών είναι γραμμικά διαχωρίσιμα ή τουλάχιστον κατέχουν έναν καλύτερο γραμμικό διαχωρισμό από ότι στο χώρο εισόδου. Ωστόσο, ο διαχωρισμός των δεδομένων στο χώρο των χαρακτηριστικών μπορεί να γίνει ακόμα χειρότερος αν εφαρμοστεί ένας ακατάλληλος πυρήνας. Η αλλαγή από έναν γραμμικό ταξινομητή σε έναν μη γραμμικό γίνεται αντικαθιστώντας το εσωτερικό γινόμενο με μία συνάρτηση πυρήνα. Αυτό δεν βλάπτει τα πλεονεκτήματα των ταξινομητών και επιτρέπει να διατηρήσουν τις ιδιότητες τους.

Οι μέθοδοι πυρήνα [26] μετασχηματίζουν τα δεδομένα σε ένα χώρο χαρακτηριστικών F, υψηλότερων διαστάσεων που συνήθως διαθέτει πάρα πολλές διαστάσεις. Το λεγόμενο φαινόμενο της κατάρας των διαστάσεων (*curse of dimensionality*) ισχυρίζεται ότι η δυσκολία σε ένα πρόβλημα εκτίμησης αυξάνεται δραστικά με τις D διαστάσεις του χώρου, αφού κατ’ αρχήν ως συνάρτηση του D, χρειαζόμαστε εκθετικά πολλά πρότυπα για να δειγματοληπτήσουμε το χώρο κατάλληλα. Αυτή η πολύ γνωστή κατάσταση περικλείει κάποιες αμφιβολίες για το αν είναι καλή ιδέα να μεταβούμε σε ένα χώρο χαρακτηριστικών υψηλών διαστάσεων, για την εκπαίδευση του ταξινομητή.

Ωστόσο, η στατιστική μηχανική μάθηση ισχυρίζεται ότι το αντίθετο μπορεί να είναι σωστό: μαθαίνοντας στο χώρο χαρακτηριστικών μπορεί να είναι πιο απλό αν χρησιμοποιήσουμε μία χαμηλή πολυπλοκότητα (γραμμικοί ταξινομητές). Δηλαδή, δεν είναι η διαστατικότητα αλλά η πολυπλοκότητα της συνάρτησης κλάσης που μετράει.

Επιπλέον, έχει αποδειχθεί ότι [36], η ικανότητα γενίκευσης εξαρτάται από τη γεωμετρία των χαρακτηριστικών των δεδομένων εκπαίδευσης και όχι από τη διάσταση του χώρου εισόδου. Οπότε, αν αυτά τα χαρακτηριστικά έχουν επιλεγεί σωστά, το αναμενόμενο σφάλμα γενίκευσης μπορεί να είναι μικρό ακόμα και αν χώρος χαρακτηριστικών έχει πάρα πολύ μεγάλο αριθμό διαστάσεων.

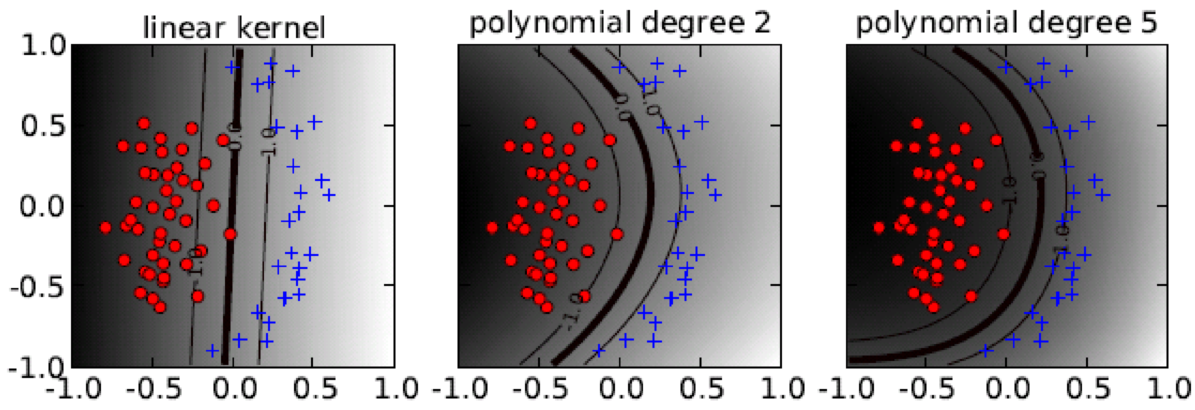
2.3.2 Εφαρμογή διαφορετικών πυρήνων

Το kernel matrix [34], επίσης γνωστό και ως Gram matrix, είναι ένας θετικά ορισμένος πίνακας που ορίζεται ως:

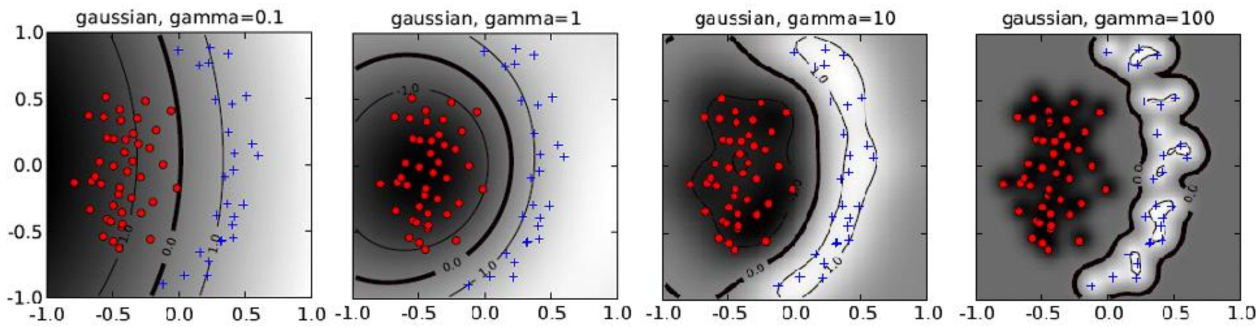
$$K = \left\{ \left(\varphi(x_i) \cdot \varphi\left(x_j\right) \right) \right\}$$

Κάποιοι από τους πιο κοινούς πυρήνες [32] που χρησιμοποιούνται ο πολυωνυμικός βαθμού p, η συνάρτηση γκαουσιανής ακτινικής βάσης (Gaussian Radial Basis Function), ο τετραγωνικός και ο σιγμοειδής. Στην εικόνα 12, αναπαρίσταται ο γκαουσιανός πυρήνας πρώτου, δευτέρου και τρίτου βαθμού, ενώ στην εικόνα 13 φαίνεται για τον ίδιο πυρήνα, η επίδραση της παραμέτρου γ, για τις τιμές 0,1, 1, 10 και 100, που αυτή λαμβάνει.

Ένα συχνό ερώτημα που τίθεται είναι ποιος πυρήνας πρέπει να επιλεγεί για να χρησιμοποιηθεί στα δεδομένα που έχουμε κάθε φορά στη διάθεση μας. Υπάρχουν πολλές απαντήσεις σε αυτή την ερώτηση. Η πρώτη λέει ότι αυτή η επιλογή εξαρτάται, όπως στις περισσότερες ερωτήσεις στη μηχανική μάθηση, από τα δεδομένα οπότε αρκετοί πυρήνες πρέπει να εφαρμοστούν. Αυτό σημαίνει ότι συνήθως ακολουθούμε την εξής διαδικασία: Αρχικά, εφαρμόζουμε έναν γραμμικό πυρήνα, και μετά ελέγχουμε αν μπορούμε να βελτιώσουμε την απόδοση εφαρμόζοντας έναν μη γραμμικό πυρήνα. Ο γραμμικός πυρήνας αποτελεί μία χρήσιμη αρχή καθώς σε πολλές εφαρμογές βιοπληροφορικής παρέχει τα καλύτερα αποτελέσματα. Η ευελιξία του γκαουσιανού και πολυωνυμικού πυρήνα, με τον πρώτο να υπερτερεί σε απόδοση και χρόνο σύγκλισης του δεύτερου, συχνά οδηγεί σε overfitting για δεδομένα υψηλών διαστάσεων με μικρό αριθμό δειγμάτων όπως είναι τα microarray σύνολα δεδομένων.



Εικόνα 12: Η επίδραση του βαθμού ενός πολυωνυμικού kernel. Ο πολυωνυμικός kernel βαθμού 1 οδηγεί σε έναν γραμμικό διαχωρισμό. Υψηλότερου βαθμού πολυωνυμικοί kernels επιτρέπουν ένα πιο ευέλικτο όριο απόφασης [34].



Εικόνα 13: Η επίδραση της παραμέτρου (γ) του Gaussian kernel. Για μικρές τιμές του γ (αριστερά) το όριο απόφασης είναι σχεδόν γραμμικό. Καθώς το γ αυξάνεται η ευελιξία του ορίου απόφασης αυξάνεται. Μεγάλες τιμές του γ οδηγούν σε overfitting (οι 2 δεξιές εικόνες) [34].

2.4 Αξιολόγηση Αποτελεσμάτων μέσω Cross Validation

Το cross-validation είναι μία τεχνική [37] για την αξιολόγηση του τρόπου με τον οποίο τα αποτελέσματα μιας στατιστικής ανάλυσης θα γενικεύσουν σε ένα ανεξάρτητο σύνολο δεδομένων, δηλαδή παρέχει μία εκτίμηση για το πόσο καλά έχει εκπαιδευτεί το μοντέλο. Χρησιμοποιείται κυρίως σε περιπτώσεις όπου στόχος είναι η πρόβλεψη καθώς και η εκτίμηση της ακρίβειας με την οποία ένα προγνωστικό μοντέλο θα εκτελεστεί στην πράξη. Ένας κύκλος του cross-validation περιλαμβάνει το διαμερισμό των δειγμάτων δεδομένων σε συμπληρωματικά υποσύνολα, την εκτέλεση της ανάλυσης στο ένα υποσύνολο που ονομάζεται σύνολο εκπαίδευσης και την επικύρωση της ανάλυσης στο άλλο υποσύνολο που ονομάζεται σύνολο επικύρωσης ή σύνολο δοκιμών. Για τη μείωση της μεταβλητότητας, εκτελούνται πολλαπλοί κύκλοι του cross validation χρησιμοποιώντας διαφορετικούς διαμερισμούς και υπολογίζοντας το μέσο όρο των αποτελεσμάτων της επικύρωσης επί των κύκλων.

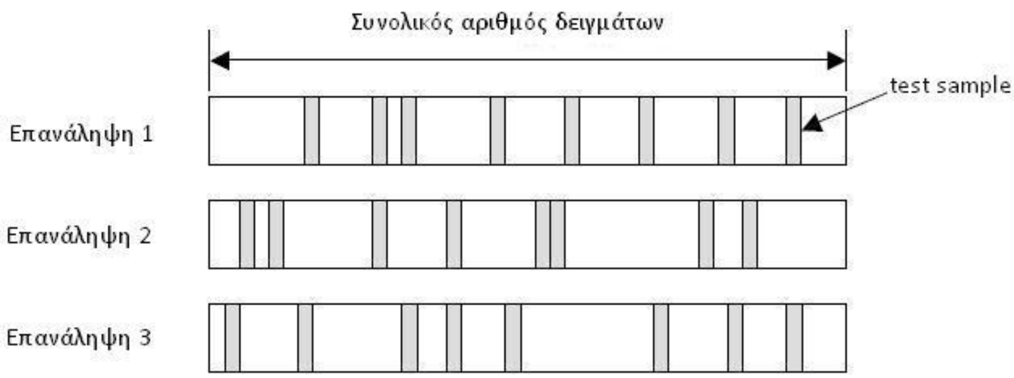
Για να εξηγηθεί ο σκοπός του cross-validation υποθέτουμε ότι έχουμε ένα μοντέλο με μία ή περισσότερες άγνωστες παραμέτρους, καθώς και ένα σύνολο δεδομένων στο οποίο το μοντέλο μπορεί να προσαρμοστεί (το σύνολο εκπαίδευσης). Η διαδικασία προσαρμογής βελτιστοποιεί τις παραμέτρους του μοντέλου για να μπορέσει το μοντέλο να προσαρμοστεί στα δεδομένα εκπαίδευσης όσο καλύτερα γίνεται. Στη συνέχεια, επιλέγοντας ένα ανεξάρτητο δείγμα του συνόλου δοκιμών από τον ίδιο πληθυσμό από τον οποίο προέρχεται και το σύνολο εκπαίδευσης, αποδεικνύεται γενικά ότι το μοντέλο δεν προσαρμόζεται στο σύνολο δοκιμών τόσο καλά όσο στο σύνολο εκπαίδευσης. Αυτό ονομάζεται *over fitting*, και συμβαίνει συνήθως όταν ο αριθμός των δεδομένων της εκπαίδευσης είναι μικρός, ή όταν ο αριθμός των παραμέτρων στο μοντέλο είναι μεγάλος. Το cross-validation είναι ένας τρόπος πρόβλεψης της προσαρμογής του μοντέλου σε ένα υποθετικό σύνολο δοκιμών όταν ένα σαφές σύνολο δοκιμών δεν είναι διαθέσιμο.

Υποθέτουμε ότι το σύνολο δεδομένων αποτελείται από Ν δείγματα (*samples*), και Μ χαρακτηριστικά (*features*), δηλαδή κάθε δείγμα είναι διάστασης Μ. Οι πιο συνηθισμένοι τύποι για cross validation είναι οι εξής [38] :

- ✓ <u>Η μέθοδος holdout.</u> Είναι ο πιο απλός τρόπος για cross validation. Το σύνολο δεδομένων χωρίζεται σε ένα σύνολο δεδομένων για την εκπαίδευση του ταξινομητή και σε ένα σύνολο δεδομένων για την εκτίμηση του ποσοστού σφάλματος του εκπαιδευμένου ταξινομητή, όπως φαίνεται στην εικόνα 14. Αυτή η μέθοδος έχει δύο σημαντικά μειονεκτήματα. Σε περίπτωση που τα δεδομένα εισόδου που διαθέτουμε είναι σποραδικά, δεν είναι εύκολα εφικτό να αποσπάσουμε μία ποσότητα δειγμάτων από το σύνολο δεδομένων για να χρησιμοποιηθεί για σύνολο δεδομένων επικύρωσης. Επιπλέον, επειδή η διαδικασία εκτελείται μία μόνο φορά για ένα σύνολο εκπαίδευσης και ένα σύνολο επικύρωσης μόνο, η εκτίμηση του ποσοστού σφάλματος μπορεί να γίνει παραπλανητική στην περίπτωση ενός «ατυχούς» διαχωρισμού των δειγμάτων.

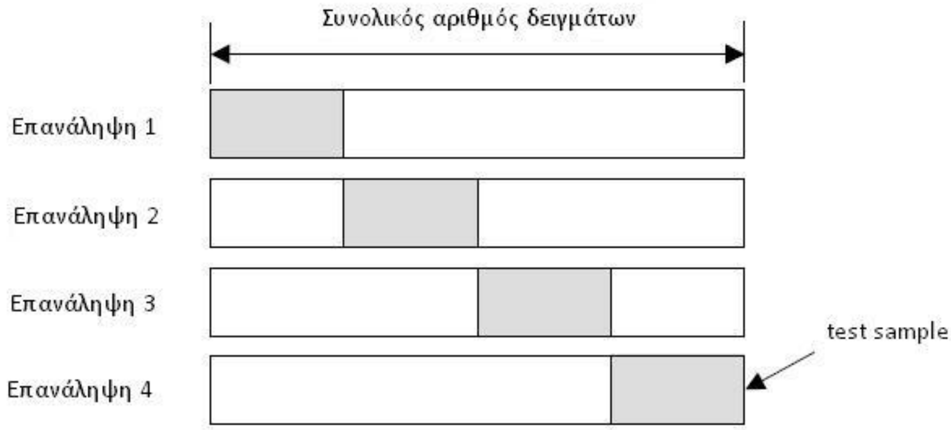
Εικόνα 14: Η μέθοδος holdout, για το διαχωρισμό του συνόλου δεδομένων σε δείγματα εκπαίδευσης και δείγματα ελέγχου [27].

- ✓ <u>Η μέθοδος repeated random sub-sampling Validation.</u> Εκτελεί K διαμερισμούς του συνόλου δεδομένων με κάθε έναν από αυτούς να αποτελείται από έναν αριθμό (ίδιος για όλους τους διαμερισμούς) τυχαίων δειγμάτων. Για κάθε τέτοιο διαμερισμό, το μοντέλο εκπαιδεύεται ξανά από το μηδέν με τα δείγματα που έχουν επιλεχθεί να αποτελούν το σύνολο εκπαίδευσης και εκτιμάται η απόδοσή του με τα δείγματα του συνόλου ελέγχου, τα οποία είναι διαφορετικά. Το πλεονέκτημα αυτής της μεθόδου (σε σχέση με την k-fold cross validation μέθοδο που αναλύεται αμέσως μετά) είναι ότι η αναλογία του διαμερισμού εκπαίδευσης/επικύρωσης δεν εξαρτάται από τον αριθμό των επαναλήψεων (*folds*). Το μειονέκτημα αυτής της μεθόδου είναι ότι κάποια δείγματα μπορεί να μην επιλεγούν ποτέ για να συμμετέχουν στο σύνολο δοκιμών ενώ άλλα μπορεί να επιλεγούν περισσότερο από μία φορά, οπότε να υπάρχει επικάλυψη στα υποσύνολα των δειγμάτων επικύρωσης. Αυτή η μέθοδος αναπαρίσταται στην εικόνα 15.



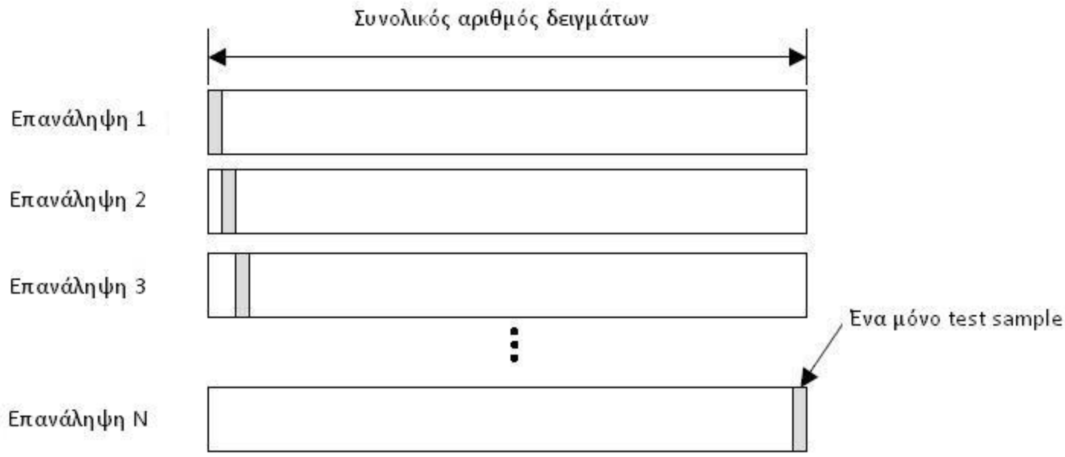
Εικόνα 15: Η μέθοδος repeated random sub-sampling Validation, για το διαχωρισμό του συνόλου δεδομένων σε δείγματα εκπαίδευσης και δείγματα ελέγχου [27].

- ✓ <u>Η μέθοδος K-Fold Cross - Validation.</u> Το σύνολο δεδομένων χωρίζεται τυχαία σε K περίπου ίσου μεγέθους (N/K) υποσύνολα. Από τα K αυτά υποσύνολα, ένα υποσύνολο διατηρείται ως δεδομένα επικύρωσης (*test set*) για την εκτίμηση του ποσοστού σφάλματος του ταξινομητή και τα υπόλοιπα K-1 υποσύνολα χρησιμοποιούνται για δεδομένα εκπαίδευσης του ταξινομητή (*train set*). Έπειτα, η διαδικασία του cross validation επαναλαμβάνεται K φορές (όσα είναι τα folds), με κάθε ένα από τα K υποσύνολα να χρησιμοποιείται ακριβώς μία φορά για δεδομένα επικύρωσης. Τα K αποτελέσματα από τα folds μπορούν να συνδυαστούν σε ένα μέσο όρο για να παράγουν μία εκτίμηση για το μέσο σφάλμα. Το πλεονέκτημα αυτής της μεθόδου, με τα επαναλαμβανόμενα τυχαία υποσύνολα, είναι ότι όλα τα δείγματα χρησιμοποιούνται τόσο στην εκπαίδευση του μοντέλου όσο και στην αξιολόγησή του, και κάθε δείγμα χρησιμοποιείται για αξιολόγηση ακριβώς μία φορά. Έτσι, με το cross validation επιτυγχάνεται το εξής παράδοξο: η εκτίμηση ενός μοντέλου γίνεται σε ανεξάρτητα δεδομένα που δε χρησιμοποιήθηκαν κατά την εκπαίδευση του, αλλά ταυτόχρονα όλα τα δεδομένα, στο τέλος των K επαναλήψεων, χρησιμοποιούνται για την εκπαίδευση του μοντέλου. Η πιο συνήθης επιλογή για το K είναι 10. Αυτή η μέθοδος φαίνεται στην εικόνα 16.



Εικόνα 16: Η μέθοδος K-Fold Cross – Validation, για το διαχωρισμό του συνόλου δεδομένων σε δείγματα εκπαίδευσης και δείγματα ελέγχου [27].

- ✓ <u>Η μέθοδος Leave-one-out.</u> Είναι μία εκφυλισμένη περίπτωση του K-Fold Cross – Validation και παρουσιάζεται στην εικόνα 17. Περιλαμβάνει τη χρήση ενός μόνο δείγματος, από το σύνολο δειγμάτων, ως δεδομένα ελέγχου και όλα τα άλλα δείγματα ως δεδομένα εκπαίδευσης. Αυτό επαναλαμβάνεται έτσι ώστε κάθε δείγμα να χρησιμοποιηθεί μία φορά για την εκτίμηση του μοντέλου. Η διαδικασία είναι η ίδια με το K-Fold Cross – Validation, με το K να είναι ίσο με τον αριθμό των δειγμάτων του αρχικού συνόλου δεδομένων. Αυτή η μέθοδος έχει συνήθως πολύ ακριβό υπολογιστικό κόστος εξαιτίας του μεγάλου αριθμού επαναλήψεων που απαιτούνται για τη διαδικασία της μάθησης.



Εικόνα 17: Η μέθοδος Leave-one-out, για το διαχωρισμό του συνόλου δεδομένων σε δείγματα εκπαίδευσης και δείγματα ελέγχου [27].

Ο στόχος του cross-validation είναι η εκτίμηση του επιθυμητού επιπέδου προσαρμογής ενός μοντέλου σε ένα σύνολο δεδομένων το οποίο είναι ανεξάρτητο από το σύνολο δεδομένων που χρησιμοποιήθηκε για να εκπαιδευτεί το μοντέλο. Μπορεί να χρησιμοποιηθεί για να εκτιμήσει κάθε ποσοτικό μέτρο προσαρμογής για τα δεδομένα και το μοντέλο. Στην περίπτωση της δυαδικής ταξινόμησης, κάθε δείγμα του συνόλου δοκιμών μπορεί να προβλεφθεί είτε σωστά είτε λάθος. Σε αυτή την κατάσταση το ποσοστό σφάλματος ταξινόμησης μπορεί να χρησιμοποιηθεί για να συνοψίσει την προσαρμογή, παρόλο που και άλλα μέτρα όπως το positive predictive value μπορούν να χρησιμοποιηθούν επίσης.

2.5 Επιλογή Χαρακτηριστικών

Στη μηχανική μάθηση και τη στατιστική [39, 40], η επιλογή χαρακτηριστικών (*feature selection*), επίσης γνωστή και ως επιλογή μεταβλητών (*variable selection*), μείωση χαρακτηριστικών (*feature reduction*), επιλογή γνωρισμάτων (*attribute selection*) ή επιλογή υποσυνόλου μεταβλητών (*variable subset selection*), είναι η τεχνική της επιλογής ενός υποσυνόλου σχετικών χαρακτηριστικών για τη δημιουργία ισχυρών μοντέλων μάθησης (για παράδειγμα ταξινομητών). Όταν κάτι τέτοιο εφαρμόζεται στον τομέα της βιολογίας ονομάζεται επίσης και διακριτή γονιδιακή υπογραφή και ανιχνεύει τα ισχυρά γονίδια βασισμένο σε DNA microarray πειράματα. Με τη διαγραφή των πιο άσχετων ή περιττών χαρακτηριστικών από τα δεδομένα, επιτυγχάνεται βελτίωση της απόδοσης των μοντέλων μάθησης γιατί:

- ✓ <u>Απαλλασσόμαστε από την επίδραση των πολλών διαστάσεων του προβλήματος</u>. Η αύξηση του αριθμού των διαστάσεων των δεδομένων έχει σαν αποτέλεσμα πολλοί τύποι δεδομένων ανάλυσης και προβλήματα ταξινόμησης να γίνονται σημαντικά δυσκολότερα, γιατί έτσι κάποιες φορές τα δεδομένα γίνονται όλο και περισσότερο σποραδικά (αραιά) στο χώρο που καταλαμβάνουν. Αυτό το φαινόμενο ονομάζεται κατάρα των διαστάσεων και οδηγεί σε μεγάλα προβλήματα για τις μεθόδους τόσο της επιβλεπόμενης όσο και της μη επιβλεπόμενης μάθησης.
- ✓ <u>Ενισχύεται η ικανότητα γενίκευσης</u>. Όταν ένα σύνολο δεδομένων έχει πολλές διαστάσεις μπορεί να περιέχει περισσότερη πληροφορία από ότι χρειάζεται για να φτιαχτεί ένα εύρωστο μοντέλο μάθησης. Στην περίπτωση της επιβλεπόμενης μάθησης ή ταξινόμησης, τα διαθέσιμα δεδομένα εκπαίδευσης συχνά αποτελούνται από έναν μικρό αριθμό δειγμάτων για εκπαίδευση, της τάξης κάποιων δεκάδων, σε σύγκριση με τον πολύ μεγάλο αριθμό των χαρακτηριστικών, της τάξης κάποιων χιλιάδων. Οπότε, η χρήση όλων των χαρακτηριστικών περικλείει τον κίνδυνο της δημιουργίας μοντέλων που αδυνατούν να αναγνωρίσουν τα ουσιαστικά πρότυπα από τα δεδομένα και κατά συνέπεια να γίνονται ανίκανα να ταξινομήσουν επακριβώς τα νέα δεδομένα. Επίσης, στη μη επιβλεπόμενη μάθηση ή ομαδοποίηση, διάφοροι ζωτικής σημασίας ορισμοί όπως η πυκνότητα ή η

απόσταση ανάμεσα στα σημεία μπορεί να γίνουν λιγότερο πειστικοί, αφού οι πολλές διαστάσεις τείνουν να κάνουν την εγγύτητα ανάμεσα στα σημεία πιο ομοιόμορφη. Σαν αποτέλεσμα, ένας μεγάλος αριθμός χαρακτηριστικών μπορεί να οδηγήσει σε χαμηλότερη ακρίβεια ταξινόμησης και σε clusters φτωχής ποιότητας.

- ✓ Επιταχύνεται η διαδικασία της μάθησης. Τα δεδομένα υψηλών διαστάσεων είναι ένα επίσης σημαντικό πρόβλημα για πολλούς αλγορίθμους ταξινόμησης εξαιτίας του υψηλού υπολογιστικού κόστους και της χρήσης της μνήμης που αυτά απαιτούν.
- ✓ Βελτιώνεται η ερμηνεία του μοντέλου μάθησης. Μία μείωση στο χώρο των χαρακτηριστικών οδηγεί σε καλύτερη κατανόηση του μοντέλου και απλοποιεί τη χρήση διαφορετικών τεχνικών οπτικοποίησης.

Η επιλογή ενός υποσυνόλου χαρακτηριστικών είναι μία τεχνική πολύ σημαντική τόσο στην επιβλεπόμενη όσο και στη μη επιβλεπόμενη μάθηση, αφού μας βοηθάει να αποκτήσουμε καλύτερη αντίληψη για τα δεδομένα μας, γνωστοποιώντας μας ποια είναι τα πιο σημαντικά χαρακτηριστικά και πώς αυτά σχετίζονται μεταξύ τους. Το πλεονέκτημα της επιλογής υποσυνόλου χαρακτηριστικών είναι ότι καμία πληροφορία σχετικά με τη σημασία των ατομικών χαρακτηριστικών δε χάνεται. Το μειονέκτημα όμως αυτής της τεχνικής προκύπτει όταν απαιτείται ένα μικρό σύνολο χαρακτηριστικών και τα αρχικά χαρακτηριστικά είναι πολύ διαφορετικά, οπότε μπορεί να χαθεί πληροφορία αφού κάποια χαρακτηριστικά πρέπει να παραλειφθούν.

2.5.1 Μέθοδοι επιλογής χαρακτηριστικών

Γενικά, υπάρχουν τρεις τύποι μεθόδων [40, 41] για την επιλογή του υποσυνόλου χαρακτηριστικών (*feature subset selection*), filters, wrappers και embedded.

- ✓ Οι Filter μέθοδοι είναι ανεξάρτητες από τον ύστερο αλγόριθμο μάθησης που εφαρμόζεται, παραμελούν τις αλληλεπιδράσεις μεταξύ των χαρακτηριστικών και χρησιμοποιούν κάποιο μέτρο διαχωριστικότητας σαν ένα κριτήριο εκτίμησης. Εστιάζουν στις εγγενείς ιδιότητες των δεδομένων χρησιμοποιώντας διάφορες στοχαστικές μετρικές όπως Fisher's ratio, T-statistics, x2 statistic, information gain και πολλές άλλες. Τα χαρακτηριστικά κατατάσσονται σύμφωνα με την τιμή τους σε αυτές τις μετρικές και τα χαρακτηριστικά εκείνα με τις υψηλότερες τιμές, που παρέχουν την καλύτερη ικανότητα διάκρισης μαζί με την υψηλότερη ακρίβεια ταξινόμησης, επιλέγονται ως τα πιο σημαντικά.

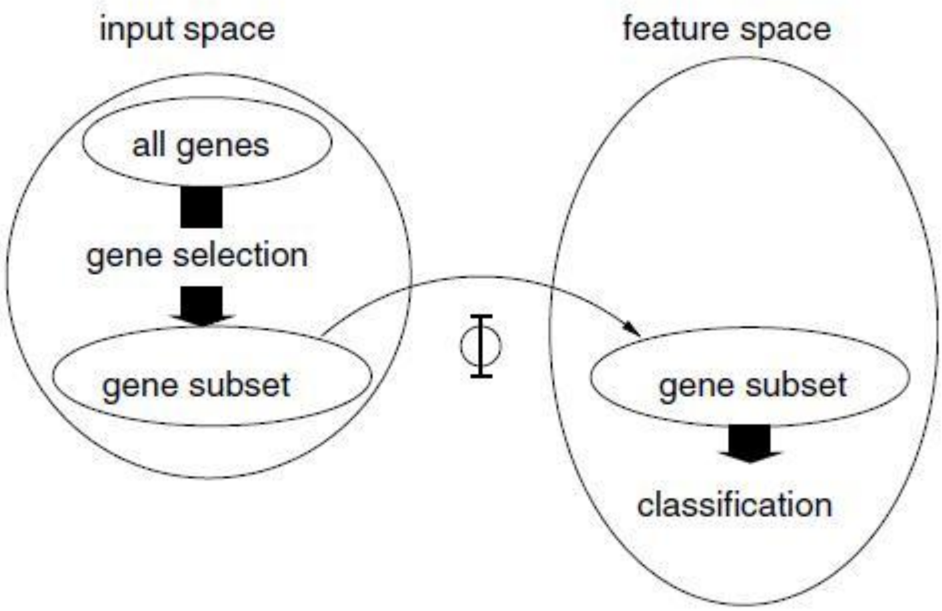
- ✓ Οι Wrapper μέθοδοι, αντίθετα, ενσωματώνουν τον αλγόριθμο μάθησης στην διαδικασία επιλογής χαρακτηριστικών αφού βασίζονται στην απόδοση ενός συγκεκριμένου ταξινομητή για να εκτιμήσουν την ποιότητα ενός συνόλου χαρακτηριστικών. Ερευνούν το χώρο των υποσυνόλων των χαρακτηριστικών και υπολογίζουν την εκτιμώμενη απόδοση ενός συγκεκριμένου αλγορίθμου μάθησης για κάθε χαρακτηριστικό που μπορεί να προστεθεί ή να διαγραφεί από το υποσύνολο χαρακτηριστικών. Ο χώρος των χαρακτηριστικών μπορεί να ερευνηθεί με πολλούς τρόπους, για παράδειγμα με προς τα εμπρός τρόπο (προσθέτοντας χαρακτηριστικά σε ένα αρχικά άδειο σύνολο από χαρακτηριστικά) ή με προς τα πίσω τρόπο ξεκινώντας με όλο το σύνολο των χαρακτηριστικών και διαγράφοντας ένα ή περισσότερα κάθε φορά χαρακτηριστικά. Αυτό επιτυγχάνεται δουλεύοντας με έναν αναδρομικό τρόπο όπου ένας ταξινομητής χρησιμοποιείται για να εκχωρήσει ένα σχετικό βάρος σε κάθε χαρακτηριστικό και έπειτα τα χαρακτηριστικά με το χαμηλότερο βάρος διαγράφονται. Περισσότερα του ενός χαρακτηριστικά μπορούν να διαγραφούν σε κάθε επανάληψη. Στον επόμενο κύκλο επανάληψης, τα βάρη επανεκτιμούνται και προσαρμόζονται δυναμικά, ενώ η διαδικασία επαναλαμβάνεται αναδρομικά. Στο τέλος, το μικρότερο σύνολο χαρακτηριστικών που επιτυγχάνει την καλύτερη ακρίβεια ταξινόμησης επιλέγεται ως το σύνολο των πιο σημαντικών χαρακτηριστικών. Συνήθως, μία τέτοια εξαντλητική αναζήτηση είναι πολύ ακριβή.
- ✓ Οι embedded μέθοδοι εκτελούν feature selection μέσα στη διαδικασία της εκπαίδευσης. Στοχεύουν στην άμεση ενσωμάτωση του feature selection ή της διαδικασίας εκχώρησης βαρών μέσα στον αλγόριθμο μάθησης του ταξινομητή, επιτυγχάνοντας έτσι να διατηρήσουν τα εγγενή χαρακτηριστικά των δεδομένων στη διαδικασία της ταξινόμησης. Συγκρινόμενες με τις wrapper μεθόδους, οι embedded είναι συνήθως πιο αποτελεσματικές, αφού

κοιτούν στη δομή του περιλαμβανομένου μοντέλου μάθησης και χρησιμοποιούν τις ιδιότητές του για να καθοδηγήσουν την εκτίμηση των χαρακτηριστικών και την αναζήτηση.

Στην προσέγγιση με filter μεθόδους, η επιλογή χαρακτηριστικών είναι ανεξάρτητη από τον αλγόριθμο μηχανικής μάθησης (ταξινομητής), όπως φαίνεται στην εικόνα 18. Αυτό είναι υπολογιστικά πιο αποτελεσματικό αλλά αγνοεί το γεγονός ότι η επιλογή των χαρακτηριστικών μπορεί να εξαρτάται από τον αλγόριθμο μάθησης. Αντίθετα, η προσέγγιση με wrapper μεθόδους έχει συχνά καλύτερη απόδοση, λαμβάνει υπόψη τις εξαρτήσεις των χαρακτηριστικών στο υποσύνολο, από τον αλγόριθμο μάθησης, απαιτεί όμως ακριβό υπολογιστικό κόστος.

Γενικά, η επιλογή χαρακτηριστικών εκτελείται πριν από την εκπαίδευση του μοντέλου, ώστε αυτόματα να επιλεγούν τα χαρακτηριστικά εκείνα του συνόλου δεδομένων που είναι τα πιο αρμόδια για να χρησιμοποιηθούν στο μοντέλο.

Στην ανάλυση δεδομένων σε μικροσυστοιχίες [12], η επιλογή γονιδίων (*gene selection*) αποτελεί ένα κεντρικό θέμα τα τελευταία χρόνια. Χρησιμοποιείται συχνά για να προσδιορίσει γονίδια που είναι πιο σχετικά ως προς ένα συγκεκριμένο θέμα ταξινόμησης, όπως για παράδειγμα τα γονίδια που διαφοροποιούνται μεταξύ των διαφορετικών συνθηκών εργαστηρίου κάτω από τις οποίες μπορεί, για παράδειγμα, ο σακχαρομύκητας να εκφράζεται. Η επιλογή των γονιδίων, παίζει σημαντικό ρόλο στα θέματα ταξινόμησης αφού βελτιώνει την ακρίβεια πρόβλεψης των ταξινομητών χρησιμοποιώντας μόνο τα σημαντικά γονίδια. Επιπλέον, εξοικονομεί υπολογιστικό κόστος μειώνοντας τον αριθμό των διαστάσεων. Πολύ σημαντικό είναι επίσης ότι αν είναι δυνατό να προσδιορίσουμε ένα μικρό υποσύνολο βιολογικά σχετικών γονιδίων, τότε αυτό μπορεί να παρέχει βαθειά γνώση για κατανόηση του θεμελιώδη μηχανισμού ενός συγκεκριμένου βιολογικού φαινομένου. Μία τέτοια πληροφορία μπορεί να είναι πολύ χρήσιμη για το σχεδιασμό στοχευμένων και παράλληλα, λιγότερο ακριβών πειραμάτων, στοχεύοντας σε ένα μικρό μόνο αριθμό γονιδίων.



Εικόνα 18: Filter προσέγγιση για την επιλογή υποσυνόλου γονιδίων [12].

ΚΕΦΑΛΑΙΟ 3: Αλγόριθμοι Ταξινόμησης και Εκτίμηση Απόδοσης

Σε αυτό το κεφάλαιο, παρουσιάζονται αναλυτικά οι αλγόριθμοι που χρησιμοποιούνται για την εκτέλεση της γονιδιακής ταξινόμησης σε αυτή την εργασία. Περιγράφεται αναλυτικά η λειτουργία των ταξινομητών support vector machines, kernel nearest centroid, kernel minimum squared error, kernel subspace καθώς και τεσσάρων μεθόδων μείγματος ειδικών που συνδυάζουν τις ατομικές αποφάσεις των ταξινομητών σε μία συνδυασμένη απόφαση με στόχο την ευρωστία και την καλύτερη ακρίβεια ταξινόμησης. Αυτές οι μέθοδοι είναι, πρώτον, η συγχώνευση ταξινομητών με συνδυασμό βασισμένο σε ασαφή βάρη, σε ένα gradient descent learning αλγόριθμο καθώς και στη μίξη των δύο τελευταίων, και δεύτερον, η επιλογή ταξινομητών με διαχωρισμό του χώρου των χαρακτηριστικών σε ομάδες μέσω του self organized map.

Επίσης, γίνεται λεπτομερής αναφορά στον αλγόριθμο RFE-LNW που χρησιμοποιείται για την επιλογή χαρακτηριστικών, αφού εκτελεί αναδρομική εξάλειψη των γονιδίων, αναδεικνύοντας εκείνα τα οποία διαφοροποιούν περισσότερο την έκφρασή τους στις δύο κλάσεις ενδιαφέροντος.

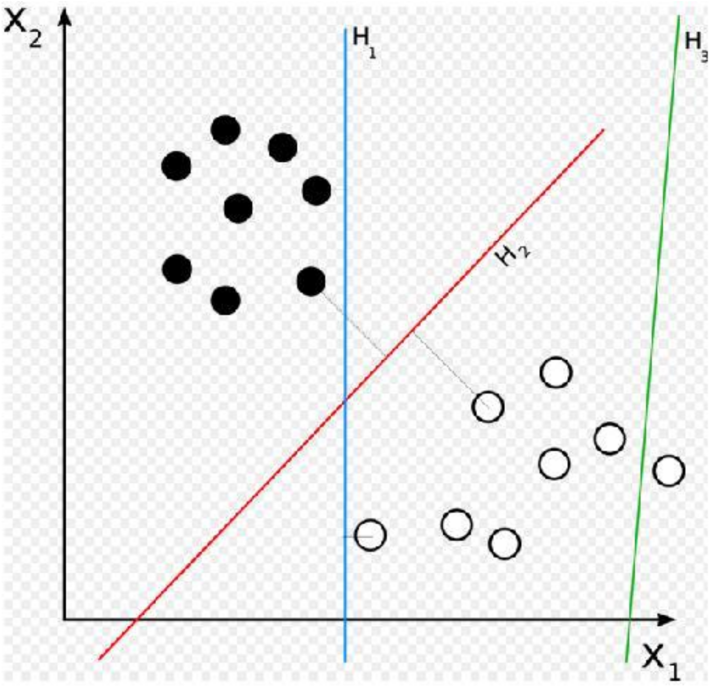
Τέλος, παρουσιάζεται το μέτρο kernel target alignment για την αξιολόγηση της καταλληλότητας των πυρήνων για το εκάστοτε σύνολο δεδομένων, καθώς και η σχέση για την εκτίμηση της απόδοσης των μεθόδων ταξινόμησης.

3.1 Μηχανές Διανυσμάτων Υποστήριξης

Μία μηχανή διανύσματος υποστήριξης (*support vector machine*) [42] είναι μία έννοια στη στατιστική και στην επιστήμη υπολογιστών, για ένα σύνολο σχετικών μεθόδων επιβλεπόμενης μάθησης που αναλύει δεδομένα και αναγνωρίζει πρότυπα, τα οποία χρησιμοποιούνται για ταξινόμηση και για regression ανάλυση. Το πρότυπο support vector machine (SVM) παίρνει ένα σύνολο δεδομένων εισόδου και προβλέπει, για κάθε δεδομένη είσοδο ποια από τις δύο δυνατές κλάσεις σχηματίζουν την είσοδο, μέσω της χρήσης πυρήνα, καθιστώντας έτσι το SVM έναν μη στατιστικό δυαδικό γραμμικό ταξινομητή.

Δεδομένου ενός συνόλου δειγμάτων εκπαίδευσης, με καθένα από αυτά να είναι κατηγοριοποιημένο στη μία από τις δύο κλάσεις, ένας SVM αλγόριθμος εκπαίδευσης δημιουργεί ένα μοντέλο που εκχωρεί νέα δείγματα στη μία ή στην άλλη κλάση. Αυτό που είναι στην ουσία ένα SVM μοντέλο είναι μία αναπαράσταση των δειγμάτων ως σημεία στο χώρο, που απεικονίζονται (*mapped*) μέσω μιας συνάρτησης πυρήνα έτσι ώστε τα δείγματα των δύο ξεχωριστών κλάσεων να χωρίζονται από ένα σαφές όριο το οποίο να είναι όσο πιο ευρύ γίνεται, όπως φαίνεται στην εικόνα 19. Μετά από αυτό, τα νέα δείγματα απεικονίζονται στον ίδιο αυτό χώρο και προβλέπεται γι' αυτά η κλάση στην οποία ανήκουν με κριτήριο την πλευρά του ορίου στην οποία βρίσκονται.

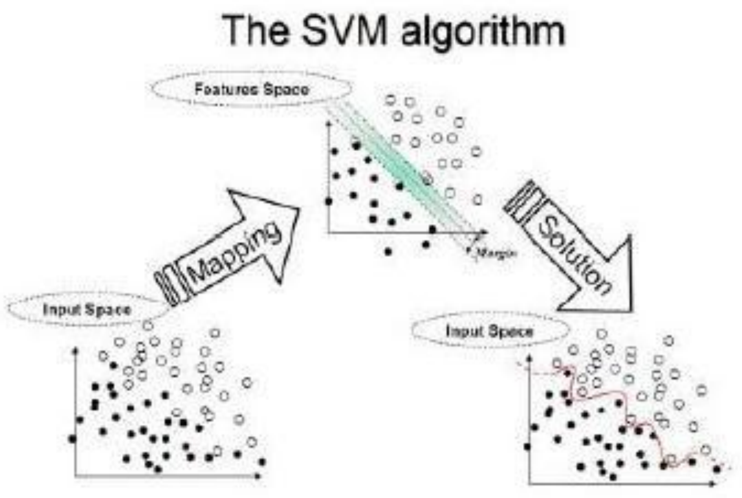
Πιο συγκεκριμένα, ένα support vector machine κατασκευάζει ένα υπερεπίπεδο ή ένα σύνολο από υπερεπίπεδα σε ένα χώρο υψηλών ή άπειρων διαστάσεων, το οποίο μπορεί να χρησιμοποιηθεί για ταξινόμηση. Διαισθητικά, ένας καλός διαχωρισμός επιτυγχάνεται από το υπερεπίπεδο που έχει τη μεγαλύτερη απόσταση από το κοντινότερο δείγμα εκπαίδευσης οποιασδήποτε κλάσης, το λεγόμενο λειτουργικό περιθώριο (*functional margin*), αφού γενικά ισχύει ότι όσο μεγαλύτερο είναι το περιθώριο, τόσο μικρότερο είναι το σφάλμα γενίκευσης του ταξινομητή.



Εικόνα 19: Η γραμμή Η3 (πράσινη) δεν διαχωρίζει τις δύο κλάσεις. Η γραμμή Η1 (μπλε) τις διαχωρίζει με μικρό περιθώριο και η γραμμή Η2 (κόκκινη) με το μέγιστο περιθώριο [41].

Ενώ το αρχικό πρόβλημα μπορεί να αναφέρεται σε ένα χώρο πεπερασμένων διαστάσεων, συχνά συμβαίνει τα σύνολα που πρέπει να ταξινομηθούν, να μην είναι γραμμικά διαχωρίσιμα σε αυτόν το χώρο. Γι' αυτό το λόγο, ο αρχικός χώρος πεπερασμένων διαστάσεων απεικονίζεται σε ένα χώρο υψηλότερων διαστάσεων, κάνοντας προφανώς το διαχωρισμό ευκολότερο σε αυτό το νέο χώρο.

Για να διατηρηθεί το υπολογιστικό κόστος σε λογικά πλαίσια, οι απεικονίσεις που χρησιμοποιούνται από τα SVM συστήματα σχεδιάζονται για να εξασφαλίσουν ότι οι υπολογισμοί στο νέο χώρο (εσωτερικά γινόμενα) μπορούν να υπολογιστούν εύκολα με βάση τις μεταβλητές στον αρχικό χώρο, μέσω της συνάρτησης πυρήνα K(x,y) που επιλέγεται στο πρόβλημα.

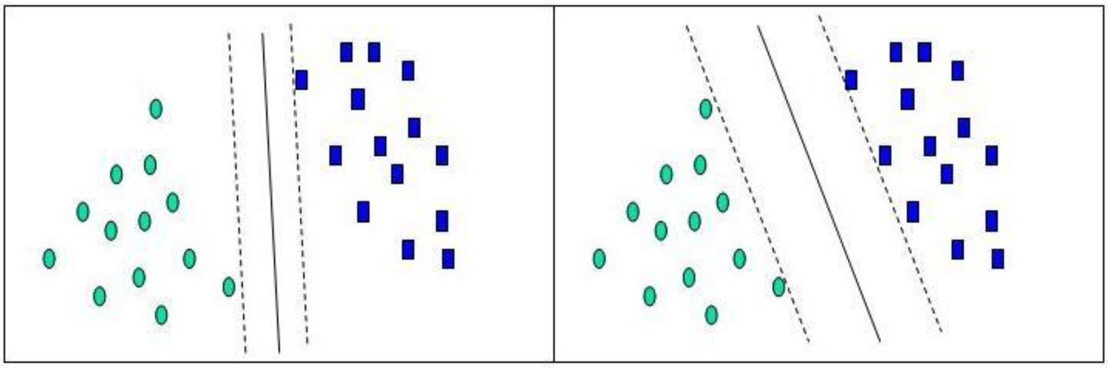


Εικόνα 20: Ανασκόπηση της διαδικασίας που ακολουθούν τα support vector machines [42].

Το υπερεπίπεδο διαχωρισμού, στο χώρο των υψηλότερων διαστάσεων, καθορίζεται ως το σύνολο των σημείων των οποίων το εσωτερικό γινόμενο με ένα διάνυσμα βαρών σε αυτό το χώρο να είναι σταθερό. Οπότε, ο στόχος του σχεδιασμού των SVM είναι να βρεθεί το βέλτιστο υπερεπίπεδο που διαχωρίζει τις ομάδες του vector με τέτοιο τρόπο ώστε τα δείγματα με τη μία κατηγορία του target variable να βρίσκονται στη μία πλευρά του επιπέδου και τα δείγματα με την άλλη κατηγορία να βρίσκονται στην άλλη πλευρά του επιπέδου. Στην εικόνα 20 φαίνεται συνοπτικά η παραπάνω διαδικασία. Τα διανύσματα-δείγματα που βρίσκονται κοντά στο υπερεπίπεδο είναι τα διανύσματα υποστήριξης (*support vectors*).

3.1.1 Γραμμικό SVM

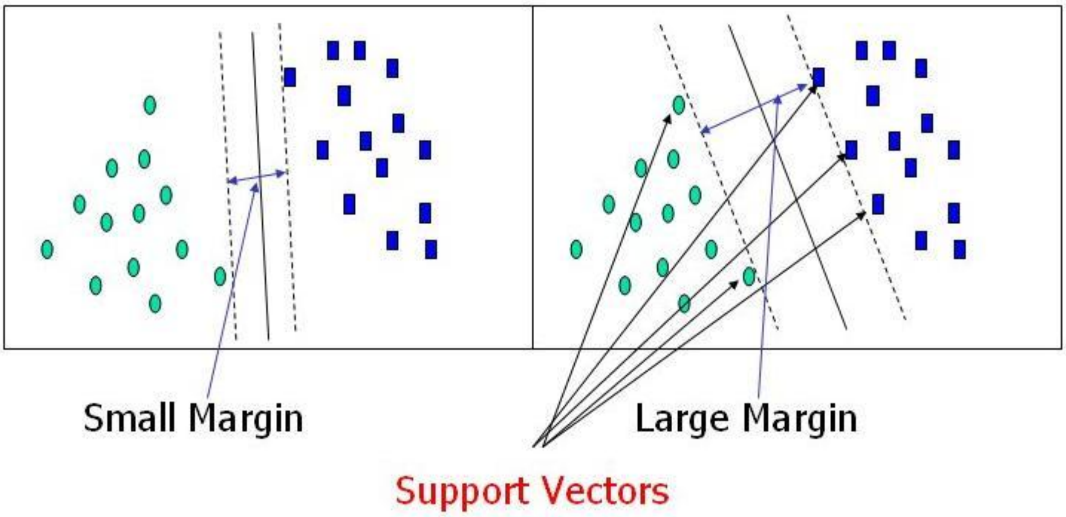
Πριν μελετήσουμε τι συμβαίνει σε Μ διαστάσεις [43], ας μελετήσουμε πρώτα τι γίνεται στις δύο. Εκτελούμε δυαδική ταξινόμηση και σχεδιάζουμε τα δείγματα χρησιμοποιώντας την τιμή του ενός χαρακτηριστικού στον άξονα x και την τιμή του άλλου χαρακτηριστικού στον άξονα y. Στο ιδανικό παράδειγμα που οι δύο κλάσεις είναι εντελώς διαχωρίσιμες, το SVM προσπαθεί να βρει ένα μονοδιάστατο υπέρ-επίπεδο (δηλαδή μία γραμμή) που να χωρίζει τα δείγματα σύμφωνα με το στόχο (target) της κλάσης τους. Υπάρχει άπειρος αριθμός τέτοιων πιθανών γραμμών, και δύο υποψήφιες φαίνονται στην εικόνα 21. Οπότε, το ερώτημα που προκύπτει είναι ποια από αυτές είναι η καλύτερη και πώς καθορίζεται μία βέλτιστη γραμμή. Οι διακεκομμένες γραμμές που έχουν σχεδιαστεί παράλληλα από τη διαχωριστική γραμμή, ορίζουν την απόσταση ανάμεσα σε αυτήν και στα κοντινότερα δείγματα ως



Εικόνα 21: Παράδειγμα δυαδικής ταξινόμησης σε δύο διαστάσεις όπου η μία κλάση αναπαριστάται με τα ορθογώνια, ενώ η άλλη με τα ωοειδή [42].

προς αυτή. Η απόσταση ανάμεσα στις δύο διακεκομμένες γραμμές ονομάζεται περιθώριο (*margin*). Τα δείγματα που περιορίζουν το εύρος του περιθωρίου αποτελούν τα διανύσματα υποστήριξης.

Μία SVM ανάλυση βρίσκει τη γραμμή που είναι προσανατολισμένη έτσι, ώστε το περιθώριο ανάμεσα στα support vectors να μεγιστοποιείται. Στην εικόνα 22 φαίνεται ότι η γραμμή στο δεξί κομμάτι της είναι καλύτερη από αυτή στο αριστερό. Δυστυχώς, για να μπορέσουμε να εξετάσουμε μία γενική περίπτωση, πρέπει να σκεφτούμε ότι τα SVM πρέπει να μπορούν να χειριστούν δείγματα με περισσότερα από δύο χαρακτηριστικά, δείγματα τα οποία χρειάζονται μη γραμμικές καμπύλες για να διαχωριστούν, και ταξινομήσεις με περισσότερες από δύο κλάσεις.



Εικόνα 22: Δείχνει ένα μικρό περιθώριο (αριστερά) και ένα μεγαλύτερο (δεξιά) διαχωρισμού των δύο κλάσεων, καθώς και τα support vectors [42].

Περνώντας στις M διαστάσεις, ένα δείγμα απεικονίζεται σαν ένα m -διάστατο διάνυσμα (αποτελείται από m χαρακτηριστικά), και ένα γραμμικό SVM ψάχνει αν υπάρχει ένα (m-1) – διάστατο υπέρ-επίπεδο που να διαχωρίζει τις δύο κλάσεις. Πάλι και σε αυτή την περίπτωση των πολλών διαστάσεων, το καλύτερο υπέρ-επίπεδο είναι αυτό που αναπαριστά το μεγαλύτερο περιθώριο ανάμεσα στις δύο κλάσεις.

3.1.2 Αναλυτικά ο τρόπος λειτουργίας των γραμμικών SVM

Θεωρούμε ένα σύνολο δεδομένων εκπαίδευσης X [41] αποτελούμενο από n δείγματα της μορφής $X = \{\, (x_i, y_i) \mid x_i \in R^m \,,\ y_i \in \{-1, +1\}\}_{i=1}^{n}$, όπου το y_i είναι είτε +1 είτε -1, υποδεικνύοντας την κλάση που ανήκει το x_i. Κάθε x_i είναι ένα m – διάστατο πραγματικό διάνυσμα.

Τα SVM προσπαθούν να βρουν το καλύτερο διαχωριστικό υπέρ-επίπεδο (με το μέγιστο περιθώριο) που να διαχωρίζει τα δείγματα που έχουν $y_i = +1$ από εκείνα που έχουν $y_i = -1$. Αυτό επιτυγχάνεται μεγιστοποιώντας την απόσταση $\frac{2}{||w||}$ ανάμεσα στις δύο παράλληλες γραμμές $(w \cdot x\,) + b = 1$ και $(w\ \cdot x) +\ b = -1$, οι οποίες σχηματίζουν το περιθώριο διαχωρισμού των δύο κλάσεων όπως φαίνεται στο σχήμα. Το τελικό διαχωριστικό υπερεπίπεδο περνά από το μέσο αυτού του περιθωρίου με εξίσωση $(w\ \cdot x) +\ b = 0$. Αυτή η διαδικασία φαίνεται στην εικόνα 23.

Για να εμποδίσουμε τα δείγματα να βρεθούν μέσα στο περιθώριο, προσθέτουμε τον ακόλουθο περιορισμό: Για κάθε i είτε $(w \cdot x_i) + b \geq 1$, για κάθε x_i που ανήκει στην πρώτη κλάση ($y_i = +1$), είτε $(w \cdot x_i) + b \leq -1$, για κάθε x_i που ανήκει στη δεύτερη κλάση ($y_i = -1$). Αυτό μπορεί να γραφτεί ως: $y_i \left((w \cdot x_i) + b\right) \geq 1$, για όλα τα $\{(x_i, y_i,)\}$, με $1 \leq i \leq n$.

Οπότε, στο στάδιο της ταξινόμησης ενός νέου δείγματος x_{test}, η συνάρτηση απόφασης είναι της μορφής:

$$f(x_{test}) = sgn((w \cdot x_{test}) + b) \qquad (3.1.2.1)$$

όπου το w αναπαριστά το διάνυσμα κατεύθυνσης του υπέρ-επιπέδου. Το πρόσημο της τιμής που επιστρέφεται από την (3.1.2.1) δείχνει την προβλεπόμενη κλάση με την οποία σχετίζεται το νέο δείγμα x_{test}, ενώ η ποσότητα |f(x_{test})| υποδεικνύει το βαθμό εμπιστοσύνης στην απόφαση που προέκυψε.

Για κάθε δείγμα εκπαίδευσης x_i , το SVM πρόβλημα μπορεί ισοδύναμα να διατυπωθεί όπως παρακάτω.

Ελαχιστοποίηση της ποσότητας :

$$\frac{1}{2} \cdot ||w||^2 + C \sum_{i=1}^{n} {\xi_\iota}^2$$

δεδομένου ότι:

$$y_i \left((w \cdot x_i) + b \right) \geq 1 - \xi_i , \xi_i \geq 0, \quad i = 1 \dots n \ . \qquad (3.1.2.2)$$

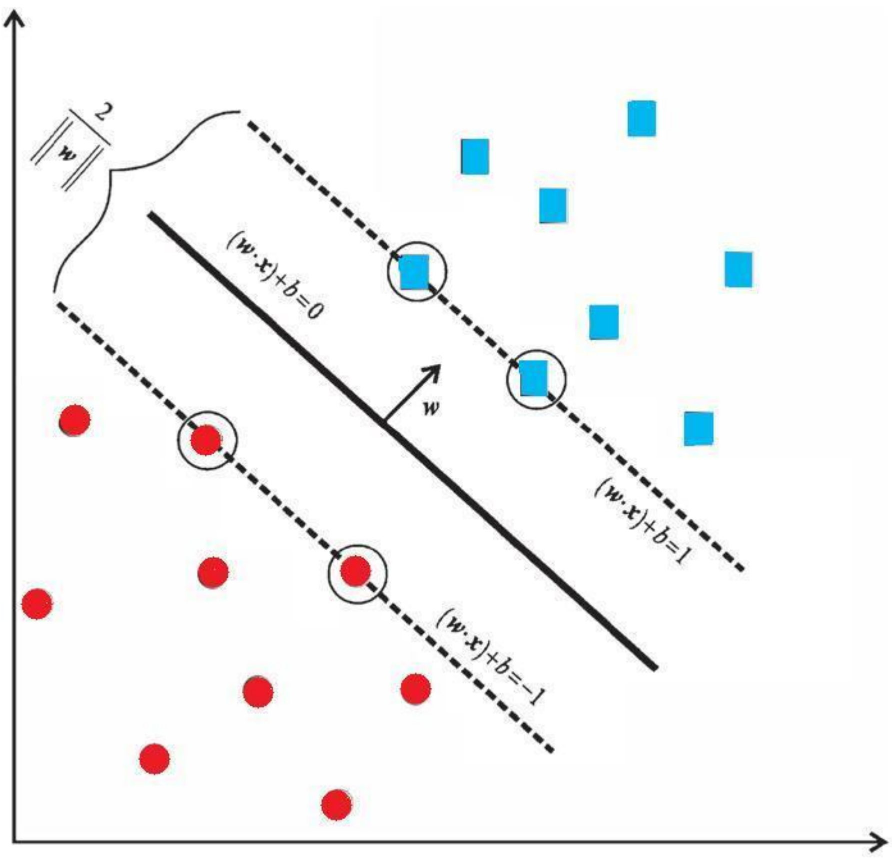
Από τη θεωρία της δυικότητας (*duality theory*), το πρόβλημα μπορεί να μετατραπεί στο ακόλουθο πρόβλημα μεγιστοποίησης, όπου το λ αντιπροσωπεύει το διάνυσμα των Lagrange πολλαπλασιαστών και το y_i αναπαριστά το label (είτε +1 είτε -1) του i-οστού δείγματος:

Μεγιστοποίηση της ποσότητας:

$$\sum_{i=1}^{n} \lambda_i - \sum_{\iota,\xi=1}^{n} \lambda_i \, \lambda_j \, y_i \, y_j \, (x_i \cdot x_j) \ , (\lambda \in R^n)$$

Δεδομένου ότι:

$$\sum_{i=1}^{n} \lambda_i \, y_i = 0 \; \kappa\alpha\iota \; 0 \leq \lambda_i \leq C \; \mu\varepsilon \; i = 1 \dots n \qquad (3.1.2.3)$$



Εικόνα 23: Επεξήγηση του προβλήματος της δυαδικής ταξινόμησης που δείχνει το περιθώριο διαχωρισμού ανάμεσα στις δύο κλάσεις. Τα κυκλωμένα εικονίδια αποτελούν τα support vectors [40].

Μέσω της λύσης αυτού του προβλήματος καταλήγουμε στο συμπέρασμα ότι το βέλτιστο (μέγιστο περιθώριο) υπερεπίπεδο περιγράφεται από τις ακόλουθες εξισώσεις, μόνο αν τα training samples x_i είναι support vectors. Αυτό προκύπτει από το γεγονός ότι αν λ_i=0, τότε το training sample x_i δεν έχει καμμία επιρροή στο υπερεπίπεδο και άρα στη συνάρτηση απόφασης. Οπότε, κάθε λ_i μη μηδενικό υποδεικνύει ότι το training sample x_i είναι support vector. Αν το πλήθος των support vectors είναι ίσο με m, τότε καταλήγουμε στην ακόλουθη έκφραση για το διάνυσμα κατεύθυνσης w:

$$w = \sum_{i=1}^{m} \lambda_i \ y_i \ x_i^{sv} \qquad (3.1.2.4)$$

Και επειδή τα support vectors βρίσκονται στα όρια των περιοχών των κλάσεων, όπως φαίνεται και από το σχήμα, αφού $y_i^{sv}\,(w \cdot x_i^{sv} + b) = 1$, μπορούν να χρησιμοποιηθούν για να βρεθεί ο παράγοντας b:

$$b = y_i^{sv} - w \cdot x_i^{sv} \qquad (3.1.2.5)$$

Ανασχηματίζοντας την εξίσωση (3.1.2.1) για τη συνάρτηση απόφασης για ένα νέο δείγμα, χρησιμοποιώντας και τους πολλαπλασιαστές Lagrange προκύπτει η εξίσωση:

$$f(x_{test}) = sgn\left(\left(\sum_{i=1}^{m} \lambda_i \; y_i \, x_i^{sv} \cdot x_{test}\right) + b\right) \qquad (3.1.2.6)$$

Επιπλέον, εφαρμόζοντας κάποιον πυρήνα διαφορετικό του γραμμικού η (3.1.2.7) μετασχηματίζεται στην:

Μεγιστοποίηση της ποσότητας:

$$\sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \, \lambda_j \, y_i \, y_j \, \varphi(x_i) \, \varphi(x_j) \;, (\lambda \in R^n)$$

Δεδομένου ότι:

$$\sum_{i=1}^{n} \lambda_i \, y_i = 0 \; \kappa\alpha\iota \; 0 \leq \lambda_i \leq C \; \mu\varepsilon \; i = 1 \dots n \qquad (3.1.2.8)$$

όπου $\varphi(x_i)$ ορίζεται ως: $R^d \rightarrow R^k$ με k>>n .

Ή ισοδύναμα οι εξισώσεις (3.1.2.3) και (3.1.2.6) μπορούν να γραφτούν:

Μεγιστοποίηση της ποσότητας:

$$\sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \, \lambda_j \, y_i \, y_j \, K(x_i \,, x_j) \;, (\lambda \in R^n)$$

Δεδομένου ότι:

$$\sum_{i=1}^{n} \lambda_i \, y_i = 0 \; \kappa\alpha\iota \; 0 \leq \lambda_i \leq C \; \mu\varepsilon \; i = 1 \dots n \qquad (3.1.2.9)$$

$$f(x_{test}) = sgn\left(\left(\sum_{i=1}^{m} \lambda_i \; y_i \, k(x_i^{sv}, x_{test})\right) + b\right) \qquad (3.1.2.10)$$

Μία σημαντική ιδιότητα των SVM είναι ότι μπορούν ταυτόχρονα να ελαχιστοποιούν το εμπειρικό σφάλμα ταξινόμησης και να μεγιστοποιούν το γεωμετρικό περιθώριο. Ως εκ τούτου, είναι γνωστοί και ως ταξινομητές μέγιστου περιθωρίου.

Τα SVM αποτελούν μία ελκυστική επιλογή για τη βιοπληροφορική εξαιτίας τριών βασικών ιδιοτήτων που εμφανίζουν:

- ✓ Η συνάρτηση απόφασής τους καθορίζεται από το υπερεπίπεδο εκείνο, που διαχωρίζει τις δύο κλάσεις με τον βέλτιστο τρόπο. Το βέλτιστο υπερεπίπεδο επιτυγχάνεται όταν τα όρια των δύο κλάσεων βρίσκονται όσο το δυνατό πιο μακριά μεταξύ τους (μέγιστο περιθώριο) και τοποθετείται στο μέσο αυτών των ορίων.

- ✓ Η συνάρτηση απόφασής τους καθορίζεται μόνο από διανύσματα και όχι από όλα τα δεδομένα. Αυτά τα διανύσματα, τα οποία τοποθετούνται πάνω στα όρια των δύο κλάσεων και κατ' επέκταση ορίζουν αυτά το όρια (περιθώρια), ονομάζονται διανύσματα υποστήριξης (support vectors).
- ✓ Η μη γραμμικότητα των δεδομένων μπορεί να αντιμετωπιστεί μέσω της εφαρμογής του 'kernel trick', η οποία απεικονίζει τα χαρακτηριστικά του χώρου εισόδου σε ένα χώρο χαρακτηριστικών υψηλότερων διαστάσεων, όπου τα δεδομένα είναι γραμμικά διαχωρίσιμα.

3.2 Ταξινομητές Βασισμένοι σε Πυρήνα

Τα support vector machines συνιστούν μία μέθοδο ταξινόμησης η οποία έχει ευρέως αναγνωριστεί ως η πιο ελπιδοφόρα στον τομέα της διάκρισης ποικίλων ασθενειών. Ωστόσο, υπάρχει μία πληθώρα άλλων ταξινομητών που βασίζονται σε πυρήνα και που συναγωνίζονται τους πρώτους σε απόδοση. Τρεις τέτοιοι ταξινομητές παρουσιάζονται σε αυτή την ενότητα.

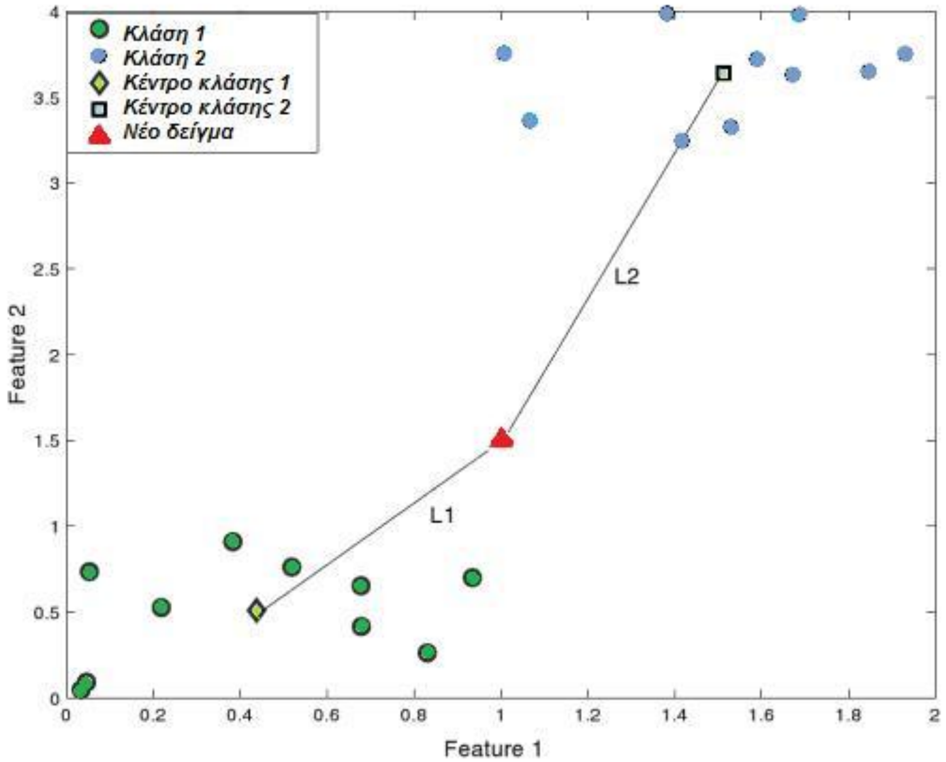
3.2.1 Μέθοδος kernel nearest centroid

Αυτός ο βασισμένος σε πυρήνα ταξινομητής [12], χρησιμοποιεί τα δεδομένα εκπαίδευσης για να υπολογίσει τα κέντρα μάζας των δύο κλάσεων, δηλαδή των δειγμάτων που ανήκουν στη θετική και των δειγμάτων που ανήκουν στην αρνητική κλάση, όπως δείχνει η εικόνα 24. Το νέο δείγμα x εκχωρείται στην κλάση της οποίας το κέντρο (*centroid*) είναι πιο κοντά (μικρότερη τετραγωνική απόσταση) στο δείγμα. Στην περίπτωση της δυαδικής ταξινόμησης με τις ετικέτες των δύο κλάσεων να είναι {+1,-1}, η συνάρτηση απόφασης ορίζεται ως εξής:

$$f(x) = \frac{1}{n_+} \sum_{\{i|y_i=+1\}} k(x, x_i) - \frac{1}{n_-} \sum_{\{i|y_i=-1\}} k(x, x_i) \; + b \qquad (3.2.1.1)$$

Όπου το b είναι το offset και μπορεί να εκφραστεί ως εξής:

$$b = \frac{1}{2}\left(\frac{1}{n_-^2} \sum_{\{(i,j)|y_i=y_j=-1\}} k(x_i, x_j) - \frac{1}{n_+^2} \sum_{\{(i,j)|y_i=y_j=+1\}} k(x_i, x_j) \right) \quad (3.2.1.2)$$



Εικόνα 24: Μία απλή επεξήγηση της kernel nearest centroid ταξινόμησης. Τα κέντρα της κλάσης με τους μπλε και πράσινους κύκλους, που εδώ αναπαριστώνται με τετράγωνο και ρόμβο αντίστοιχα, χρησιμοποιούνται για την ταξινόμηση του νέου δείγματος [44].

Έτσι, η κλάση του x προβλέπεται να είναι η +1 αν $f(x) \geq 0$ και -1 αλλιώς. Στην εικόνα 24, η απόσταση L1 είναι μικρότερη από την L2, οπότε το νέο δείγμα εκχωρείται στην κλάση, με τους πράσινους κύκλους.

3.2.2 Μέθοδος kernel minimum squared error machine

Ο μηχανισμός kernel minimum squared error [12] είναι μία έκδοση πυρήνα για το γραμμικό μηχανισμό ελαχίστου τετραγωνικού σφάλματος, που βασίζεται στον ψευδοανάστροφο πίνακα. Ορίζουμε με το σύμβολο $(\cdot)^T$ τον ανάστροφο και με το σύμβολο $(\cdot)^+$ τον ψευδοανάστροφο ενός πίνακα. Για την περίπτωση της δυαδικής ταξινόμησης με τις ετικέτες των δύο κλάσεων να είναι {+1,-1}, η διακρίνουσα συνάρτηση μπορεί να οριστεί ως εξής:

$$f(x) = \left(n\,\overrightarrow{1_n} + t(x)\right)^T \left(n^2\overrightarrow{1_n}\,\overrightarrow{1_n}^T + K\right)^+ b \qquad (3.2.2.1)$$

Όπου το $\overrightarrow{1_n}$ υποδηλώνει ένα n διαστάσεων διάνυσμα με όλες τις συνιστώσες του ίσες με $1/n$, το K είναι ο kernel matrix των δειγμάτων εκπαίδευσης και :

$$t(x) = \left(k(x, x_1), \dots, k(x, x_n)\right)^T \qquad (3.2.2.2)$$

Και το b δηλώνει ένα n διαστάσεων διάνυσμα που φέρει τις ετικέτες των δύο σχετικών κλάσεων. Τότε, η κλάση του x προβλέπεται να είναι η +1 αν $f(x) \geq 0$ και η -1 αλλιώς. Οπότε:

$$f(x) = sgn\left(\left(n\,\overrightarrow{1_n} + t(x)\right)^T \left(n^2\overrightarrow{1_n}\,\overrightarrow{1_n}^T + K\right)^+ b\right) \qquad (3.2.2.3)$$

3.2.3 Μέθοδος kernel subspace

Η μέθοδος kernel subspace [12] είναι μία μη γραμμική προέκταση της subspace μεθόδου. Προηγούμενες μελέτες σε ταξινόμηση καρκίνου σε microarray έδειξαν ότι η απόδοση του είναι συγκρίσιμη με αυτή των support vector machines. Η γενική μορφή αυτής της μεθόδου περιγράφεται παρακάτω.

Υποθέτουμε ότι το $s_{ki}(i = 1, \dots, n_k)$ υποδηλώνει τα δείγματα εκπαίδευσης της κλάσης k, και ορίζουμε το $g(x) \in R^{n_k}$ ως εξής:

$$g(x) = \left(k(x, s_{k1}), \dots, k\left(x, s_{kn_k}\right)\right)^T \qquad (3.2.3.1)$$

Το $\lambda_{ki}\ (i = 1, \dots, n_k)$ και το $a_{ki}\ (i = 1, \dots, n_k)$ είναι οι ιδιοτιμές του πίνακα $G = \left(\,g(s_{k1}), \dots, g\left(s_{kn_k}\right)\right) \in R^{n_k \times n_k}$ και τα ιδιοδιανύσματα που αντιστοιχούν στις ιδιοτιμές λ_{ki} , και r_k η διάσταση του υποχώρου (*subspace*) για την κλάση k. Οπότε, η διακρίνουσα συνάρτηση $f_k(x)$ για την κλάση k ορίζεται ως εξής:

$$f_k(x) = \ \|L_k^{-1}\, A_k^T\, g(x)\|^{\,2} \qquad (3.2.3.2)$$

Όπου,

$$L_k = diag\left(\sqrt{\lambda_{k1}}, \dots, \sqrt{\lambda_{kr_k}}\right) \tag{3.2.3.3}$$

$$A_k = \left(\alpha_{k1}, \dots, \alpha_{kr_k}\right) \tag{3.2.3.4}$$

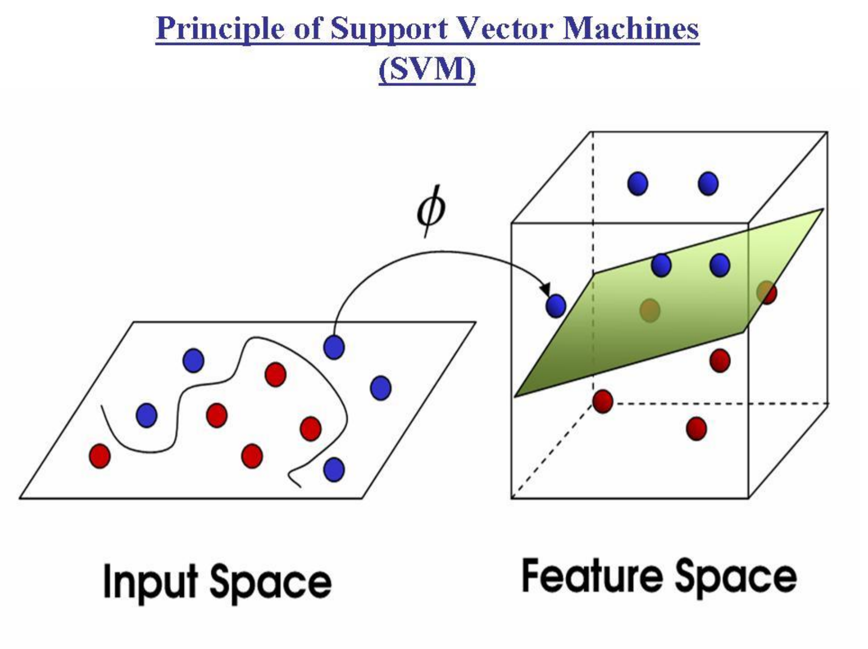
Τελικά, το νέο δείγμα x εκχωρείται στην κλάση με το μεγαλύτερο $f_k(x)$. Με σεβασμό στην επιλογή της διάστασης r_k , εφαρμόζουμε το αθροιστικό ποσοστό (*cumulative proportion*) το οποίο ορίζεται:

$$c.p. = \sum_{i=1}^{r_k} \lambda_{ki} \Big/ \sum_{i=1}^{n_k} \lambda_{ki} \tag{3.2.3.5}$$

Η διάσταση είναι καθορισμένη έτσι ώστε η αθροιστική αναλογία να μην υπερβαίνει μία προκαθορισμένη τιμή.

Όπως έχει ήδη προαναφερθεί στην ενότητα 2.3, οι μέθοδοι πυρήνα χρησιμοποιούνται για να μετασχηματίσουν το χώρο εισόδου σε ένα χώρο χαρακτηριστικών υψηλότερων διαστάσεων όπου ένας γραμμικός διαχωρισμός μπορεί να επιτευχθεί (εικόνα 25). Κάποιοι από τους πιο κοινούς πυρήνες [26] που χρησιμοποιούνται στους ταξινομητές οι οποίοι βασίζονται σε πυρήνα είναι οι ακόλουθοι:

Gaussian Radial Basis Function (RBF):	$K(x_i, x_j) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right)$, $\mu\varepsilon\ \gamma > 0\ \kappa\alpha\iota\ \kappa\acute{\alpha}\pi o\iota\varepsilon\varsigma\ \varphi o\rho\acute{\varepsilon}\varsigma\ \acute{\iota}\sigma o\ \mu\varepsilon\ \gamma = \frac{1}{2\sigma^2}$	(3.2.3.6)
Polynomial βαθμού p:	$K(x_i, x_j) = \left(c + (x_i \cdot x_j)\right)^p, \mu\varepsilon\ c \geq 0$	(3.2.3.7)
Quadratic:	$K(x_i, x_j) = (x_i \cdot x_j)\left(1 + (x_i \cdot x_j)\right)$	(3.2.3.8)
Polynomial βαθμού p για svm:	$K = (x_i \cdot x_j)$ $K = K(1 + (x_i \cdot x_j))$	(3.2.3.9)
Sigmoid:	$K(x_i, x_j) = \tanh(kx_i \cdot x_j + c)$, $\gamma\iota\alpha\ k > 0\ \kappa\alpha\iota\ c < 0$	(3.2.3.10)



Εικόνα 25: Μετασχηματισμός του χώρου εισόδου σε ένα χώρο χαρακτηριστικών, υψηλότερων διαστάσεων, όπου ένας γραμμικός διαχωρισμός είναι εφικτός [45].

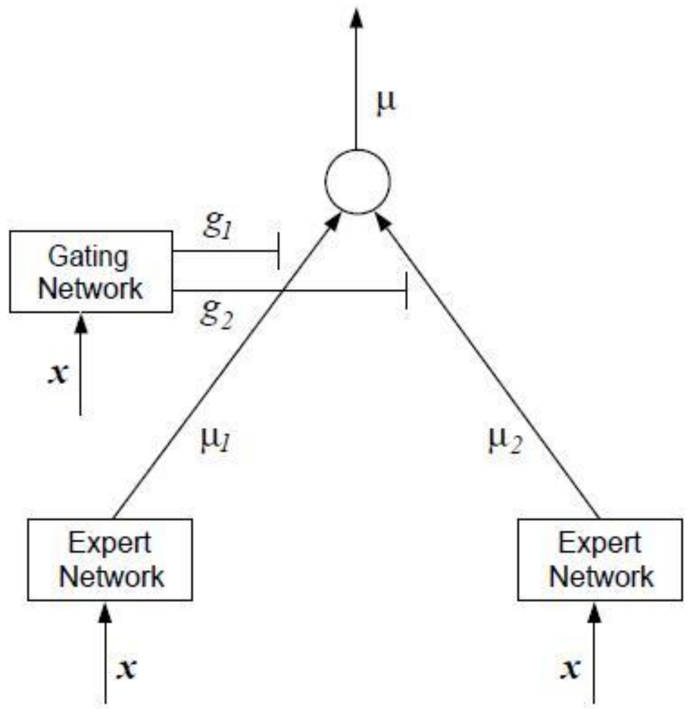
3.3 Μοντέλα Μείγματος Ειδικών

Τα ensembles είναι [46] σύνολα μηχανών μάθησης, των οποίων οι αποφάσεις συνδυάζονται για να βελτιωθεί η απόδοση του γενικού συνόλου. Στη βιβλιογραφία, υπάρχει μία πληθώρα όρων όπως committee, classifier fusion, combination, aggregation οι οποίοι χρησιμοποιούνται για να υποδηλώσουν σύνολα μηχανών μάθησης που εργάζονται μαζί για να λύσουν ένα πρόβλημα μηχανικής μάθησης. Αυτή η ποικιλία στους όρους και στις προδιαγραφές αντικατοπτρίζει την απουσία μίας ενοποιημένης θεωρίας για τις ensemble μεθόδους καθώς και το γεγονός ότι αυτή η περιοχή έρευνας είναι ακόμα νέα.

Η ανάπτυξη συστημάτων πολλαπλών ταξινομητών έχει λάβει ολοένα και αυξανόμενη προσοχή τελευταία, καθώς έχει διαπιστωθεί ότι τέτοια συστήματα μπορούν συχνά να είναι πιο εύρωστα, δηλαδή λιγότερο ευαίσθητα στους συντονισμούς των εσωτερικών παραμέτρων τους, σε ανακρίβειες, σε ελαττώματα των δεδομένων, και πιο ακριβή σε σχέση με έναν μόνο ατομικό ταξινομητή. Ωστόσο, δεν εξασφαλίζεται ότι τα συστήματα πολλαπλών ταξινομητών θα είναι πιο εύρωστα ή πιο ακριβή είτε από τους ατομικούς ταξινομητές που τα αποτελούν είτε από τον καλύτερο ατομικό ταξινομητή (*best single classifier*) που διαφορετικά θα μπορούσε να κατασκευαστεί. Όπως συχνά προκύπτει σε προβλήματα αναγνώρισης προτύπων όπου, για παράδειγμα, συμπεριλαμβάνονται άσχετα χαρακτηριστικά που συχνά μειώνουν την απόδοση ενός ταξινομητή, «το περισσότερο δεν είναι πάντα το καλύτερο».

Είναι γνωστό ότι η αποτελεσματικότητα των ensemble μεθόδων εξαρτάται από την ακρίβεια και την ποικιλία των ταξινομητών βάσης καθώς και από την ανεξαρτησία των σφαλμάτων που προκύπτουν από αυτούς, δηλαδή από το αν παρουσιάζουν χαμηλά ποσοστά σφάλματος ή από το αν παράγουν διαφορετικά σφάλματα.

Η προσέγγιση του μείγματος ειδικών (*mixture of experts*) [15], όπως φαίνεται στην εικόνα 26, είναι μία ensemble μέθοδος, στην οποία οι έξοδοι του συνόλου των ατομικών ταξινομητών, συνδυάζονται για να παραχθεί η τελική απόφαση ταξινόμησης σύμφωνα με κάποιο κανόνα συνυπολογισμού. Υπάρχουν δύο τύποι συνδυασμού των ειδικών: επιλογή ταξινομητών (*classifier selection*) και συγχώνευση ταξινομητών (*classifier fusion*). Στην πρώτη προσέγγιση κάθε ταξινομητής ειδικεύεται σε μία συγκεκριμένη περιοχή του χώρου χαρακτηριστικών και κάθε τοπικός ταξινομητής αποφασίζει μόνος του για την έξοδο του συνόλου. Στη δεύτερη προσέγγιση, όλη η πληροφορία του χώρου χαρακτηριστικών παρέχεται σε κάθε ατομικό ταξινομητή, οπότε όλοι οι ταξινομητές εκπαιδεύονται πάνω σε όλο το χώρο χαρακτηριστικών και οι έξοδοι τους συνδυάζονται. Έτσι, στην πρώτη περίπτωση υπάρχει συμπληρωματικότητα ανάμεσα στους ταξινομητές ενώ στη δεύτερη ανταγωνισμός.



Εικόνα 26: Μία αρχιτεκτονική μείγματος ειδικών ενός επιπέδου. Η έξοδος μ είναι η συνδυασμένη απόφαση του μείγματος των ειδικών, που προκύπτει από τις ατομικές αποφάσεις μ1 και μ2 [44].

3.3.1 Μείγμα ειδικών βασισμένο σε συγχώνευση με συνδυασμό ασαφών βαρών

Σε αυτό το μείγμα ειδικών [15], εφαρμόζεται η προσέγγιση της συγχώνευσης ταξινομητών, οπότε, οι ατομικοί ταξινομητές που συνδυάζονται είναι καθολικοί εμπειρογνώμονες. Αρχικά, όλοι οι ειδικοί εκπαιδεύονται ατομικά σε ολόκληρο το σύνολο δεδομένων εκπαίδευσης και έπειτα εκτελούν ταξινόμηση στο σύνολο δεδομένων ελέγχου. Επόμενο βήμα είναι ο συνδυασμός των αποφάσεών τους σε μία συνολική weighted averaged απόφαση.

Πιο συγκεκριμένα, αν υποθέσουμε ότι διαθέτουμε ένα σύνολο δεδομένων ελέγχου με k=1…N πλήθος δειγμάτων, ατομικούς ταξινομητές με πλήθος i=1...c και δύο κλάσεις ταξινόμησης j=1 ή j=2, τότε οι αποφάσεις των ειδικών συνδυάζονται ως εξής:

- Αρχικά, υπολογίζεται το κέντρο μάζας (μέσος όρος) και η διασπορά των δειγμάτων που ταξινομούνται από κάθε ταξινομητή στην πρώτη και στη δεύτερη κλάση σύμφωνα με τους τύπους:

$$v^{(i)(j)} = \sum_{k=1}^{N_j} u_{ik}^{(j)}\, x_k \Bigg/ \sum_{k=1}^{N_j} u_{ik}^{(j)} \tag{3.3.1.1}$$

$$s^{(i)(j)} = \sum_{k=1}^{N_j} u_{ik}^{(j)} \left[x_k - v^{(i)(j)}\right]^2 \Bigg/ \sum_{k=1}^{N_j} u_{ik}^{(j)} \tag{3.3.1.2}$$

Όπου $v^{(i)(j)}$ είναι το κέντρο μάζας των δειγμάτων που ο i-οστός ταξινομητής αποφάσισε να καταχωρήσει στην κλάση j, $s^{(i)(j)}$ είναι η διασπορά των δειγμάτων που ο i-οστός ταξινομητής αποφάσισε να καταχωρήσει στην κλάση j, x_k είναι το k-οστό δείγμα του συνόλου δεδομένων ελέγχου, N_j είναι το πλήθος των δειγμάτων που ο i-οστός ταξινομητής αποφάσισε ότι ανήκουν στην κλάση j, και $u_{ik}^{(j)}$ είναι η απόφαση (-1 ή +1) του i-οστού ταξινομητή για το δείγμα k το οποίο ταξινομήθηκε στην κλάση j.

- Στη συνέχεια, υπολογίζονται οι c συναρτήσεις πύλης (*gate functions*), μία για κάθε ταξινομητή, οι οποίες δρουν σαν μεσολαβητές ανάμεσα στους ειδικούς. Έτσι, για τον i-οστό ταξινομητή, για κάθε νέο δείγμα που είναι προς ταξινόμηση, υπολογίζονται οι αποστάσεις του από την πρώτη και τη δεύτερη κλάση γι’ αυτόν τον ταξινομητή. Για το σχηματισμό του i-οστού gate function επιλέγεται η μικρότερη από τις δύο αποστάσεις, δηλαδή στον παρακάτω τύπο, λόγω του αρνητικού εκθετικού που περιέχεται, επιλέγεται το μέγιστο των δύο ποσοτήτων.

$$A^{(i)}(x_{test}) = \max\left\{\exp\left[-\frac{1}{2}\sum_{p=1}^{t}\left(\frac{x_p - v_p^{(i)(1)}}{s_p^{(i)(1)}}\right)^2\right], \exp\left[-\frac{1}{2}\sum_{p=1}^{t}\left(\frac{x_p - v_p^{(i)(2)}}{s_p^{(i)(2)}}\right)^2\right]\right\} \tag{3.3.1.3}$$

Κάνοντας διερεύνηση επί των αποτελεσμάτων, προκύπτουν οι εξής περιπτώσεις τιμών για τους τύπους (3.3.1.1), (3.3.1.2) και (3.3.1.3).

- ✓ Αν κάποιος ταξινομητής καταχωρήσει όλα τα δείγματα στη μία κλάση, τότε το κέντρο μάζας για την άλλη κλάση, όπως και η αντίστοιχη διασπορά προκύπτουν ίσα με μηδέν.
- ✓ Αν κάποιος ταξινομητής καταχωρήσει ένα μόνο δείγμα στη μία κλάση και όλα τα άλλα στην άλλη κλάση, τότε το κέντρο των δειγμάτων γι' αυτή την κλάση ισούται με το ίδιο το δείγμα, οπότε η αντίστοιχη διασπορά είναι μηδέν.
- ✓ Αν δύο ή περισσότερα δείγματα, που έχουν εκχωρηθεί σε μία κλάση από έναν ταξινομητή, είναι ίσα μεταξύ τους, τότε το κέντρο αυτών των δειγμάτων ισούται με ένα από αυτά και η διασπορά τους είναι μηδέν.
- ✓ Αν και οι c ταξινομητές έχουν ταξινομήσει τα δείγματα του συνόλου ελέγχου με τον ίδιο ακριβώς τρόπο στις δύο κλάσεις, τότε τα κέντρα και οι διασπορές των δειγμάτων και για τους c ταξινομητές είναι ίδια, άρα και τα c gate functions είναι ίδια (ιδανική περίπτωση όπου όλοι οι ταξινομητές συμφωνούν).
- ✓ Οι αποστάσεις που σχηματίζονται από τον τύπο (3.3.1.3) και αποτελούν το όρισμα στη συνάρτηση του αρνητικού εκθετικού, λαμβάνουν πολύ υψηλές τιμές λόγω του αθροίσματος και της τετραγωνικής δύναμης που υπάρχει στον τύπο. Όμως, για πολύ μεγάλα ορίσματα, η συνάρτηση του αρνητικού εκθετικού γίνεται ασύμπτωτη στο μηδέν με αποτέλεσμα και οι δύο ποσότητες του τύπου (3.3.1.3) και άρα και όλα τα gate functions να προκύπτουν ίσα με μηδέν.

Οπότε, για να υπάρχει ακρίβεια στις τιμές που λαμβάνουν τα gate functions, εκτελείται μία παραλλαγή του τύπου (3.3.1.3).

- Λαμβάνοντας υπόψη τη γραφική παράσταση της συνάρτησης του αρνητικού εκθετικού, πλέον αναζητείται ο ελάχιστος των δύο εκθετών κατά αντιστοιχία στο μέγιστο των δύο ποσοτήτων των τιμών της εκθετικής συνάρτησης.

$$A^{(i)}(x_{test}) = min\left\{\sum_{p=1}^{t}\left(\frac{x_p - v_p^{(i)(1)}}{s_p^{(i)(1)}}\right)^2, \sum_{p=1}^{t}\left(\frac{x_p - v_p^{(i)(2)}}{s_p^{(i)(2)}}\right)^2\right\} \qquad (3.3.1.4)$$

- Η συνολική έξοδος για κάθε δείγμα $x_{test,k}$, λαμβάνεται από το σταθμισμένο μέσο όρο (*weighted average*):

$$y_k = \sum_{i=1}^{c} A^{(i)}(x_{test,k}) u_{ik} \Big/ \sum_{i=1}^{c} A^{(i)}(x_{test,k}) \qquad (3.3.1.5)$$

Υλοποιείται μία παραλλαγή του παραπάνω τύπου θέλοντας να γίνει ενίσχυση του καλύτερου ταξινομητή δηλαδή εκείνου που αντιστοιχεί στο μικρότερο gate function. Κάτι τέτοιο επιτυγχάνεται με την προσθήκη κατάλληλων βαρών για κάθε ταξινομητή με την εξής λογική:

- Αν a_i είναι ένα από τα c gate functions, τότε υπολογίζεται το $\hat{a} = \max\{a_i\}$ και έπειτα τα $b_i = \frac{a_i}{\hat{a}}$ με $0 \leq b_i \leq 1$. Έπειτα, υπολογίζονται τα $b_{i_{new}} = 1.1 - b_i$ (και όχι $b_{i_{new}} = 1 - b_i$ γιατί έτσι χάνεται τελείως το μεγαλύτερο από τα gate functions) μετά το $\hat{b} = sum\{b_{i_{new}}\}$ και τέλος τα $w_i = \frac{b_{i_{new}}}{\hat{b}}$. Άρα το μικρότερο a_i αντιπροσωπεύεται στον τύπο (3.3.1.5) από το μεγαλύτερο w_i .
- Οπότε, η (3.3.1.5) μετασχηματίζεται στην:

$$y_k = \sum_{i=1}^{c} w_i(x_{test,k})\, u_{ik} \qquad (3.3.1.6)$$

Αν $y_k \geq 0$ τότε το $x_{test,k}$ εκχωρείται στην κλάση +1, αλλιώς το $x_{test,k}$ εκχωρείται στην κλάση -1. Αυτός ο ταξινομητής ονομάζεται ταξινομητής μείγματος ειδικών.

3.3.2 Μείγμα ειδικών βασισμένο σε συγχώνευση με gradient descent

Αυτός ο συνδυασμός των ειδικών [14, 47] χρησιμοποιεί την προσέγγιση της συγχώνευσης ταξινομητών. Αν υποθέσουμε ότι το πλήθος των ατομικών ταξινομητών που συνδυάζονται είναι K και το πλήθος των γονιδίων κάθε δείγματος του συνόλου εκπαίδευσης είναι M, τότε η διαδικασία μάθησης συνοψίζεται ως εξής:

- Αρχικά, εκχωρούνται τυχαίες τιμές βαρών w_{ki} και a_{ik} με k=1…K και i=1…M, για τους ταξινομητές και για τα gate functions, μόνο στα γονίδια του πρώτου δείγματος του συνόλου εκπαίδευσης. Αυτό το βήμα γίνεται μόνο μία φορά στην αρχή του αλγορίθμου και δεν επαναλαμβάνεται ξανά όσο αυτός εκτελείται.
- Αν το δείγμα εισόδου είναι το (x,d), με x να είναι ένα διάνυσμα μεγέθους M, τότε το d είναι το αντίστοιχο classification target με τιμές 0 ή 1.
- Κάθε ειδικός υπολογίζει την έξοδο του από τον τύπο:

$$v_k(t) = w_{ki}x_i(t) \tag{3.3.2.1}$$

$$y_k(t) = logsig(v_k(t)) \tag{3.3.2.2}$$

- Για τα gate functions, μία softmax activation συνάρτηση είναι απαραίτητη:

$$g_k = \left. e^{u_k} \middle/ \sum_{j=1}^{K} e^{u_k} \right. \tag{3.3.2.3}$$

Όπου $u_k = \alpha_k^T x_i$, και k=1,2,…K.

- Η συνολική έξοδος του μείγματος ειδικών για το δείγμα συνδυάζει τα gate functions και τις εξόδους των ατομικών ταξινομητών:

$$y = \sum_{k=1}^{K} g_k y_k \tag{3.3.2.4}$$

Όπου k=1,2,…K.

- Σύμφωνα με την τιμή του target d, το σφάλμα e ορίζεται:

$$e = d - y \tag{3.3.2.5}$$

- Η εκ των υστέρων πιθανότητα ορίζεται ως εξής:

$$h = \frac{g_i y^d (1-y)^{1-d}}{\sum_{j=1} g_j y_j^d \left(1 - y_j\right)^{1-d}} \tag{3.3.2.6}$$

Όπου g_i είναι ο μέσος όρος των δύο gate functions και j=1,2,…K.

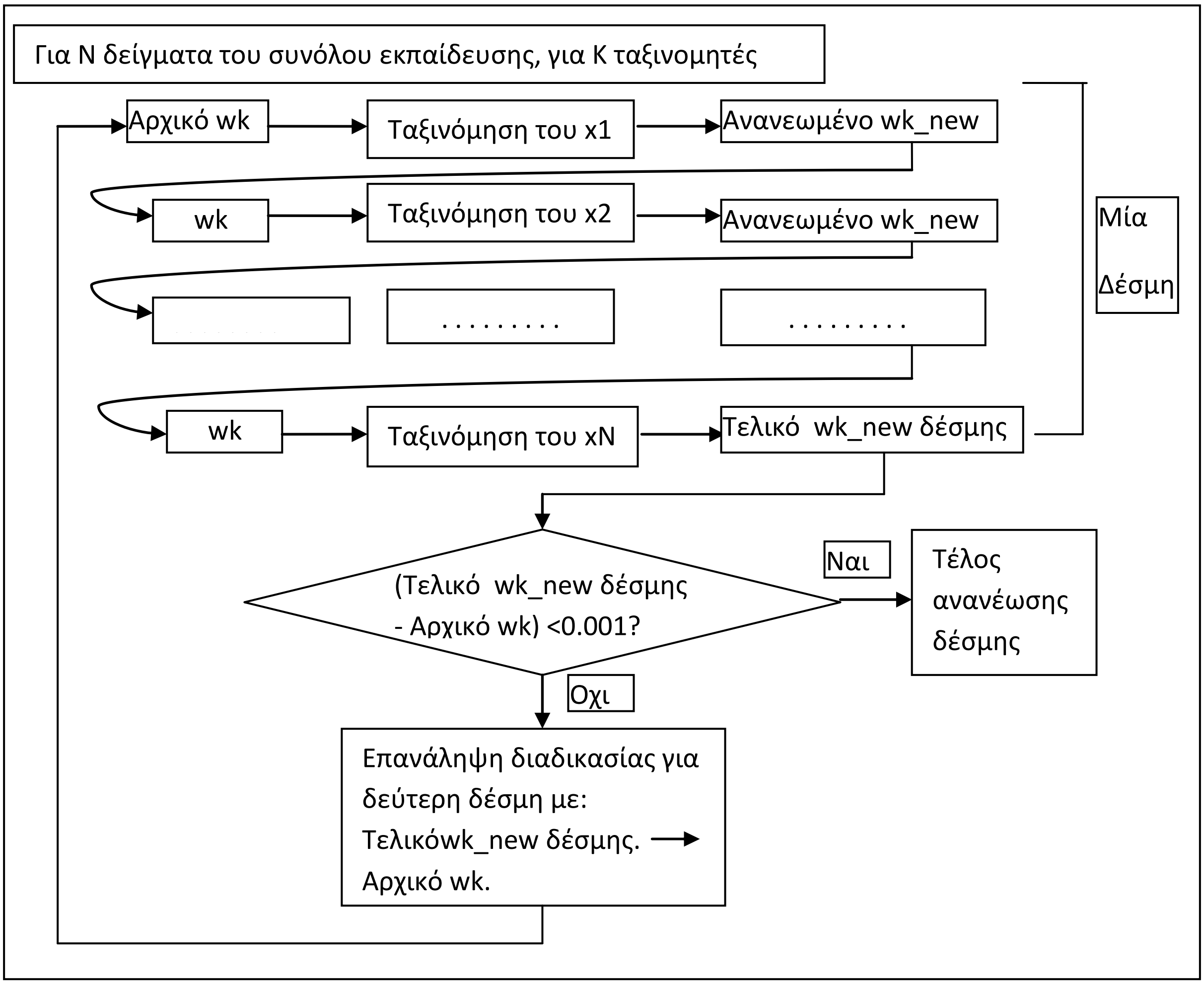
- Τα βάρη w_{ki} και a_{ik}των ταξινομητών και των gate functions για όλα τα γονίδια του δείγματος ανανεώνονται από τους ακόλουθους κανόνες:

$$w_i(t+1) = w_i(t) + \eta h(t) e(t) x_i \tag{3.3.2.7}$$

$$a_i(t+1) = a_i(t) + \eta(h(t) - g_i(t)) x_i \tag{3.3.2.8}$$

Όπου η είναι ο ρυθμός μάθησης. Μετά από την πρώτη ανανέωση των βαρών των γονιδίων του πρώτου δείγματος, εκτελείται ανανέωση δέσμης (*batch update*), δηλαδή επαναλαμβάνεται ολόκληρη η παραπάνω διαδικασία, μόνο από το βήμα δύο και κάτω, για όλα τα υπόλοιπα δείγματα του συνόλου εκπαίδευσης, τα οποία θεωρούμε ότι αποτελούν μία δέσμη, με κάθε δείγμα από το δεύτερο και μετά να παίρνει σαν αρχικό βάρος το ανανεωμένο βάρος του προηγούμενου δείγματος.

- Όταν η ανανέωση των βαρών εκτελεστεί για όλα τα δείγματα του συνόλου εκπαίδευσης, τα βάρη που έχουν προκύψει μετά την ανανέωση και του τελευταίου δείγματος, αποτελούν την πρώτη ολοκληρωμένη ανανέωση των βαρών των ταξινομητών και των gate functions. Στη συνέχεια ελέγχεται αν αυτά τα βάρη θα ανανεωθούν εκ νέου, για τη δέσμη όλων των δειγμάτων, συγκρίνοντας αν απέχουν λιγότερο από 0,001 από το αρχικό βάρος του πρώτου δείγματος. Αν ναι, τότε η διαδικασία μάθησης τερματίζεται. Αν όχι, τότε η παραπάνω διαδικασία επαναλαμβάνεται ξανά για όλη τη δέσμη, με το βάρος του τελευταίου δείγματος να εισάγεται σαν αρχικό βάρος στο πρώτο



Εικόνα 27: Η διαδικασία της ανανέωσης βαρών με batch update για το μείγμα ειδικών που βασίζεται συγχώνευση ταξινομητών με gradient descent.

δείγμα της επόμενης δέσμης, ενώ στο τέλος της ανανέωσης για όλη τη δέσμη γίνεται πάλι ο ίδιος έλεγχος. Επίσης, ο ρυθμός μάθησης η μειώνεται. Οι ανανεώσεις στα βάρη σταματούν όταν οι επαναλήψεις γίνουν 1000 ή όταν το βάρος του τελευταίου δείγματος από το βάρος του πρώτου δείγματος, της ίδιας πάντα δέσμης, απέχει λιγότερο από 0,001. Όλη η διαδικασία της ανανέωσης βαρών με batch update φαίνεται στην εικόνα 27.

Όσο η διαδικασία μάθησης λαμβάνει χώρα, τα βάρη w_{ki} των ταξινομητών και τα βάρη a_{ik} των gate functions ανανεώνονται συνεχώς, βασισμένα τόσο στο σφάλμα της απόφασης της ταξινόμησης του μείγματος ειδικών για όλα τα δείγματα του συνόλου εκπαίδευσης, όσο και στην εκ των υστέρων πιθανότητα τους.

Αφού ολοκληρωθεί η διαδικασία μάθησης, τα εν λόγω βάρη έχουν εκπαιδευτεί πάνω σε όλα τα δείγματα του συνόλου εκπαίδευσης, πολλαπλές φορές, έχουν σταθεροποιήσει τις τιμές τους και μπορούν να χρησιμοποιηθούν για την ταξινόμηση των νέων δειγμάτων του συνόλου ελέγχου.

Ο αλγόριθμος για την ταξινόμηση των νέων δειγμάτων του συνόλου ελέγχου περιλαμβάνει τα βήματα 3-5 της διαδικασίας μάθησης, δηλαδή τις εξισώσεις (3.3.2.1) έως (3.3.2.4) για κάθε δείγμα εισόδου x_{test}, και με w_{ki} και a_{ik} να αποτελούν τα τελικά βάρη της διαδικασίας μάθησης για τους ταξινομητές και τα gate functions. Επειδή η έξοδος των ατομικών ταξινομητών προέρχεται από μία logsig συνάρτηση, οι τιμές που προκύπτουν από αυτήν βρίσκονται στο διάστημα [0,1]. Οπότε, για την απόφαση του μείγματος ειδικών, μετά την εφαρμογή του τύπου (3.3.2.4), εφαρμόζεται ένα κατώφλι σύμφωνα με το οποίο αν $y < 0.5$ τότε το δείγμα x_{test} εκχωρείται στην κλάση -1, ενώ αν $y \geq 0.5$ τότε το δείγμα x_{test} εκχωρείται στην κλάση +1.

3.3.3 Μείγμα ειδικών βασισμένο σε συγχώνευση με συνδυασμό ασαφών βαρών και gradient descent

Αυτός ο συνδυασμός των ειδικών χρησιμοποιεί την προσέγγιση της συγχώνευσης ταξινομητών και αποτελεί μίξη της μεθόδου που περιγράφηκε στην ενότητα 3.3.1 και αυτής που αναφέρθηκε στην ενότητα 3.3.2.

Πιο συγκεκριμένα, χρησιμοποιεί logsig ταξινομητές οι οποίοι εκπαιδεύονται και ανανεώνουν τα βάρη τους βασισμένοι στο σφάλμα της δικής τους απόφασης ταξινόμησης. Έπειτα, οι αποφάσεις τους συνδυάζονται σύμφωνα με τον fuzzy mixture. Η διαδικασία μάθησης για κάθε ταξινομητή είναι η ακόλουθη:

- Αρχικά, εκχωρούνται τυχαίες τιμές βαρών w_{ki} και a_{ik} με k=1…K και i=1…M, για τους ταξινομητές και για τα gate functions, μόνο στα γονίδια του πρώτου δείγματος του συνόλου εκπαίδευσης. Αυτό το βήμα γίνεται μόνο μία φορά στην αρχή του αλγορίθμου και δεν επαναλαμβάνεται ξανά όσο αυτός εκτελείται.

- Αν το δείγμα εισόδου είναι το (x,d), με x να είναι ένα διάνυσμα μεγέθους M, τότε το d είναι το αντίστοιχο classification target με τιμές 0 ή 1.
- Ο κάθε ειδικός υπολογίζει την έξοδο του από τον τύπο:

$$v_k(t) = w_{ki}x_i(t) \tag{3.3.3.1}$$

$$y_k(t) = logsig(v_k(t)) \tag{3.3.3.2}$$

- Σύμφωνα με την τιμή του target d, το σφάλμα e του κάθε ταξινομητή ορίζεται

$$e_k = d - y_k \tag{3.3.3.3}$$

- Τα βάρη w_{ki} των ταξινομητών για όλα τα γονίδια του δείγματος ανανεώνονται από τον ακόλουθο κανόνα:

$$w_{ki}(t+1) = w_{ki}(t) + \eta e_k(t)x_i \tag{3.3.3.4}$$

Όπου η είναι ο ρυθμός μάθησης. Μετά από την πρώτη ανανέωση των βαρών των γονιδίων του πρώτου δείγματος, εκτελείται ανανέωση δέσμης (*batch update*), όπως αυτή περιγράφεται στην ενότητα 3.3.2. Όσο η διαδικασία μάθησης λαμβάνει χώρα, τα βάρη w_{ki} των ταξινομητών ανανεώνονται συνεχώς, βασισμένα στο σφάλμα της δικής τους απόφασης ταξινόμησης για όλα τα δείγματα του συνόλου εκπαίδευσης, με στόχο, τα εν λόγω βάρη να εκπαιδευτούν πάνω σε όλα τα δείγματα του συνόλου εκπαίδευσης, πολλαπλές φορές, και να σταθεροποιήσουν, με αυτό τον τρόπο, τις τιμές τους για να μπορέσουν να χρησιμοποιηθούν για την ταξινόμηση των νέων δειγμάτων του συνόλου ελέγχου.

Ο αλγόριθμος για την ταξινόμηση των νέων δειγμάτων του συνόλου ελέγχου είναι ο εξής:

- Για κάθε δείγμα εισόδου x_{test} από το σύνολο δεδομένων ελέγχου, ο κάθε ειδικός υπολογίζει την έξοδο του από τον τύπο:

$$v_k(t) = w_{ki} x_{test}(t) \tag{3.3.3.5}$$

$$y_k(t) = logsig(v_k(t)) \tag{3.3.3.6}$$

- Επειδή η έξοδος των ατομικών ταξινομητών προέρχεται από μία logsig συνάρτηση, οι τιμές που προκύπτουν από αυτήν βρίσκονται στο διάστημα [0,1]. Οπότε, για την απόφαση τους, εφαρμόζεται ένα κατώφλι σύμφωνα με το οποίο, αν $y_k < 0.5$ τότε το δείγμα x_{test} εκχωρείται από τον k-οστό ταξινομητή στην κλάση -1, ενώ αν $y_k \geq 0.5$ τότε το δείγμα x_{test} εκχωρείται στην κλάση +1.

Στη συνέχεια, αφού έχουν σχηματιστεί οι αποφάσεις των ατομικών ταξινομητών, ακολουθείται η διαδικασία που περιγράφεται στην ενότητα 3.3.1, για το συνδυασμό τους σε μία weighted averaged απόφαση.

3.3.4 Μείγμα ειδικών βασισμένο σε εκπαίδευση ειδικών σε διαφορετικά υποσύνολα δεδομένων

Αυτός ο συνδυασμός των ειδικών [14] χρησιμοποιεί την προσέγγιση της επιλογής ταξινομητών και βασίζεται στη λογική αποστάσεων. Το πρώτο πράγμα που γίνεται εδώ είναι ο διαχωρισμός του συνόλου δεδομένων εκπαίδευσης σε περιοχές (ομάδες), τόσες όσες ο αριθμός των ειδικών και άρα και των gate functions που θα υλοποιηθούν. Εργαζόμαστε προς αυτή την κατεύθυνση εκτελώντας μία τεχνική της μάθησης χωρίς επίβλεψη που ονομάζεται ομαδοποίηση των δεδομένων και υλοποιείται με το self organized map (*SOM*).

- Αρχικά, εκτελώντας ομαδοποίηση, γίνεται διαχωρισμός των δειγμάτων του συνόλου δεδομένων εκπαίδευσης σε ομάδες (*clusters*) με τη χρήση του SOM και προκύπτουν οι ομάδες με τα κέντρα τους. Τα δείγματα εκπαίδευσης είναι διανύσματα σε ένα πολυδιάστατο χώρο χαρακτηριστικών, με καθένα από αυτά να διαθέτει ετικέτα που υποδηλώνει την κλάση του.

  Ο αλγόριθμος μάθησης περιλαμβάνει τα εξής βήματα:

  - Εκχώρηση τυχαίων τιμών βαρών σε κάθε κόμβο w_{ij} του δικτύου.
  - Για ένα δείγμα εισόδου x, υπολογίζεται η απόσταση ανάμεσα στο δείγμα x και κάθε νευρώνα j, ο οποίος αναπαρίσταται από το βάρος

w_j, και ορίζεται ο νικητής νευρώνας ως $argmin_j \{\|x - w_j\|\}$, όπου $\|\cdot\|$ η ευκλείδεια απόσταση.

- Ανανεώνονται τα βάρη της γειτονιάς του νικητή νευρώνα ως εξής: $w_{ij}(t+1) = w_{ij}(t) + \eta(t)K(j,t)\{x_i(t) - w_{ij}(t)\}$, όπου το η(t) είναι ο ρυθμός μάθησης στην εποχή t και K(j,t) είναι μία κατάλληλη συνάρτηση γειτονιάς, γκαουσιανής φύσης.
- Επαναλαμβάνονται τα βήματα 2 και 3 μέχρι να συμβεί μία σύγκλιση, για παράδειγμα μέχρι οι απόλυτες τετραγωνικές αλλαγές να είναι μικρότερες από 0.02 στις πάνω από 2500 εποχές.

Για κάθε δείγμα υπολογίζεται η ευκλείδεια απόσταση του από όλα τα κέντρα και τελικά, αυτό εκχωρείται στην ομάδα από το κέντρο της οποίας απέχει λιγότερο. Με αυτό τον τρόπο καταχωρούνται όλα τα δείγματα στις ομάδες.

- Στη συνέχεια αποδίδονται σχετικά βάρη στα δείγματα κάθε ομάδας σύμφωνα με το τύπο:

$$wf_{jc_i} = d'(w_{j\,c_i} - c_i) \Big/ \sum_{j=1\ to\ |c_i|} d'(w_{j\,c_i} - c_i) \tag{3.3.4.1}$$

Όπου το $d'(w_{j\,c_i}, c_i)$ είναι η απόσταση κάθε δείγματος από το κέντρο της ομάδας c_i στην οποία ανήκει και υπολογίζεται ως εξής:

$$d'(w_{j\,c_i}, c_i) = \|w_{j\,c_i} - c_i\|^{-1} \tag{3.3.4.2}$$

Όπου το $w_{j\,c_i}$ είναι το j-οστό δείγμα της ομάδας i και c_i είναι το κέντρο της ομάδας i. Αυτά τα σχετικά βάρη εκφράζουν το βαθμό στον οποίο κάθε δείγμα ανήκει στην ομάδα του, δηλαδή τα δείγματα που βρίσκονται στην άκρη της ομάδας έχουν μικρότερο βάρος σε σχέση με αυτά που βρίσκονται πιο κοντά στο κέντρο της.

Αφού τα δείγματα του συνόλου εκπαίδευσης έχουν μοιραστεί σε ομάδες, το επόμενο βήμα είναι η υλοποίηση της αρχιτεκτονικής του μείγματος ειδικών για την ταξινόμηση των νέων δειγμάτων του συνόλου ελέγχου. Κάθε ταξινομητής ειδικεύεται τοπικά, δηλαδή γίνεται ειδικός (*local expert*) μόνο σε μία από τις ομάδες και άρα εκπαιδεύεται μόνο με τα δείγματα που ανήκουν στην ομάδα του, χωρίς να «βλέπει» αυτά που ανήκουν στις υπόλοιπες.

- Για κάθε νέο δείγμα εισόδου x_{test}, κάθε ειδικός εκτελεί ταξινόμηση σύμφωνα με τη λογική του κοντινότερου γείτονα. Δηλαδή, υπολογίζει την ευκλείδεια απόσταση του x_{test} από όλα τα δείγματα της ομάδας του, βρίσκει το δείγμα, της ομάδας του, που απέχει λιγότερο από το x_{test} και τελικά εκχωρεί το x_{test} στην κλάση στην οποία ανήκει το κοντινότερο δείγμα που μόλις βρήκε.

Αφού καθένας από τους ατομικούς ταξινομητές αποφασίσει για την κλάση στην οποία ανήκει το x_{test}, βασιζόμενος μόνο στα δείγματα εκπαίδευσης της δικής του ομάδας, σειρά έχει η δημιουργία των gate functions για το σχηματισμό του μείγματος των ειδικών.

- Ο συνδυασμός των αποφάσεων των ατομικών ταξινομητών γίνεται από τις gate functions ως εξής:

$$gate_i(x_{test}) = \sum_{j=1}^{|c_i|} wf_{jc_i} \times d(x_{test}, w_{j\,c_i}) \qquad (3.3.4.3)$$

  Το $gate_i$ υπολογίζει το άθροισμα των αποστάσεων του κάθε x_{test} από κάθε δείγμα της ομάδας i πολλαπλασιασμένων με το σχετικό βάρος του αντίστοιχου δείγματος.

- Η απόφαση του μείγματος των ειδικών υπολογίζεται από τον τύπο:

$$y = \sum_{k=1}^{K} gate_k\, y_k \qquad (3.3.4.4)$$

  Όπου K=1...k το πλήθος των ταξινομητών και των gate functions και y_k η απόφαση του k ταξινομητή για το εκάστοτε δείγμα ελέγχου.

Κάνοντας διερεύνηση επί των αποτελεσμάτων, προκύπτουν οι εξής περιπτώσεις τιμών:

- ✓ Τα gate functions είναι θετικά πάντα αφού εκφράζουν αποστάσεις και τα y_k παίρνουν τις τιμές -1 ή +1.

- ✓ Αν οι ατομικοί ταξινομητές συμφωνούν, τα y_k είναι ίδια, δηλαδή έχουν και τα δύο την τιμή -1 ή και τα δύο την τιμή +1, τότε $y_k < 0$ ή $y_k > 0$ αντίστοιχα, άρα το νέο δείγμα του συνόλου ελέγχου x_{test} εκχωρείται στην αρνητική ή στη θετική κλάση αντίστοιχα, χωρίς οι τιμές των δύο gate functions να επηρεάζουν αυτή την απόφαση.
- ✓ Αν οι ατομικοί ταξινομητές διαφωνούν και τα y_k έχουν διαφορετικές τιμές μεταξύ τους, τότε συγκρίνουμε τα gate functions, και η απόφαση του μείγματος ειδικών υιοθετεί την απόφαση του ταξινομητή με το μικρότερο gate function. Οπότε, το νέο δείγμα x_{test} εκχωρείται σύμφωνα με τη λογική των αποστάσεων στην κλάση που υπέδειξε ο ταξινομητής εκείνος που ειδικεύεται στην ομάδα στην οποία βρίσκεται πιο κοντά το δείγμα x_{test}.

Σε αυτή την προσέγγιση μείγματος ειδικών, εφαρμόζεται η λογική των αποστάσεων τόσο στη δημιουργία των ομάδων SOM όσο και στην απόφαση των επιμέρους ταξινομητών αλλά και του μείγματος των ειδικών.

3.4 Αναδρομική Εξάλειψη Χαρακτηριστικών

Το πρόβλημα στην επιλογή δεικτών στην ανάλυση μικροσυστοιχιών DNA [41], μπορεί να αντιμετωπιστεί με δύο βασικούς τύπους μεθόδων, τις λεγόμενες filter και wrapper μεθόδους. Οι wrapper μέθοδοι λειτουργούν με αναδρομικό τρόπο, όπου τα βάρη των χαρακτηριστικών (γονιδίων) επανεκτιμώνται και αλλάζουν δυναμικά από επανάληψη σε επανάληψη, ενώ στις filter μεθόδους τα βάρη των χαρακτηριστικών παραμένουν σταθερά. Ο αλγόριθμος RFE-LNW που παρουσιάζεται αναλυτικά σε αυτή την ενότητα, είναι μία μέθοδος που ενσωματώνει τις δύο αυτές μεθόδους. Εφαρμόζει τα filter κριτήρια με έναν αναδρομικό τρόπο, όπου τα βάρη ενδεχομένως ρυθμίζονται από κύκλο σε κύκλο, παράγοντας αξιοσημείωτη βελτίωση στην μετρούμενη απόδοση της γενίκευσης σε ανεξάρτητα σύνολα ελέγχου.

3.4.1 Διαφορικά εκφρασμένα γονίδια

Η βασική ιδέα πίσω από την ανάπτυξη της μεθόδου για την αναδρομική εξάλειψη γονιδίων [41], είναι η αναγνώριση και τελικά η επιλογή διαφορικά εκφρασμένων γονιδίων, ιδέα που δεν είναι καινούρια στο πεδίο της επιλογής δεικτών (*marker selection*). Στις μελέτες με filter μεθόδους επιλογής, χρησιμοποιούνται παραλλαγές των συντελεστών Fisher που δίνονται από την ακόλουθη εξίσωση:

$$f_1(g_i) = \frac{(\mu_+(g_i) - \mu_-(g_i))^2}{\sigma_+(g_i)^2 + \sigma_-(g_i)^2} \qquad (3.4.1.1)$$

Μία παραλλαγή των συντελεστών Fisher μπορεί να εκφραστεί ως εξής:

$$f_2(g_i) = \frac{\sum_{j=1}^{n} \left| g_{ij} - c(g_i) \right|}{\sigma_+(g_i) + \sigma_-(g_i)} \qquad (3.4.1.2)$$

Όπου $\mu_+(g_i)$, $\mu_-(g_i)$, $\sigma_+(g_i)\ και\ \sigma_-(g_i)$ είναι οι μέσες τιμές και οι τυπικές αποκλίσεις των εκφράσεων του γονιδίου g_i στη θετική και αρνητική κλάση αντίστοιχα, n είναι ο αριθμός των δειγμάτων και το c ισούται με:

$$c(g_i) = \frac{(\mu_+(g_i) + \mu_-(g_i))}{2} \qquad (3.4.1.3)$$

Μία άλλη παραλλαγή, με χαμηλό υπολογιστικό κόστος δίνεται από την εξίσωση:

$$f_3(g_i) = \frac{|\mu_+(g_i) - \mu_-(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \qquad (3.4.1.4)$$

Εύκολα διαπιστώνουμε ότι οι εξισώσεις (3.4.1.1), (3.4.1.2) και (3.4.1.4) εκφράζουν στην ουσία την ίδια ιδέα. Όταν χρησιμοποιούνται αυτές οι εξισώσεις για την εκχώρηση βαρών, σε ένα σύνολο από δεδομένα γονίδια, είναι προφανές ότι στα γονίδια που διαφοροποιούν περισσότερο την έκφρασή τους στις δύο κλάσεις, εκχωρούνται υψηλότερες τιμές βαρών από ότι σε εκείνα που σημειώνουν μικρότερη διαφοροποίηση ανάμεσα στις δύο κλάσεις. Στα γονίδια τα οποία εκφράζονται με τον ίδιο ακριβώς τρόπο και στις δύο κλάσεις, εκχωρείται το μικρότερο δυνατό βάρος που είναι ίσο με το μηδέν. Στις επόμενες υποενότητες χρησιμοποιείται αυτό το μέτρο με ένα wrapper τρόπο, ενσωματωμένο κατάλληλα στην διαδικασία μάθησης των γραμμικών νευρώνων και των μηχανών διανυσμάτων υποστήριξης (SVM) με σκοπό να συνδράμει στο πεδίο της επιλογής δεικτών.

3.4.2 Η μέθοδος RFE-LNW

Οι περισσότερες μέθοδοι επιλογής δεικτών [41], που εφαρμόζονται στο πεδίο των DNA μικροσυστοιχιών (*microarray*), εξαιτίας των πολλών διαστάσεων των δεδομένων, χρησιμοποιούν γραμμικά εργαλεία για να αξιολογήσουν το πρόβλημα. Το RFE-SVM είναι μία τέτοια μέθοδος, όπου ένας γραμμικός πυρήνας χρησιμοποιείται για να εκτιμήσει το διάνυσμα βαρών του διαχωριστικού υπέρ-επιπέδου, η απόλυτη τιμή του οποίου χρησιμοποιείται μετά για το κριτήριο κατάταξης των γονιδίων (*gene ranking*).

Από την άλλη πλευρά, λόγω του σχεδιασμού του (ένας γραμμικός συνδυασμός των εισόδων), ένας γραμμικός νευρώνας μπορεί επίσης να προσεγγίσει κάθε γραμμική συνάρτηση. Έτσι, χρησιμοποιείται ένας τέτοιος γραμμικός νευρώνας (*linear neuron*) για να προσεγγίσουμε το διαχωριστικό υπέρ-επίπεδο ανάμεσα στις θετικές και αρνητικές κλάσεις. Εκμεταλλευόμενοι μία τέτοια ανοιχτή αρχιτεκτονική, μπορούμε να επιλέξουμε ανάμεσα σε μία ποικιλία από σχήματα εκμάθησης ή εύκολα να ενσωματώσουμε μία νέα διαδικασία μάθησης κατάλληλα προσαρμοσμένη στο υπογραμμισμένο πρόβλημα, ενώ μπορούμε να το επεκτείνουμε και άλλο σε multilayer ή multi-class σχηματισμούς. Αυτό εφαρμόζεται σε ένα δίκτυο αποτελούμενο από έναν νευρώνα (*single neuron network*) με m εισόδους, όπως αυτός που φαίνεται στην εικόνα 28, όπου το m αντιστοιχεί στον αριθμό των γονιδίων. Μελετώντας δύο πιθανά αποτελέσματα στο επίπεδο εξόδου (*output layer*), δηλαδή έξοδος 0 για την αρνητική κλάση και έξοδος 1 για τη θετική κλάση, μπορούμε να χρησιμοποιήσουμε έναν τέτοιο γραμμικό νευρώνα για να προσεγγίσουμε το διαχωριστικό υπέρ-επίπεδο που διακρίνει τις δύο κλάσεις ενδιαφέροντος (θετική και αρνητική).

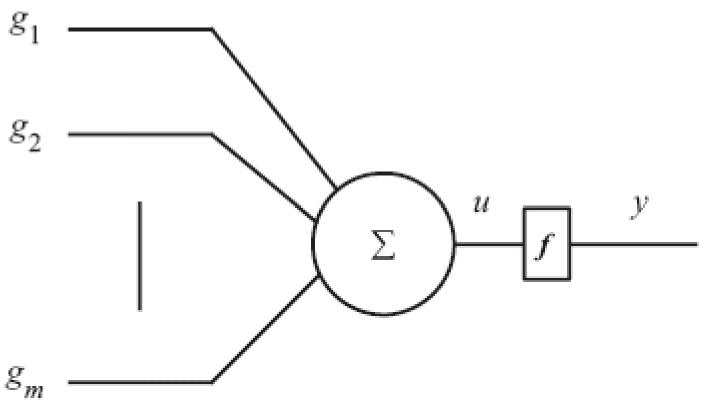
Πιο συγκεκριμένα, χρησιμοποιώντας τη sigmoid συνάρτηση f(u) λαμβάνουμε:

$$y = \frac{1}{1 + e^{u}} = f(u) \qquad (3.4.2.1)$$

$$u = \sum_{i=1}^{m} w_i \, g_i \qquad (3.4.2.2)$$

$$f'(u) = y(1 - y) \qquad (3.4.2.3)$$

Σημειώνουμε ότι $f'(u) \geq 0$, αφού το y εκτείνεται από 0 έως 1.



Εικόνα 28: Ένας μονός νευρώνας προσαρμοσμένος στο πρόβλημα της επιλογής δεικτών [40].

3.4.3 Εκπαιδεύοντας τον RFE-LNW

Το βασικό μαθηματικό υπόβαθρο [41] της διαδικασίας μάθησης που χρησιμοποιείται για την ανανέωση των βαρών του γραμμικού νευρώνα είναι το εξής:

Η συνάρτηση σφάλματος ενός μονού νευρώνα που πρέπει να ελαχιστοποιηθεί δίνεται από την εξίσωση:

$$E = \frac{1}{2} \sum_{j=1}^{n} (d_j - y_j)^2 \qquad (3.4.3.1)$$

όπου το n αντιστοιχεί στον αριθμό των δειγμάτων, το d_j αναπαριστά την επιθυμητή έξοδο του νευρώνα για το δείγμα j και το y_j είναι η πραγματική έξοδος που παράγεται από αυτόν το νευρώνα για το δεδομένο δείγμα. Μέσω της gradient descent μεθόδου, για την ελαχιστοποίηση της εξίσωσης (3.4.3.1), ενημερώνουμε το βάρος w_i που σχετίζεται με το γονίδιο g_i ακολούθως:

$$w_i(t+1) = w_i(t) - \left(\mu \frac{\partial E}{\partial w_i} \right) = w_i(t) - \mu \sum_{j=1}^{n} \left(\frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial u} \frac{\partial u}{\partial w_i} \right) \qquad (3.4.3.2)$$

Επηρεάζοντας την ανανέωση μέσω του μέτρου Fisher από την εξίσωση (3.4.1.2) προκύπτει:

$$w_i(t+1) = w_i(t) - \frac{\mu}{2} \sum_{j=1}^{n} \left[\left(\frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial u_j} \frac{\partial u_j}{\partial w_i} \right) \frac{|g_{ij} - c(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \right]$$

$$= w_i(t) - \frac{\mu}{2} \sum_{j=1}^{n} \left(-2(d_j - y_j) y_j (1 - y_j) g_{ij} \right) f_2(g_i)$$

$$= w_i(t) + \mu \sum_{j=1}^{n} (d_j - y_j) y_j (1 - y_j) g_{ij} f_2(g_i)$$

$$= w_i(t) + \mu \sum_{j=1}^{n} (d_j - y_j) f'(u_j) g_{ij} f_2(g_i) \qquad (3.4.3.3)$$

Τελικά,

$$w_i(t+1) = w_i + \mu \sum_{j=1}^{n} e_j f'(u_j) g_{ij} \frac{|g_{ij} - c(g_i)|}{\sigma_+(g_i) + \sigma_-(g_i)} \qquad (3.4.3.4)$$

Όπου το t αναπαριστά την τρέχουσα επανάληψη, το μ είναι ο ρυθμός μάθησης (*learning rate*) και

$$e_j = (d_j - y_j) \qquad (3.4.3.5)$$

Δουλεύοντας με πρόσημα, μία ιδέα που παρουσιάζεται στο ελαστικό back-propagation learning η εξίσωση (3.4.3.4) γράφεται ισοδύναμα:

- ✓ Στην περίπτωση όπου $f_2(g_i) = 1$, το οποίο είναι παρόμοιο με την πρότυπη back-propagation διαδικασία, λαμβάνουμε:

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{n} e_j f'(u_j)\, g_{ij} \qquad (3.4.3.6)$$

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{n} sign\left(e_j f'(u_j)\right) sign(g_{ij}) \qquad (3.4.3.7)$$

- ✓ Ή γενικά:

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{n} sign\left(e_j\, f'(u_j)\right) \times sign(\, g_{ij}\,)\, f_2(g_i) \qquad (3.4.3.8)$$

$$w_i(t+1) = w_i(t) + \sum_{j=1}^{n} \left|d_j - y_j\right| sign\left(e_j\, f'(u_j\,)\right) \times sign(\, g_{ij}\,) f_2(g_i) \qquad (3.4.3.9)$$

Η εξίσωση (3.4.3.6) είναι ο βασικός gradient descent learning αλγόριθμος, αποδεδειγμένος να οδηγεί τη συνάρτηση σφάλματος (3.4.3.1) σε ένα ελάχιστο. Επίσης, η εξίσωση (3.4.3.7) συγκλίνει, αφού κρατώντας το πρόσημο του gradient οδηγούμαστε προς την κατεύθυνση του ελαχίστου, στο οποίο τελικά θα φτάσουμε (με εξαίρεση τις περιπτώσεις που παγιδευόμαστε σε ένα τοπικό ελάχιστο) χρησιμοποιώντας ένα κατάλληλο ρυθμό μάθησης. Στην πραγματικότητα, περιμένουμε ότι η εξίσωση (3.4.3.7) να συγκλίνει πιο γρήγορα από την (3.4.3.6), αφού το e_j και το $f'(u_j)$ στην (3.4.3.6) μπορούν να πάρουν πολύ μικρές τιμές με αποτέλεσμα χαμηλές μεταβολές του w, που συνεπάγεται χαμηλή μεταβολή της συνάρτησης σφάλματος και επιβράδυνση της σύγκλισης.

Παίρνοντας μόνο το πρόσημο του gradient στην εξίσωση (3.4.3.7) κατευθυνόμαστε προς την κατεύθυνση του ελαχίστου κάνοντας «μεγαλύτερα σταθερά βήματα» και επιταχύνοντας τη σύγκλιση τουλάχιστον όταν η διαδικασία είναι μακριά από το ελάχιστο. Αυτή η εξαγωγή είναι πολύ χρήσιμη στην εφαρμογή της στη διαδικασία επιλογής δεικτών, αφού αναγκάζει τον αλγόριθμο να συγκλίνει πολύ γρήγορα, ειδικότερα στα πρώτα βήματα της διαδικασίας όπου ο αριθμός των χαρακτηριστικών (γονίδια) είναι εξαιρετικά μεγάλος.

Η εξίσωση (3.4.3.8) διαφέρει από την (3.4.3.7) μόνο στο συντελεστή $f_2(\cdot)$. Ακολουθώντας τον ίδιο συλλογισμό, η (3.4.3.8) μπορεί να αποδειχθεί ότι συγκλίνει σε ένα ελάχιστο, αλλά επιπρόσθετα λαμβάνει υπόψη και τελικά μετρά, μέσω του όρου άθροισης, μία προσέγγιση του μέτρου Fisher. Οι λίγες διαστάσεις των δεδομένων, ωστόσο, μπορεί να καθυστερήσουν τη σύγκλιση. Καθώς η διαδικασία προχωρά και το πρόβλημα των διαστάσεων μειώνεται σημαντικά, ο λόγος των δειγμάτων προς τα χαρακτηριστικά αυξάνεται και το πρόβλημα της εκτίμησης του διαχωριστικού υπέρ-επιπέδου γίνεται πιο δύσκολο, καθυστερώντας τη σύγκλιση. Κάτι τέτοιο κάνει απαραίτητη την αύξηση είτε του αριθμού των εποχών (επαναλήψεων), είτε του ρυθμού μάθησης.

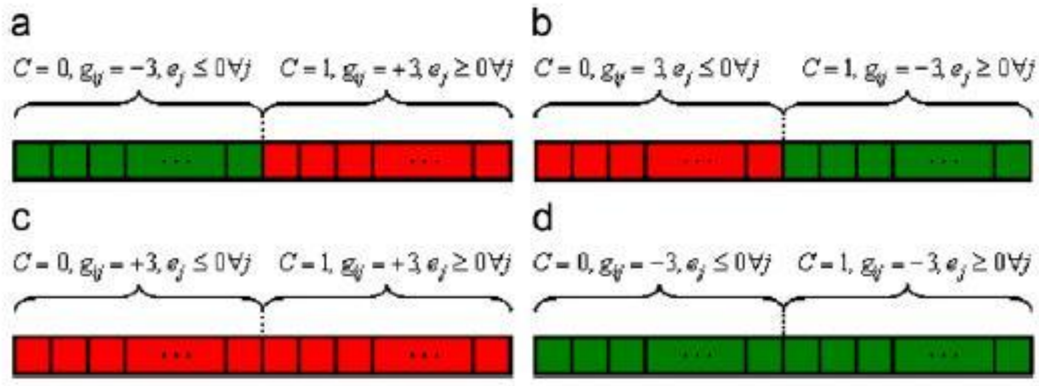
Σε αυτά τα τελευταία βήματα, η εξίσωση (3.4.3.9) μπορεί να χρησιμοποιηθεί για επιταχύνει τη σύγκλιση χρησιμοποιώντας ένα παραλλαγμένο ρυθμό μάθησης $| d_j - y_j |$. Είναι εύκολο να φανεί ότι όσο είμαστε μακριά από το στόχο (*target*), η ποσότητα $| d_j - y_j |$ θα παίρνει μία μεγάλη τιμή, αυξάνοντας τη σύγκλιση, αλλά όσο προσεγγίζουμε το στόχο, η ποσότητα $| d_j - y_j |$ θα αρχίζει να παίρνει χαμηλότερες τιμές, επιβραδύνοντας έτσι τη σύγκλιση. Δηλαδή, στα τελευταία βήματα της διαδικασίας της επιλογής χαρακτηριστικών, κατευθυνόμαστε προς το στόχο γρήγορα όταν είμαστε μακριά ακόμα από αυτόν, αλλά επιβραδύνουμε όταν τον προσεγγίζουμε για το μικροσυντονισμό του διαχωριστικού υπερεπιπέδου. Η εξίσωση (3.4.3.8) σε συνδυασμό με την (3.4.3.9) είναι οι τελικοί κανόνες ανανέωσης βαρών που χρησιμοποιούνται για την εκχώρηση των βαρών, ενώ οι εξισώσεις (3.4.3.6) και (3.4.3.7) χρησιμοποιούνται πιο πολύ για επεξήγηση και για λόγους αιτιολόγησης.

Συνοπτικά, από τα παραπάνω φαίνεται ότι εκπαιδεύοντας έναν νευρώνα με μία κατάλληλη διαδικασία μάθησης, μπορούμε τελικά να εφαρμόσουμε ένα filter κριτήριο, όπως το Fisher's ratio, σε ένα wrapper τρόπο.

3.4.4 Υπογραμμίζοντας τα διαφορικά εκφρασμένα γονίδια

Η επιλογή διαφορικά εκφρασμένων γονιδίων [41] είναι ένας επιθυμητός στόχος σε κάθε προσέγγιση επιλογής δεικτών. Έτσι, χρειαζόμαστε να αποδείξουμε ότι ανάμεσα στον εξαιρετικά μεγάλο αριθμό από γονίδια, που τροφοδοτεί τη μέθοδο RFE-LNW, θα μπορέσουμε τελικά να επιλέξουμε δείκτες που εκφράζονται με διαφορετικό τρόπο ανάμεσα στις δύο κλάσεις ενδιαφέροντος.

Η εικόνα 29 δείχνει το επίπεδο έκφρασης ενός υποθετικού γονιδίου g_i στην αρνητική (C=0) και στη θετική (C=1) κλάση, αντίστοιχα.



Εικόνα 29: Διαφορικά εκφρασμένα έναντι μη διαφορικά εκφρασμένων γονιδίων [40].

Στις περιπτώσεις (a) και (b) το υποθετικό γονίδιο εκφράζεται διαφορετικά στις δύο κλάσεις ενδιαφέροντος, με πράσινο (αρνητικές τιμές) στην αρνητική κλάση και με κόκκινο (θετικές τιμές) στη θετική κλάση ή αντίστροφα. Από την άλλη πλευρά, οι περιπτώσεις (c) και (d) δε δείχνουν καμία διαφοροποίηση στο επίπεδο έκφρασης του συγκεκριμένου γονιδίου στις δύο καταστάσεις ενδιαφέροντος. Εξετάζοντας την περίπτωση (a), σε συνδυασμό με την εξίσωση (3.4.3.7), και εστιάζοντας στην αρνητική κλάση (πράσινο τμήμα), παρατηρούμε ότι ο όρος $sign\left(e_j\, f'(u_j)\right) sign(g_{ij}) \ge 0$ διατηρείται. Πράγματι, ισχύει ότι $e_j \ge 0$, (αφού $d_j = 0$ από την εξίσωση (21), και $y_j \in [0 \dots 1]$), και το $f'(u_j)$ από την εξίσωση (3.4.3.5) είναι θετικό και $g_{ij} = -3$.

Τώρα εστιάζοντας στη θετική κλάση (κόκκινο τμήμα) της εικόνας 22.a και χρησιμοποιώντας τον ίδιο συλλογισμό, παρατηρούμε πάλι ότι $sign\left(e_j\, f'(u_j)\right) sign(g_{ij}) \ge 0$. Αφού ο όρος e_j , σε ένα ρεαλιστικό σενάριο, είναι πιο συχνά μη μηδενικός, ο όρος άθροισης της εξίσωσης (3.4.3.7) παράγει ένα θετικό αποτέλεσμα. Ακολουθώντας περίπου την ίδια λογική για την περίπτωση (b) της εικόνας 22, μπορεί να φανεί ότι η εξίσωση (3.4.3.7) παράγει ένα αρνητικό αποτέλεσμα, ενώ ο όρος άθροισης στις περιπτώσεις (c) και (d) παράγει αποτελέσματα κοντά στο μηδέν αφού οι τιμές των όρων στις δύο κλάσεις αυτοαναιρούνται.

Αν τώρα εξετάσουμε τις απόλυτες τιμές των εκχωρημένων βαρών, διαπιστώνουμε ότι τα διαφορετικά εκφρασμένα γονίδια (περιπτώσεις (a) και (b)) λαμβάνουν μεγαλύτερες τιμές από τα γονίδια που δε διαφοροποιούν τις εκφράσεις τους (περιπτώσεις (c) και (d)) στις δύο καταστάσεις ενδιαφέροντος. Αντίθετα, παρατηρούμε ότι η εξίσωση (3.4.3.6) δε μπορεί να παράγει το ίδιο αποτέλεσμα, αφού εξαρτάται από την τιμή και όχι από το πρόσημο του όρου e_j , το οποίο ενδεχομένως να γίνει πολύ μικρό και να μειώσει το αναμενόμενο αποτέλεσμα.

Από την άλλη πλευρά, η εξίσωση (3.4.3.7) είναι άδικη στα διαφορικά εκφρασμένα γονίδια, αφού λαμβάνοντας υπόψη μόνο την τιμή του πρόσημου, στα γονίδια που εκφράζονται πιο διαφορετικά θα εκχωρηθεί το ίδιο βάρος με εκείνα τα γονίδια που εκφράζονται λιγότερο διαφορικά. Μία πιο δίκαιη λύση περιλαμβάνει την εκχώρηση μεγαλύτερων βαρών στα περισσότερο διαφορετικά εκφρασμένα γονίδια, κάτι το οποίο επιτυγχάνεται με την εξίσωση (3.4.3.8) όπου ο όρος μεγέθους $f_2(g_i)$ παρουσιάζεται.

3.4.5 Σταδιακή μάθηση έναντι μάθησης δέσμης

Οι εξισώσεις (3.4.3.6) έως (3.4.3.9) ανανεώνουν [41] τον παράγοντα του βάρους w(t+1) μετά από την εμφάνιση όλων των δειγμάτων στο δίκτυο, δηλαδή αφού έχουν υπολογιστεί οι όροι των αθροισμάτων. Αυτό αναφέρεται ως εκπαίδευση δέσμης (*batch training*) στη θεωρία των νευρωνικών δικτύων. Εναλλακτικά, μπορούμε να ανανεώσουμε τα βάρη σταδιακά, εξετάζοντας ένα δείγμα τη φορά, οπότε οι όροι της άθροισης μπορούν να παραλειφθούν. Για τονίσουμε αυτή τη στρατηγική ανανέωσης, η παραπάνω ομάδα εξισώσεων παρουσιάζεται μετασχηματισμένη ως εξής:

$$w_i(t+1) = w(t) + \mu e f'(u) g_i \tag{3.4.5.1}$$

$$w_i(t+1) = w(t) + \mu \cdot sign(e \cdot f'(u)) \cdot sign(g_i) \tag{3.4.5.2}$$

$$w_i(t+1) = w(t) + \mu \cdot sign(e \cdot f'(u)) \cdot sign(g_i) \cdot f_2(g_i) \tag{3.4.5.3}$$

$$w_i(t+1) = w(t) + |d-y| \cdot sign(e \cdot f'(u)) \cdot sign(g_i) \cdot f_2(g_i) \tag{3.4.5.4}$$

Στον αλγόριθμο RFE-LNW, τα βάρη ανανεώνονται με σταδιακό τρόπο, αφού στο συγκεκριμένο τομέα, έχει αποδειχτεί ότι παράγει καλύτερα αποτελέσματα από ότι η ανανέωση βαρών με batch training. Σημειώνουμε ότι η εξίσωση (3.4.5.4) εφαρμόζεται από το σημείο των 100 επιζώντων γονιδίων και κάτω μέχρι το τέλος της διαδικασίας.

3.4.6 Αλγοριθμική αναπαράσταση του RFE-LNW

Ο πίνακας 1 παρέχει μία αλγοριθμική σύνοψη της RFE-LNW μεθόδου [41], όπως αυτή περιγράφεται στις παραπάνω υποενότητες. Σημειώνουμε ότι εκτός από το πλεονέκτημα της χρήσης ενός μονού νευρώνα ως το μόνο συστατικό μάθησης, καταφέρνει επίσης να εφαρμόζει filter κριτήρια σε ένα πραγματικό wrapper τρόπο, όπου τα βάρη επανεκτιμώνται και δυνητικά προσαρμόζονται από επανάληψη σε επανάληψη.

Πράγματι, εξαλείφοντας γονίδια μειώνονται οι διαστάσεις του προβλήματος και έτσι, το νέο εκτιμημένο υπερεπίπεδο επαναϋπολογίζεται σε ένα νέο μειωμένο χώρο, με ένα νέο διάνυσμα κατεύθυνσης w. Στις εφαρμογές του πραγματικού κόσμου φαίνεται κατάλληλο τα βάρη των γονιδίων να αλλάζουν από επανάληψη σε επανάληψη, αφού ένας μεγάλος χώρος χαρακτηριστικών, με πολλά μη σημαντικά γονίδια, μπορεί να επισκιάσει την επιρροή των πραγματικά σημαντικών, τα οποία γίνονται πιο σχετικά, όσο οι διαστάσεις του προβλήματος μειώνονται. Σημειώνουμε τη διαφορά με τη filter μέθοδο όπου το μέτρο του Fisher, για τα επιζώντα γονίδια, παραμένει σταθερό σε ολόκληρη τη φάση της διαδικασίας της επιλογής χαρακτηριστικών και το ίδιο συμβαίνει και με το διαχωριστικό όριο που αυτό καθορίζει.

Πίνακας 1: Η αλγοριθμική παρουσίαση του RFE-LNW.

1. Έστω m ο αρχικός αριθμός των χαρακτηριστικών (γονίδια).
2. Όσο ($m \geq 0$)
3. Ανανέωσε το διάνυσμα βάρους w χρησιμοποιώντας την εξίσωση (3.4.5.3) και (3.4.5.4). (Η εξίσωση (3.4.5.3) χρησιμοποιείται όσο ο αριθμός των επιζώντων χαρακτηριστικών είναι μεγαλύτερος από 100,αλλιώς χρησιμοποιείται η εξίσωση (3.4.5.4)).
4. Κατάταξε τα χαρακτηριστικά σύμφωνα με τις απόλυτες τιμές του διανύσματος w.
5. Αφαίρεσε το χαρακτηριστικό με το μικρότερο βάρος σε απόλυτη τιμή ($m \leftarrow m-1$). Περισσότερα από ένα χαρακτηριστικά μπορούν να διαγραφούν σε κάθε επανάληψη.
6. Εκτίμησε την ακρίβεια της ταξινόμησης των m επιζώντων γονιδίων χρησιμοποιώντας έναν γραμμικό svm ταξινομητή.
7. Τέλος όσο.
8. Εξήγαγε σαν marker γονίδια το σύνολο των επιζώντων γονιδίων που επιτυγχάνουν

3.5 Μέτρα Εκτίμησης της Απόδοσης Μεθόδων

Η εκτίμηση της απόδοσης ενός ταξινομητή [24] εξαρτάται σε μεγάλο βαθμό από τα χαρακτηριστικά των δεδομένων που είναι προς ταξινόμηση, αλλά και από τα χαρακτηριστικά του ταξινομητή, ένα από τα οποία είναι ο πυρήνας. Δεν υπάρχει ατομικός ταξινομητής που να μπορεί να γενικεύσει καλά σε όλα τα δοθέντα προβλήματα. Διάφοροι εμπειρικοί έλεγχοι εκτελούνται για να συγκρίνουν την απόδοση της ταξινόμησης και να βρουν τα χαρακτηριστικά των δεδομένων που καθορίζουν την απόδοση του ταξινομητή. Ο καθορισμός ενός κατάλληλου ταξινομητή για ένα δεδομένο πρόβλημα είναι, ωστόσο, ακόμα τέχνη και όχι επιστήμη.

Οι αλγόριθμοι μάθησης [48] που βασίζονται σε πυρήνα (*kernel based*) είναι αρθρωτά συστήματα που σχηματίζονται από ένα στοιχείο μάθησης γενικού σκοπού και από μία συνάρτηση πυρήνα συγκεκριμένου προβλήματος. Είναι θέμα ζωτικής σημασίας, για την απόδοση του συστήματος, η συνάρτηση του πυρήνα να προσαρμόζεται κάπως στο στόχο εκμάθησης (*learning target*). Αυτό σημαίνει ότι στο χώρο των χαρακτηριστικών, η κατανομή των δεδομένων σχετίζεται με κάποιο τρόπο με την κατανομή των ετικετών (*label*). Πολλά αποτελέσματα δείχνουν ότι η γενίκευση λαμβάνει χώρα μόνο όταν ένας τέτοιος συσχετισμός υφίσταται, και πολλοί κλασσικοί εκτιμητές της απόδοσης μπορούν να κατανοηθούν εκτιμώντας αυτή τη σχέση. Με άλλα λόγια, η επιλογή ενός πυρήνα, σε αυτή την τάξη συστημάτων, ισοδυναμεί στη μηχανική μάθηση με προβλήματα της κλασσικής επιλογής χαρακτηριστικών και μοντέλου. Οπότε, μετρώντας το βαθμό συμφωνίας ανάμεσα σε έναν πυρήνα και σε μία δεδομένη συνάρτηση στόχου (*target function*), είναι ένα σημαντικό πρόβλημα τόσο για εννοιολογικούς όσο και για πρακτικούς λόγους.

3.5.1 Εκτίμηση της απόδοσης των ταξινομητών

Ο τελικός στόχος της διαδικασίας της ταξινόμησης είναι να μπορούν οι ταξινομητές να προβλέπουν με όσο το δυνατό πιο υψηλή ακρίβεια, αποφεύγοντας παράλληλα το φαινόμενο του overfitting, την κλάση των νέων δειγμάτων. Όπως ήδη αναφέρθηκε, ο ταξινομητής εκπαιδεύεται πάνω στα δεδομένα εκπαίδευσης και μόνο μετά το τέλος της διαδικασίας μάθησης χρησιμοποιούνται τα δεδομένα ελέγχου για να μετρηθεί η ακρίβεια της ταξινόμησης. Το φαινόμενο του overfitting μπορεί να προκύψει όταν ο ταξινομητής προσαρμόζεται τόσο πολύ στα δεδομένα εκπαίδευσης, που αδυνατεί να γενικεύσει καλά σε νέα αθέατα δεδομένα, με αποτέλεσμα η ακρίβεια του να είναι χαμηλή. Δυστυχώς, στην πραγματική ζωή και στην ιατρική είναι συχνά δύσκολο να έχουμε στη διάθεση μας ένα μεγάλο σύνολο δεδομένων, όπως επίσης και κάποια σωστά δεδομένα ελέγχου. Για το λόγο αυτό χρησιμοποιούνται μέθοδοι όπως το Cross Validation (CV), που εξηγείται στην ενότητα (2.4), η οποία χωρίζει το σύνολο δεδομένων σε ένα μεγάλο τμήμα που αποτελεί τα δεδομένα εκπαίδευσης και σε ένα μικρότερο που αποτελεί τα δεδομένα ελέγχου.

Γενικά, η απόδοση της διαδικασίας της δυαδικής ταξινόμησης [40] μπορεί να εκτιμηθεί από τις ακόλουθες ποσότητες: True positives (TP), false positives (FP), true negatives (TN), false negatives (FN). Στο πεδίο του σακχαρομύκητα και στο σύνολο δεδομένων που διαθέτουμε γι' αυτήν την εργασία, το θετικό (*positive*) υποδηλώνει ένα δείγμα το οποίο ταξινομείται στην κλάση heat-cold-dtt και το αρνητικό (*negative*) ένα δείγμα το οποίο ταξινομείται στην κλάση elutriation. Συνεπώς, ένα true positive είναι ένα δείγμα της κλάσης heat-cold-dtt το οποίο σωστά ταξινομήθηκε σε αυτή την κλάση, ενώ ένα false positive είναι ένα δείγμα της κλάσης elutriation το οποίο (λανθασμένα) ταξινομήθηκε στην κλάση heat-cold-dtt. Αντίστοιχα, ένα true negative είναι ένα δείγμα της κλάσης elutriation το οποίο σωστά ταξινομήθηκε σε αυτή την κλάση, ενώ ένα false negative είναι ένα δείγμα της κλάσης heat-cold-dtt το οποίο (λανθασμένα) ταξινομήθηκε στην κλάση elutriation.

Πίνακας 2: Πραγματικές και προβλεπόμενες, από έναν ταξινομητή, κλάσεις.

	ΠΡΟΒΛΕΠΟΜΕΝΕΣ ΚΛΑΣΕΙΣ	
ΠΡΑΓΜΑΤΙΚΕΣ ΚΛΑΣΕΙΣ	ELUTRIATION	HEAT-COLD-DTT
ELUTRIATION	TN	FP
HEAT-COLD-DTT	FN	TP

Η ακρίβεια του ταξινομητή υπολογίζεται από την σχέση:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \qquad (3.5.1.1)$$

Που αποτελεί την ποσότητα των true positives και true negatives στον πληθυσμό όλων των δειγμάτων.

Επίσης, υπάρχουν και άλλα μέτρα [49] για την εκτίμηση της απόδοσης ενός ταξινομητή όπως:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.5.1.2}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{3.5.1.3}$$

$$PPV\ (Positive\ Predictive\ Value) = \frac{TP}{(TP + FP)} \tag{3.5.1.4}$$

$$NPV\ (\ Negative\ Predictive\ Value\ ) = \frac{TN}{(TN + FN)} \tag{3.5.1.5}$$

3.5.2 Εκτίμηση πυρήνα

Η εκτίμηση της ποιότητας ενός πυρήνα βασίζεται σε ένα μέτρο ομοιότητας ανάμεσα σε δύο συναρτήσεις πυρήνα ή ανάμεσα σε έναν πυρήνα και σε μία συνάρτηση στόχου. Αυτή η ποσότητα περικλείει το βαθμό συμφωνίας ανάμεσα σε έναν πυρήνα και σε ένα δεδομένο θέμα ταξινόμησης (*learning task*), και έχει πολύ φυσικές ερμηνείες στη μηχανική μάθηση (*machine learning*), οδηγώντας επίσης σε απλούς αλγορίθμους για την επιλογή μοντέλου και μάθησης.

Η ευθυγράμμιση πυρήνα-στόχου (*Kernel Target Alignment*) [34] είναι ένα κριτήριο που αξιολογεί το πόσο κατάλληλος είναι ένας συγκεκριμένος πυρήνας για έναν συγκεκριμένο τομέα εφαρμογής και σύνολο δεδομένων, υπολογίζοντας την εναρμόνιση-ευθυγράμμιση (*alignment*) ανάμεσα στον *kernel matrix* των δειγμάτων εκπαίδευσης του πυρήνα και στον *target matrix* των ετικετών των δειγμάτων εκπαίδευσης (στόχος). Όσο μεγαλύτερη είναι η ομοιότητα ανάμεσα σε αυτούς τους δύο πίνακες, τόσο καλύτερη αναμένεται να είναι η απόδοση ταξινόμησης ενός SVM ταξινομητή.

Πιο συγκεκριμένα, δεδομένου ενός συνόλου δεδομένων εκπαίδευσης X αποτελούμενο από n δείγματα της μορφής $X = \{\ (x_i, y_i)\ |\ x_i\ \in R^m\ , y_i\ \in \{-1, +1\}\ \}_{i=1}^{n}$, όπου το y_i είναι είτε +1 είτε -1 και αποτελεί το στόχο της ταξινόμησης, υποδεικνύοντας την κλάση που ανήκει το x_i, το target matrix σχηματίζεται από τη σχέση:

$$K_y = yy^T \tag{3.5.2.1}$$

Και το συγκρίνουμε για την ομοιότητα του με το kernel matrix K_x των m x m χαρακτηριστικών χρησιμοποιώντας το Frobenius εσωτερικό γινόμενο (*inner product*) που, ανάμεσα σε δύο πίνακες, ορίζεται ως εξής:

$$\langle K , K' \rangle_F = \sum_{i,j=1}^{m} k_{ij} \, k'_{ij} = Tr(KK') \tag{3.5.2.2}$$

Το KTA ορίζεται ως το κανονικοποιημένο Frobenius εσωτερικό γινόμενο:

$$\begin{aligned} A\big(\langle K_x , K_y \rangle \big) &= \frac{\langle K_x , K_y \rangle_F}{\sqrt{\langle K_x , K_x \rangle_F \, \langle K_y , K_y \rangle_F}} \\ &= \frac{\langle K_x , yy^T \rangle_F}{\sqrt{\langle K_x , K_x \rangle_F \, \langle yy^T , yy^T \rangle_F}} \end{aligned} \tag{3.5.2.3}$$

Και παίρνει τιμές στο σύνολο [0,1]. Μία υψηλή τιμή υποδεικνύει ένα καλό kernel matrix, άρα αναμένεται και καλή απόδοση του ταξινομητή, ενώ μία χαμηλή τιμή αποκαλύπτει μία χαμηλή αναμενόμενη ακρίβεια και άρα έναν ακατάλληλο πυρήνα.

Για να συγκριθεί το KTA με άλλα κριτήρια, παρέχεται μία εκτίμηση για το σφάλμα του ταξινομητή:

$$KTA_{err} = 1 - A(\langle K_x , K_y \rangle) \tag{3.5.2.4}$$

ΚΕΦΑΛΑΙΟ 4: Προτεινόμενη Μεθοδολογία

Ένα μεγάλο πρόβλημα [13] στην εφαρμογή των προφίλ της γονιδιακής έκφρασης γενικά και του σακχαρομύκητα ειδικότερα, για την ταξινόμηση και την πρόβλεψη, αποτελεί το γεγονός ότι ο αριθμός των χαρακτηριστικών (γονιδίων) υπερβαίνει κατά πολύ τον αριθμό των δειγμάτων. Μία μικρή, όμως, συλλογή γονιδίων, επιλεγμένων σωστά, μπορεί να οδηγήσει σε καλά αποτελέσματα ταξινόμησης. Οπότε, η επιλογή γονιδίων είναι ένα θέμα ζωτικής σημασίας στη μοριακή ταξινόμηση. Ο στόχος της επιλογής γονιδίων είναι να επιλεγεί ένα υποσύνολο από πληροφοριακά γονίδια που μειώνουν το χρόνο εξεργασίας και παρέχουν υψηλότερη ακρίβεια ταξινόμησης από εκείνη που σημειώνεται στο αρχικό σύνολο όλων των γονιδίων.

Σε αυτό το κεφάλαιο παρουσιάζεται αναλυτικά η προτεινόμενη μεθοδολογία για την επιλογή γονιδίων στην οποία υπόκεινται τα πειράματα αυτής της εργασίας, καθώς και ο τρόπος με τον οποίο αυτή εφαρμόζεται στα διαφορετικά μοντέλα ταξινόμησης τα οποία υλοποιούνται. Στόχος αυτής της μεθοδολογίας είναι η εξαγωγή μίας γονιδιακής υπογραφής για κάθε έναν από τους αλγορίθμους ταξινόμησης της εργασίας, η σύγκριση των υπογραφών αυτών μεταξύ τους και η εξαγωγή ενός συμπεράσματος σχετικά με την επιρροή που έχουν οι διαφορετικοί αλγοριθμικοί παράμετροι τόσο στη δύναμη της πρόβλεψης μίας υπογραφής όσο και στη διαδικασία επιλογής της.

4.1 Επιλογές στο Σύνολο Δεδομένων

Όπως αναφέρεται στην ενότητα 1.6, το σύνολο δεδομένων που διαθέτουμε αποτελείται από 2467 γονίδια και 79 δείγματα τα οποία μοιράζονται σε 8 διεργασίες. Αυτές οι διεργασίες αποτελούν τις κλάσεις ενδιαφέροντος για την ταξινόμηση. Επειδή κάποια από τα δείγματα έχουν ελλιπή δεδομένα, δηλαδή κάποια χαρακτηριστικά των γονιδίων του σακχαρομύκητα δε διαθέτουν τιμή, εφαρμόζεται απόδοση τιμών σε αυτά (*imputation*) σύμφωνα με τη λογική του κοντινότερου γείτονα. Με αυτό τον τρόπο, κάθε κενό χαρακτηριστικό αντικαθίσταται από την αντίστοιχη τιμή που υπάρχει στο δείγμα του κοντινότερου γείτονα. Το δείγμα του κοντινότερου γείτονα είναι το κοντινότερο δείγμα σε ευκλείδεια απόσταση, και αν η αντίστοιχη τιμή σε αυτό είναι επίσης κενή τότε χρησιμοποιείται το αμέσως επόμενο κοντινότερο δείγμα.

Στη συνέχεια, για την υλοποίηση της δυαδικής ταξινόμησης που εκτελείται σε αυτή την εργασία, απομονώνονται μόνο δύο από τις οκτώ διεργασίες του συνόλου δεδομένων. Από το νέο σύνολο δεδομένων, με τα 2467 γονίδια και τις δύο κλάσεις ενδιαφέροντος, γίνεται εξαγωγή ενός συνόλου δεδομένων εκπαίδευσης και ενός συνόλου δεδομένων ελέγχου (το 1/10 του συνόλου δεδομένων), με την εφαρμογή του ten fold cross validation, που αναπτύσσεται στην ενότητα 2.4, σε αυτό. Τα δείγματα όμως που ανήκουν στο σύνολο δεδομένων ελέγχου, που προκύπτει από τη διαδικασία του ten fold cross validation, είναι πολύ λίγα (περίπου 3), οπότε ο αριθμός τους κρίνεται ανεπαρκής για τη εκτίμηση της απόδοσης των ταξινομητών.

Για το λόγο αυτό, οι οκτώ διεργασίες συγχωνεύονται σε ζεύγη, με σεβασμό πάντα στη βιολογική τους συμπεριφορά. Έτσι, προκύπτουν ζευγάρια, τα οποία αποτελούνται από δύο σύνολα ευρύτερων διεργασιών, κάθε ένα από τα οποία μπορεί να περικλείει περισσότερες της μίας διεργασίες, από τις 8 αρχικές. Για παράδειγμα οι 3 διεργασίες, που προκαλούν στον σακχαρομύκητα σοκ λόγω υψηλής θερμοκρασίας, αναγωγικών μέσων και χαμηλής θερμοκρασίας, συγχωνεύονται σε μία ευρύτερη διεργασία. Παράλληλα επιτυγχάνεται μείωση του αριθμού των folds σε πέντε, εφαρμόζοντας five fold cross validation, οπότε αυξάνεται ο αριθμός των δειγμάτων σε κάθε fold και άρα και στο σύνολο δεδομένων ελέγχου.

Πολλά σύνολα δεδομένων που ανακύπτουν στη βιοπληροφορική και σε άλλες περιοχές εφαρμογών είναι ασύμμετρα [32], δηλαδή η μία κλάση περιλαμβάνει πολλά περισσότερα δείγματα από την άλλη. Τα ασύμμετρα σύνολα δεδομένων μπορεί να παρουσιάσουν μία πρόκληση κατά τη διάρκεια της εκπαίδευσης ενός ταξινομητή. Ένα πρώτο πρόβλημα είναι ότι σχεδιάζονται ταξινομητές κλάσης πλειοψηφίας, δηλαδή ταξινομητές που εκχωρούν κάθε νέο δείγμα στην κλάση με τα περισσότερα δείγματα και γι’ αυτό το λόγο δεν είναι χρήσιμοι όταν αναμένουμε υψηλή ακρίβεια ταξινόμησης από αυτούς. Ένα δεύτερο πρόβλημα που προκύπτει είναι το ότι η έννοια της ακρίβειας, έτσι όπως περιγράφεται από τη σχέση (3.5.1.1) εκφράζοντας το ποσοστό των επιτυχώς ταξινομημένων δειγμάτων, δεν είναι ένας καλός τρόπος για να μετρήσουμε την επιτυχία ενός ταξινομητή που εφαρμόζεται σε ασύμμετρα δεδομένα. Αυτό γίνεται φανερό από το γεγονός ότι ο ταξινομητής κλάσης πλειοψηφίας αποδίδει καλά για την κλάση πλειοψηφίας και μόνο.

Επειδή το δικό μας σύνολο δεδομένων δεν αποτελεί εξαίρεση ως προς την ασυμμετρία των κλάσεων, και για να αποφύγουμε τα παραπάνω προβλήματα, επιλέξαμε ένα ζεύγος ευρύτερων διεργασιών με ίσο αριθμό δειγμάτων και στις δύο διεργασίες. Το ζευγάρι των διεργασιών, το οποίο χρησιμοποιήθηκε σε όλη την εργασία και που πάνω σε αυτό έχουν εξαχθεί όλα τα αποτελέσματά της, συνιστώντας τις δύο κλάσεις ενδιαφέροντος της δυαδικής ταξινόμησης, είναι το elutriation (ο κύκλος της κυτταρικής διαίρεσης συγχρονίζεται με το φυγόκεντρο διαχωρισμό επίπλευσης) και το heat-cold-dtt (θερμικά και αναγωγικά σοκ), με 28 δείγματα συνολικά και 2467 γονίδια. Η πρώτη κλάση, που είναι το elutriation, διαθέτει 14 δείγματα και το target της χαρακτηρίζεται από το -1, ενώ η δεύτερη κλάση, που είναι το heat-cold-dtt και διαθέτει επίσης 14 δείγματα, χαρακτηρίζεται από target με +1.

4.2 Δομή Χρήσης των Μοντέλων Μάθησης και Πρόβλεψης

Τα μοντέλα μάθησης και πρόβλεψης, που υλοποιούνται, είναι ταξινομητές που βασίζονται σε πυρήνα (*kernel based classifiers*), καθώς και συνδυασμοί αυτών. Αρχικά, υλοποιούνται οι μέθοδοι support vector machine, kernel nearest centroid, kernel minimum squared error machine και kernel subspace που αποτελούν ταξινομητές και που περιγράφονται στις ενότητες 3.1 και 3.2. Σε κάθε έναν από αυτούς μπορεί να εφαρμοστεί οποιοσδήποτε από τους πυρήνες Gaussian, γραμμικός, πολυωνυμικός δευτέρου και τρίτου βαθμού και τετραγωνικός που αναφέρονται στην ενότητα 3.2. Επιλέγουμε έναν πυρήνα για κάθε ταξινομητή με τη λογική που περιγράφεται αναλυτικά στην επόμενη ενότητα 4.3.

Παράλληλα με τη χρήση αυτών των ταξινομητών εξετάζεται, επίσης, η προσέγγιση του μείγματος ειδικών, η οποία αναλύεται στην ενότητα 3.3, δηλαδή ο συνδυασμός των ατομικών ταξινομητών με στόχο την καλύτερη πρόβλεψη. Το πρώτο μείγμα ειδικών που βασίζεται στη συγχώνευση των επιμέρους ειδικών με ένα συνδυασμό ασαφών βαρών και που περιγράφεται στην υποενότητα 3.3.1, συνδυάζει τους τέσσερις ταξινομητές που αναφέρονται παραπάνω. Επειδή το support vector machine χρησιμοποιείται σε αυτή την εργασία με δύο διαφορετικούς πυρήνες, όπως θα αναφερθεί αναλυτικά στην υποενότητα 4.3.2, στην ουσία οι ταξινομητές που αποτελούν αυτό το πρώτο μείγμα ειδικών είναι πέντε.

Για το δεύτερο και για το τρίτο μείγμα ειδικών των υποενοτήτων 3.3.2 και 3.3.3, που βασίζονται στη συγχώνευση με gradient descent learning αλγόριθμο, υλοποιούνται δύο διαφορετικές συναρτήσεις logistic-sigmoid, που αποτελούν τους επιμέρους ειδικούς.

Το τέταρτο μείγμα ειδικών, που βασίζεται στην εκπαίδευση των επιμέρους ειδικών σε διαφορετικά υποσύνολα του χώρου των χαρακτηριστικών, ο οποίος διαχωρίζεται με som και περιγράφεται στην υποενότητα 3.3.4, υλοποιείται με δύο επιμέρους ταξινομητές και με δύο gate functions, που βασίζονται στη λογική των αποστάσεων για να εκτελέσουν ταξινόμηση. Τα δείγματα του συνόλου εκπαίδευσης μοιράζονται σε δύο μόνο ομάδες επειδή ο αριθμός τους είναι αρκετά μικρός και έτσι καθίσταται σχεδόν απαγορευτικό να μοιραστούν σε περισσότερες.

Στο υπόλοιπο της εργασίας, για λόγους ευκολίας, με τον όρο μοντέλο θα εννοείται οποιοσδήποτε από τους ατομικούς ταξινομητές ή οποιοδήποτε από τα τέσσερα μείγματα ειδικών. Όπου είναι απαραίτητο, θα γίνεται κατάλληλη αναφορά σε συγκεκριμένο μοντέλο.

Η στρατηγική που ακολουθούμε για την εφαρμογή της προτεινόμενης μεθοδολογίας επιλογής χαρακτηριστικών για την επιλογή γονιδίων, αναπτύσσεται στις επόμενες ενότητες, ενώ στο διάγραμμα ροής της εικόνας 30, φαίνεται η γενική εικόνα των βημάτων της.

Ο αλγόριθμος ο οποίος χρησιμοποιείται για την αξιολόγηση χαρακτηριστικών είναι ο RFE-LNW, που περιγράφεται στην ενότητα 3.4. Η μέθοδος αυτή δίνει από ένα βάρος στα γονίδια που εξετάζονται κάθε φορά. Η επιλογή του αριθμού γονιδίων που διατηρείται σε κάθε βήμα της μεθοδολογίας μας αξιολογείται από ένα ή περισσότερα μοντέλα ταξινόμησης.

Στην εργασία εξετάζουμε τις επιπτώσεις πολλών διαφορετικών ταξινομητών. Επίσης, εξετάζεται η δυνατότητα καλύτερης επίδοσης μέσω χρήσης μείγματος ταξινομητών. Προς αυτή την κατεύθυνση, εξετάζονται διάφορα σχήματα μίξης ειδικών. Για την καλύτερη γενίκευση των αποτελεσμάτων, η διαδικασία επιλογής/αξιολόγησης γίνεται μέσα σε ένα κύκλο cross validation.

Η προτεινόμενη μεθοδολογία επιλογής γονιδίων ακολουθεί δύο επίπεδα. Στο πρώτο επίπεδο εξετάζεται ολόκληρο το σύνολο των δεδομένων για μία πρώτη επιλογή πολλών χαρακτηριστικών γονιδίων. Στο δεύτερο επίπεδο γίνεται λεπτομερέστερη επιλογή γονιδίων βασισμένη στην αξιολόγηση των δεδομένων που απαρτίζονται όχι από το συνολικό αριθμό γονιδίων αλλά από το μέρος των γονιδίων που προκύπτουν από την πρώτη επιλογή. Κάθε επίπεδο αξιολογεί τον κατάλληλο αριθμό γονιδίων που θα επιλεγούν μέσα σε ένα κύκλο 5 fold cross validation και έπειτα προχωρά στην επιλογή συγκεκριμένων γονιδίων υλοποιώντας 20 κύκλους 5 fold cross validation.

<u>Στο πρώτο επίπεδο αξιολόγησης δεδομένων</u>, *πραγματοποιείται αξιολόγηση της απόδοσης ταξινόμησης των 2467 γονιδίων ολόκληρου του συνόλου δεδομένων.* Μέσα σε ένα σύστημα five fold cross validation, εφαρμόζεται ο αλγόριθμος RFE-LNW για την

αναδρομική εξάλειψη των χαρακτηριστικών, πάνω στο σύνολο δεδομένων με τα 2467 γονίδια που περιγράφεται στην ενότητα 4.1, και σε κάθε βήμα του εκτιμάται η προγνωστική δύναμη του συνόλου των υπολοίπων γονιδίων από τους παραπάνω ταξινομητές καθώς και από τα προαναφερθέντα μείγματα αυτών. Με αυτό τον τρόπο, όταν όλα τα γονίδια σταδιακά εξαλειφθούν, είμαστε σε θέση να εξάγουμε μία γενική εικόνα για τη μέση απόδοση που σημείωσαν τα μοντέλα μας για όλα τα υποσύνολα γονιδίων από 2467 έως και 1, και να βρούμε ποιο ήταν το μικρότερο υποσύνολο στο οποίο σημειώθηκε η μέση μέγιστη ακρίβεια από το κάθε μοντέλο.

Οπότε, σε αυτό το πρώτο επίπεδο αξιολόγησης των δεδομένων πραγματοποιείται μία διαλογή γονιδίων για το κάθε μοντέλο, που αναδεικνύει τον αριθμό των ελαχίστων γονιδίων (έστω A) για τα οποία σημειώθηκε η μέση μέγιστη απόδοση ταξινόμησης. Σε αυτό τον αριθμό, που αποφέρει μία μεγάλη απόδοση ταξινόμησης και που φυσικά είναι διαφορετικός για κάθε μοντέλο, βασίζεται το πρώτο επίπεδο επιλογής γονιδίων για να ξεκινήσει τη διαδικασία για τη μείωση των γονιδίων. Πιο συγκεκριμένα, ο αριθμός αυτός (A) χρησιμοποιείται ως κριτήριο τερματισμού του κάθε κύκλου επανάληψης (fold) στο πρώτο επίπεδο επιλογής γονιδίων.

Ακολουθεί <u>το πρώτο επίπεδο επιλογής γονιδίων</u>, στο οποίο *πραγματοποιείται επιλογή για τα 2467 γονίδια σύμφωνα με τη συχνότητα εμφάνισής τους.* Εκτελούνται 20 επαναλήψεις της διαδικασίας του πρώτου επιπέδου αξιολόγησης δεδομένων, δηλαδή εφαρμόζεται 100 φορές (20επαναλήψεις x 5folds = 100 runs) ο αλγόριθμος FRE-LNW πάνω στο σύνολο δεδομένων με τα 2467 γονίδια, προσαρμοσμένος όμως για κάθε μοντέλο, με τρόπο τέτοιο ώστε η αναδρομική εξάλειψη να μην πραγματοποιείται για όλα τα γονίδια από 2467 μέχρι 1, αλλά μέχρι τον αριθμό A των ελαχίστων γονιδίων που θεωρήθηκε βέλτιστος σύμφωνα με την αξιολόγηση της απόδοσης που έγινε γι' αυτά κατά το πρώτο επίπεδο αξιολόγησης δεδομένων.

Στο τέλος αυτού του πρώτου επιπέδου επιλογής γονιδίων επιτυγχάνεται επιλογή των C συγκεκριμένων γονιδίων βασισμένη στις σχετικές συχνότητες εμφάνισής τους, σύμφωνα με ένα όριο τιμών (γόνατο γραφικής παράστασης συχνοτήτων-γονιδίων). Αυτά τα συγκεκριμένα C γονίδια θα αξιολογηθούν στο δεύτερο επίπεδο αξιολόγησης δεδομένων, ως προς την απόδοση ταξινόμησης που αποφέρουν. Επιπλέον, εξάγεται ένας αριθμός D που δηλώνει τον πιθανό αριθμό γονιδίων με τις υψηλότερες *συχνότητες εμφάνισης για να χρησιμοποιηθεί σαν κριτήριο τερματισμού στο δεύτερο επίπεδο επιλογής γονιδίων.*

Στο δεύτερο επίπεδο αξιολόγησης δεδομένων *πραγματοποιείται αξιολόγηση της απόδοσης ταξινόμησης των C γονιδίων που έχουν επιλεγεί στο πρώτο επίπεδο επιλογής.* Το σύνολο δεδομένων μειώνεται ώστε να περιλαμβάνει μόνο αυτά τα C γονίδια και στη συνέχεια εκτελείται πάνω σε αυτό ένα σύστημα five fold cross validation, μέσα στο οποίο εφαρμόζεται ο αλγόριθμος RFE-LNW, που περιγράφεται στην ενότητα 3.4 για την αναδρομική εξάλειψη των χαρακτηριστικών. Σε κάθε βήμα του η προγνωστική δύναμη του συνόλου των υπολοίπων γονιδίων εκτιμάται από τους παραπάνω ταξινομητές καθώς και από τα προαναφερθέντα μείγματα αυτών και συγκρίνεται η απόδοση με αυτή των αρχικών δεδομένων.

Στο δεύτερο επίπεδο επιλογής γονιδίων *πραγματοποιείται επιλογή για τα C γονίδια που έχουν προκύψει από το πρώτο επίπεδο επιλογής, σύμφωνα με τη συχνότητα εμφάνισής τους.* Εκτελούνται 20 επαναλήψεις της διαδικασίας του αρχικού επιπέδου διαλογής γονιδίων, δηλαδή εφαρμόζεται 100 φορές (20επαναλήψεις x 5folds = 100 runs) ο αλγόριθμος FRE-LNW πάνω στο σύνολο δεδομένων με τα C γονίδια, προσαρμοσμένος όμως για κάθε μοντέλο, με τρόπο τέτοιο ώστε η αναδρομική εξάλειψη να μην πραγματοποιείται για όλα τα γονίδια από C μέχρι 1, αλλά μέχρι τον αριθμό γονιδίων D, που αποτελείται από τα γονίδια με τις υψηλότερες συχνότητες εμφάνισης από όλα τα C γονίδια του πρώτου επιπέδου επιλογής. *Στο τέλος αυτού του επιπέδου επιτυγχάνεται επιλογή των τελικών F γονιδίων βασισμένη στις σχετικές συχνότητες εμφάνισής τους, σύμφωνα με ένα όριο τιμών (γόνατο γραφικής παράστασης συχνοτήτων-γονιδίων).*

Στο τελικό επίπεδο αξιολόγησης των γονιδιακών υπογραφών *πραγματοποιείται αξιολόγηση της απόδοσης ταξινόμησης των F γονιδίων που έχουν επιλεγεί στο δεύτερο επίπεδο επιλογής γονιδίων.* Το σύνολο δεδομένων μειώνεται ώστε να περιλαμβάνει μόνο αυτά τα F γονίδια και στη συνέχεια εκτελείται 10 φορές πάνω σε αυτό, ένα σύστημα five fold cross validation (5folds x 10 επαναλήψεις =50 runs), μέσα στο οποίο τα μοντέλα ταξινόμησης εκπαιδεύονται και ελέγχεται η ικανότητα γενίκευσής τους για το σύνολο των συγκεκριμένων F γονιδίων. Σε αυτή την τελική φάση, δε χρησιμοποιείται ο αλγόριθμος RFE-LNW αφού δε θέλουμε να μετρήσουμε τη μέση ακρίβεια της ταξινόμησης για όλα τα υποσύνολα γονιδίων από F έως και 1, αλλά μόνο για το σύνολο των F συγκεκριμένων γονιδίων. Τέλος, γίνεται σύγκριση της μέσης απόδοσης στο τελικό επίπεδο αξιολόγησης των γονιδιακών υπογραφών σε σχέση με τη μέση απόδοσή τους στο πρώτο επίπεδο αξιολόγησης των γονιδίων.

Στις επόμενες ενότητες αυτού του κεφαλαίου αναπτύσσονται περισσότερο αυτά τα πέντε στάδια της προτεινόμενης μεθοδολογίας της εργασίας.

ΠΡΩΤΟ ΕΠΙΠΕΔΟ ΑΞΙΟΛΟΓΗΣΗΣ ΔΕΔΟΜΕΝΩΝ:

- ***Αρχική Dataset_1 : 28 δείγματα και 2467 γονίδια***
- Αξιολόγηση της απόδοσης ταξινόμησης των μοντέλων στο αρχικό σύνολο δεδομένων.
- Για κάθε μοντέλο σημειώνεται ο ελάχιστος αριθμός **A** γονιδίων με την μέγιστη ακρίβεια ταξινόμηση. Ο αριθμός αυτός θα χρησιμοποιηθεί ως κριτήριο τερματισμού στο πρώτο επίπεδο επιλογής των γονιδίων.

⇩

ΠΡΩΤΟ ΕΠΙΠΕΔΟ ΕΠΙΛΟΓΗΣ ΓΟΝΙΔΙΩΝ:

- Σύμφωνα με τον αριθμό **A** του κριτηρίου τερματισμού του πρώτου επιπέδου αξιολόγησης, εκτελείται διαλογή των 2467 γονιδίων και επιλέγονται μόνο τα **C** γονίδια για να αξιολογηθούν στο δεύτερο επίπεδο αξιολόγησης δεδομένων.
- Σημειώνεται ο αριθμός **D** γονιδίων με τις υψηλότερες συχνότητες για να χρησιμοποιηθεί ως κριτήριο τερματισμού στο δεύτερου επίπεδο επιλογής γονιδίων.

⇩

ΔΕΥΤΕΡΟ ΕΠΙΠΕΔΟ ΑΞΙΟΛΟΓΗΣΗΣ ΔΕΔΟΜΕΝΩΝ :

- ***Νέα Dataset_2: 28 δείγματα και C γονίδια. (C<<2467).***
- Αξιολόγηση της απόδοσης ταξινόμησης των μοντέλων στο νέο σύνολο δεδομένων και σύγκριση με την αντίστοιχη του πρώτου επιπέδου αξιολόγησης δεδομένων.

ΔΕΥΤΕΡΟ ΕΠΙΠΕΔΟ ΕΠΙΛΟΓΗΣ ΓΟΝΙΔΙΩΝ:

- Σύμφωνα με τον αριθμό **D** του κριτηρίου τερματισμού του πρώτου επιπέδου επιλογής γονιδίων, εκτελείται διαλογή των C γονιδίων και επιλέγονται μόνο τα **F** γονίδια.

ΤΕΛΙΚΟ ΕΠΙΠΕΔΟ ΑΞΙΟΛΟΓΗΣΗΣ ΓΟΝΙΔΙΑΚΩΝ ΥΠΟΓΡΑΦΩΝ:

- ***Νέα Dataset 3: 28 δείγματα και F γονίδια. (F<<C , F<<<2467).***
- Αξιολόγηση της απόδοσης ταξινόμησης των μοντέλων στο νέο σύνολο δεδομένων και σύγκριση με την αντίστοιχη του πρώτου επιπέδου αξιολόγησης δεδομένων.

Εικόνα 30: Τα 5 βήματα της προτεινόμενης μεθοδολογίας της επιλογής χαρακτηριστικών για την επιλογή γονιδίων. Διακρίνονται το πρώτο επίπεδο αξιολόγησης δεδομένων, το πρώτο επίπεδο επιλογής γονιδίων, το δεύτερο επίπεδο αξιολόγησης δεδομένων, το δεύτερο επίπεδο επιλογής γονιδίων και το τελικό επίπεδο της αξιολόγησης των τελικών γονιδιακών υπογραφών.

4.3 Πρώτο Επίπεδο Αξιολόγησης Δεδομένων

Αρχικά, όπως έχει ήδη αναφερθεί, το σύνολο δεδομένων διαθέτει 2467 γονίδια και 28 δείγματα τα οποία είναι μοιρασμένα στις δύο κλάσεις, elutriation και heat-cold-dtt, με ίσο αριθμό δειγμάτων η καθεμία. Σε αυτό το σύνολο δεδομένων επιχειρούμε να εκπαιδεύσουμε κάθε μοντέλο μάθησης που υλοποιείται στα πλαίσια αυτής της εργασίας, και έπειτα να το ελέγξουμε ως προς τη σημαντικότερη ιδιότητα του, που σχετίζεται με την ταξινόμηση, την ικανότητα γενίκευσης σε νέα δείγματα.

<u>Βήμα 1: Διαχωρισμός δεδομένων.</u> Για το λόγο αυτό, εφαρμόζεται η τεχνική του five fold cross validation, η οποία αναπτύσσεται στην ενότητα 2.4, και τα δείγματα του συνόλου δεδομένων χωρίζονται σε 5 υποσύνολα (folds) αποτελούμενα από 5 ή 6 δείγματα το καθένα (28/5=5,6). Από τα 5 αυτά υποσύνολα, τα 4 χρησιμοποιούνται ως δεδομένα εκπαίδευσης με συνολικό αριθμό δειγμάτων 22 ή 23, ενώ το ένα διατηρείται ως σύνολο δεδομένων ελέγχου διαθέτοντας 5 ή 6 δείγματα. Τόσο το σύνολο δεδομένων εκπαίδευσης όσο και το σύνολο δεδομένων ελέγχου αποτελούνται από 2467 γονίδια. Εκτελούνται 5 κύκλοι επανάληψης του cross validation ώστε το μοντέλο μάθησης να εκπαιδευτεί 5 φορές και σε κάθε έναν κύκλο να χρησιμοποιηθεί ακριβώς μία φορά ένα από τα 5 υποσύνολα ως σύνολο δεδομένων ελέγχου.

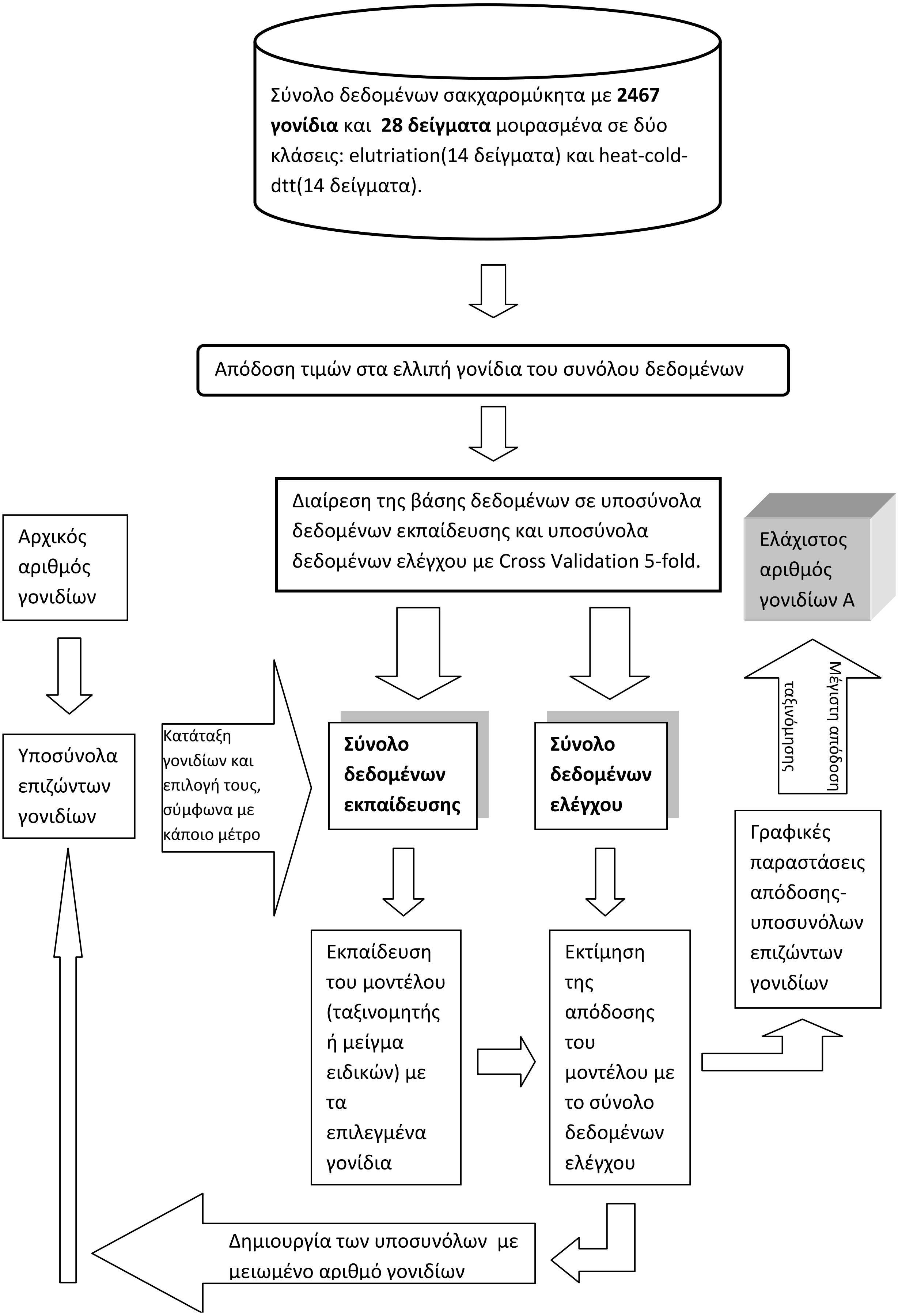
<u>Βήμα 2: Μείωση γονιδίων</u>. Μετά από τη διαδικασία δημιουργίας των συνόλων εκπαίδευσης και ελέγχου, σειρά έχει η διαδικασία της επιλογής των γονιδίων. Σε κάθε κύκλο επανάληψης του cross validation, εκτελείται ο αλγόριθμος recursive feature elimination linear neuron weight, όπως αυτός περιγράφεται στην ενότητα 3.4. Πιο συγκεκριμένα, σε κάθε κύκλο επανάληψης του cross validation, ο αλγόριθμος FRE-LNW αποδίδει αρχικά βάρη στα 2467 γονίδια. Έπειτα, εκτελούνται επαναλήψεις αυτού του αλγορίθμου και σε κάθε μία από αυτές, τα βάρη ανανεώνονται και τα γονίδια κατατάσσονται σύμφωνα με τις απόλυτες τιμές αυτών των βαρών, με σκοπό να αφαιρεθούν, τόσο από το σύνολο δεδομένων εκπαίδευσης όσο και από το σύνολο δεδομένων ελέγχου, το ή τα γονίδια με το μικρότερο βάρος καθώς θεωρούνται τα λιγότερο σημαντικά. Στη συνέχεια, το μοντέλο μάθησης εκπαιδεύεται και εκτιμάται η ακρίβεια ταξινόμησης του, για το υποσύνολο των υπολοίπων γονιδίων, από τη σχέση (3.5.1.1). Οι επαναλήψεις του αλγορίθμου σταματούν όταν αφαιρεθούν σταδιακά όλα

τα γονίδια. Όλη η παραπάνω διαδικασία επαναλαμβάνεται 5 φορές, με σκοπό να εκπαιδευτεί το μοντέλο σε όλα τα πιθανά σύνολα εκπαίδευσης και να δοκιμαστεί και για τα πέντε σύνολα ελέγχου.

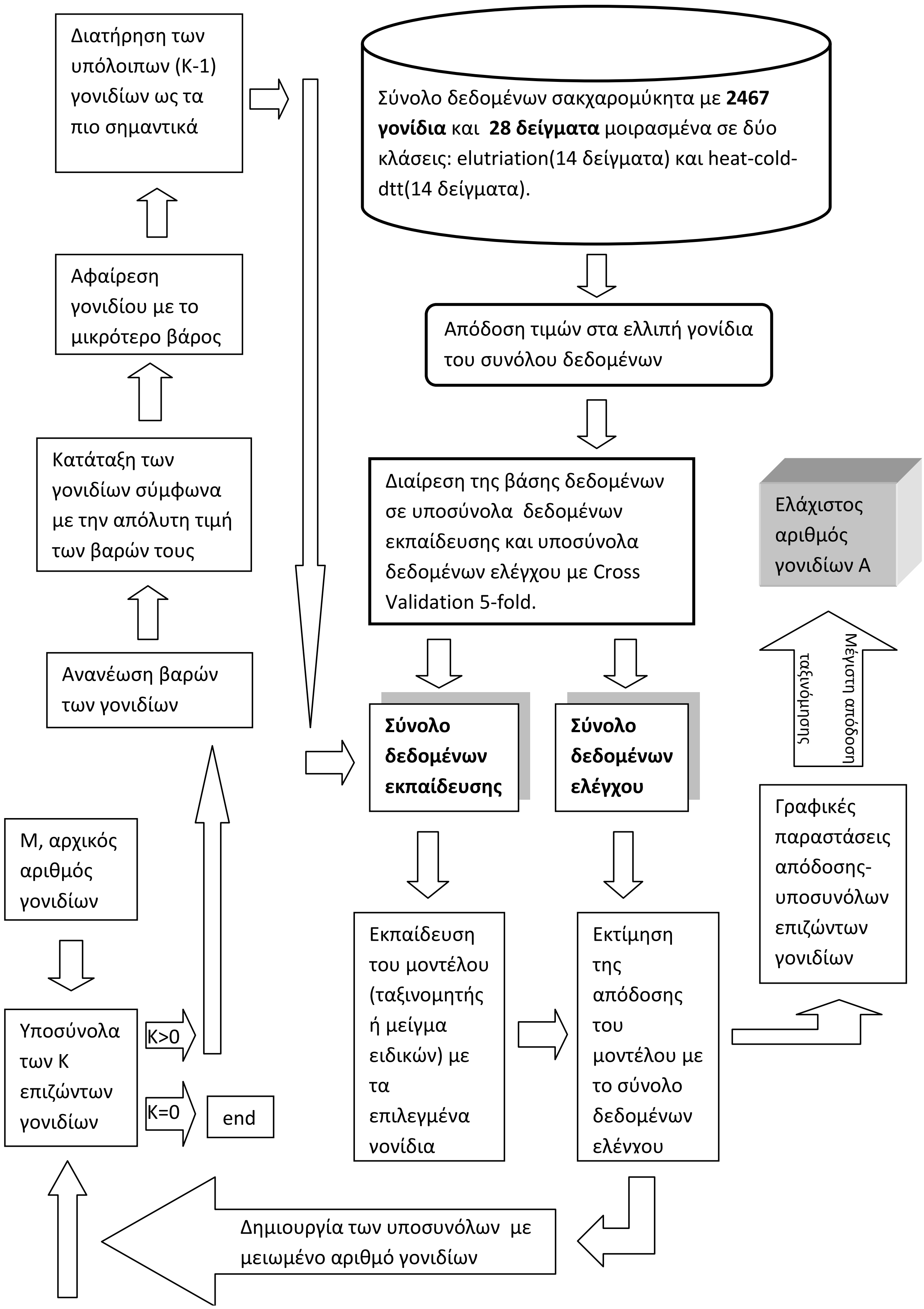
Βήμα 3: Εκτίμηση αποτελεσμάτων. Για να εξάγουμε αποτελέσματα από την παραπάνω διαδικασία, προσανατολιζόμαστε σε δύο κατευθύνσεις. Πρώτον, στοχεύουμε στην απόκτηση μίας γενικής εικόνας για την ακρίβεια που σημείωσε το μοντέλο, για όλα τα υποσύνολα γονιδίων και για όλα τα folds στα οποία εξετάστηκε. Αφού σε κάθε επανάληψη του FRE-LNW μετριέται το ποσοστό επιτυχίας της ταξινόμησης του μοντέλου για το τρέχον υποσύνολο γονιδίων, μετά το πέρας όλων των επαναλήψεων αυτού του αλγορίθμου, υπολογίζεται ο μέσος όρος των επιμέρους ποσοστών επιτυχίας για όλα τα υποσύνολα γονιδίων, από 2467 έως και 0 γονίδια, που έχουν προκύψει. Επειδή ο αλγόριθμος FRE-LNW εκτελείται μέσα σε five fold cross validation, προκύπτουν πέντε τέτοιοι μέσοι όροι. Τέλος, υπολογίζεται ο τελικός μέσος όρος και των πέντε και με αυτό τον τρόπο προκύπτει μία μέση τιμή για το μοντέλο που αντιπροσωπεύει το γενικό ποσοστό επιτυχίας που σημείωσε για όλα τα υποσύνολα γονιδίων που κλήθηκε να ταξινομήσει σε όλα τα folds.

Δεύτερον, ενδιαφερόμαστε να αναζητήσουμε το μικρότερο δυνατό υποσύνολο γονιδίων στο οποίο σημειώνεται η μεγαλύτερη ακρίβεια ταξινόμησης από το μοντέλο. Σκοπός της αναδρομικής εξάλειψης γονιδίων, που εφαρμόζεται στο μοντέλο, για κάθε fold, είναι στο τέλος της, να βρεθεί ο μικρότερος αριθμός γονιδίων με τον οποίο το μοντέλο μπόρεσε να ταξινομήσει καλύτερα τα νέα δείγματα του συνόλου ελέγχου, σε σύγκριση με τις αποδόσεις ταξινόμησης που σημείωσε σε όλα τα άλλα υποσύνολα γονιδίων, μικρότερα ή μεγαλύτερα. Επειδή προκύπτουν πέντε τέτοιοι αριθμοί λιγότερων δυνατών γονιδίων με την καλύτερη απόδοση, ένας σε κάθε fold, υπολογίζεται ο μέσος όρος τους (έστω A). Ο αριθμός, στον οποίο καταλήγουμε, θεωρείται ο αριθμός των πιο σημαντικών γονιδίων από τα αρχικά 2467 για το μοντέλο.

Όλη η διαδικασία που μόλις περιγράφηκε απεικονίζεται στα δύο διαγράμματα ροής των εικόνων 31 και 32 που ακολουθούν. Στην εικόνα 31 φαίνεται παραστατικά ο τρόπος με τον οποίο συνδέονται τα βασικά τμήματα της εργασίας, δηλαδή το cross



Εικόνα 31: Πρώτο επίπεδο αξιολόγησης δεδομένων για επιλογή γονιδίων. Βασική ροή της εργασίας, στην οποία αναπαρίσταται ο τρόπος με τον οποίο συνδέονται τα βασικά τμήματα της, όπως το cross validation, η εκπαίδευση και ο έλεγχος της απόδοσης των μοντέλων, η μείωση γονιδίων.

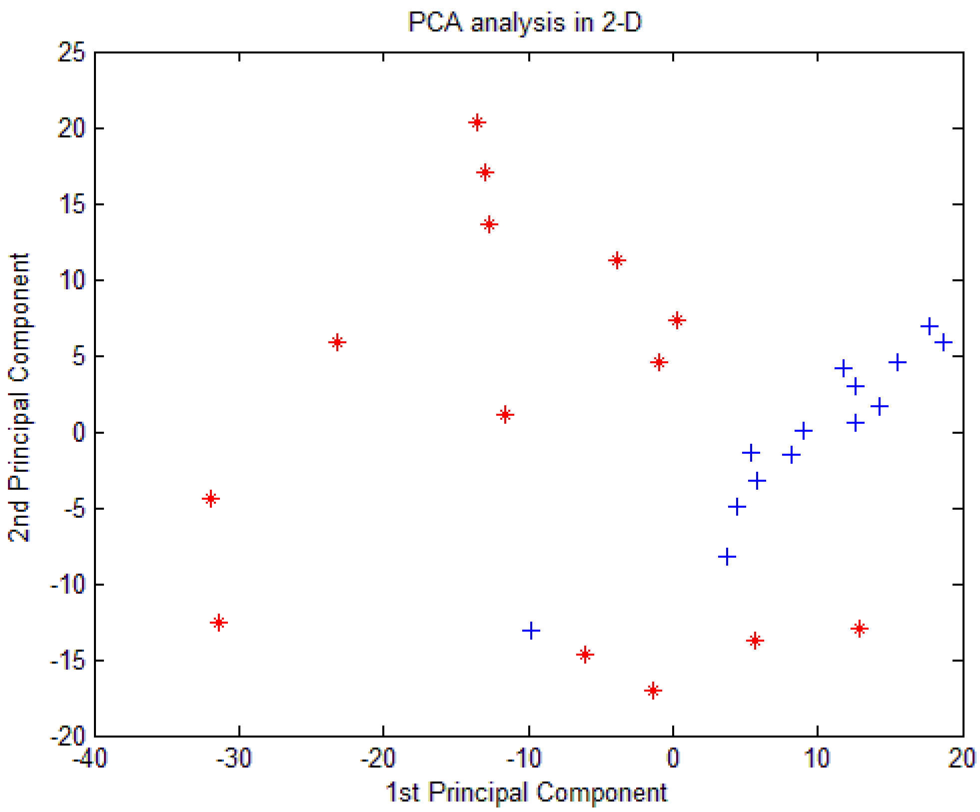


Εικόνα 32: Πρώτο επίπεδο αξιολόγησης δεδομένων για επιλογή γονιδίων. Επανάληψη της βασικής ροής της εργασίας, εμπλουτίζοντάς την με την αναλυτική περιγραφή του αλγορίθμου RFE-LNW για τη μείωση των γονιδίων.

validation, η εκπαίδευση του μοντέλου ταξινόμησης, ο έλεγχος της απόδοσής του, η μείωση γονιδίων. Η εικόνα 32 επαναλαμβάνει το διάγραμμα ροής της εικόνας 31, προσθέτοντας όμως αναλυτική περιγραφή του αλγορίθμου RFE-LNW, για τη μείωση των γονιδίων.

4.3.1 Οπτική επιβεβαίωση των αποτελεσμάτων

Η μέθοδος ανάλυσης κυρίων συνιστωσών (*principal component analysis-pca*) [50], αναλύει ένα χώρο δεδομένων σε ένα γραμμικό συνδυασμό μιας μικρής συλλογής από βάσεις, οι οποίες είναι ανά ζεύγη ορθογώνιες και οι οποίες συλλαμβάνουν τις κατευθύνσεις της μέγιστης διακύμανσης στο σύνολο δεδομένων. Όπως δείχνει η εικόνα 33, εφαρμόζοντας την PCA ανάλυση στο αρχικό σύνολο δεδομένων με τα 2467 γονίδια, παρατηρούμε ότι τα δείγματα της κλάσης elutriation βρίσκονται συγκεντρωμένα σε μία περιοχή, ενώ εκείνα της κλάσης heat-cold-dtt βρίσκονται διάσπαρτα στο χώρο, καθιστώντας, με αυτό τον τρόπο, δύσκολη τη διαχωρισιμότητα των δύο κλάσεων.

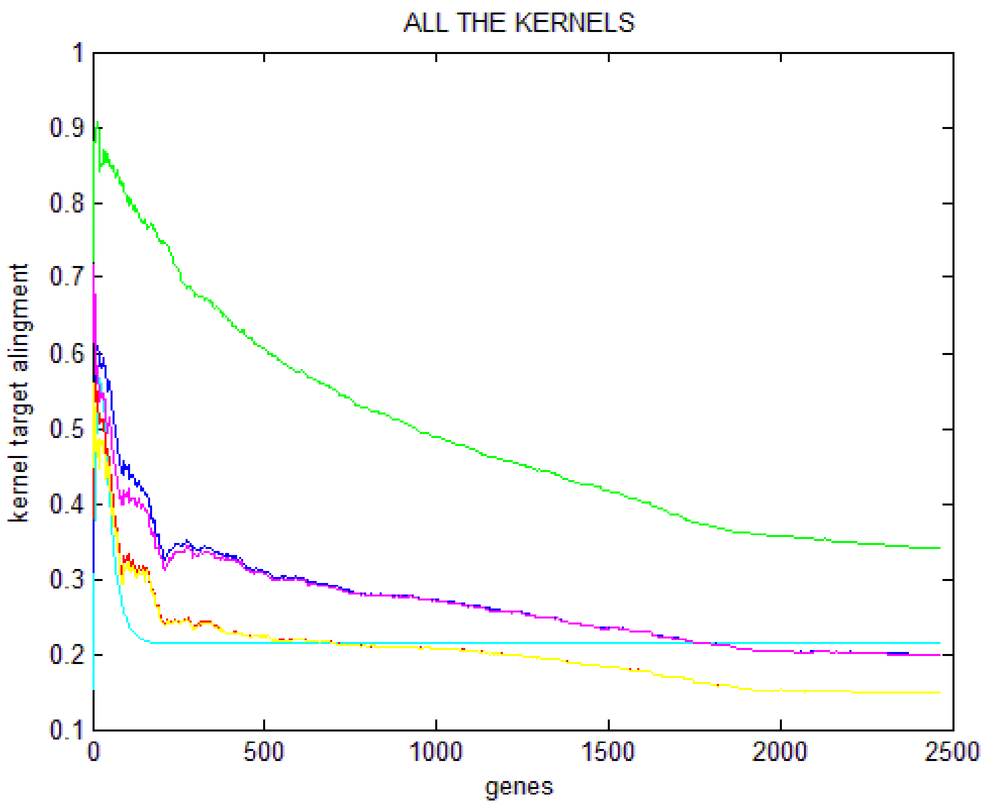


Εικόνα 33: Αποτέλεσμα της PCA ανάλυσης στο αρχικό σύνολο δεδομένων. Με τους μπλε σταυρούς παρουσιάζεται η κλάση elutriation, ενώ με τα κόκκινα αστέρια η κλάση heat-cold-dtt.

4.3.2 Επιλέγοντας έναν πυρήνα για τον κάθε ταξινομητή

Επειδή οι ταξινομητές που υλοποιούνται σε αυτή την εργασία βασίζονται σε πυρήνες, σχεδιάζονται οι πυρήνες Gaussian Radial Basis Function (RBF), γραμμικός, πολυωνυμικός δευτέρου και τρίτου βαθμού, τετραγωνικός και πολυωνυμικός τρίτου βαθμού για SVM, για να εφαρμοστούν σε αυτούς, όπως αυτοί προκύπτουν από τις σχέσεις (3.2.3.6) έως και (3.2.3.9), αντίστοιχα. Σκοπός είναι να αποφασιστεί ποιος πυρήνας θα εφαρμοστεί σε κάθε ταξινομητή και παράλληλα να απαντηθεί το ερώτημα αν το kernel target alignment, που περιγράφεται στην ενότητα 3.5.2, αποτελεί ένα καλό μέτρο εκτίμησης της απόδοσης ενός ταξινομητή.

Για το λόγο αυτό, αρχικά εξετάζουμε όλους τους ταξινομητές χρησιμοποιώντας σε καθέναν από αυτούς όλους τους πυρήνες. Μία πρώτη εκτίμηση για το ποιος,



Εικόνα 34: Γραφική αναπαράσταση του KTA για έξι πυρήνες. Με πράσινο χρώμα αναπαριστάται ο γραμμικός, με μπλε ο πολυωνυμικός δευτέρου βαθμού, με μοβ ο τετραγωνικός, με κόκκινο ο πολυωνυμικός τρίτου βαθμού, με κίτρινο ο πολυωνυμικός τρίτου βαθμού για svm ταξινομητή, και με κυανό ο Gaussian πυρήνας.

από τους διαθέσιμους πυρήνες, ευθυγραμμίζεται καλά με τα δεδομένα μας προκύπτει αν συγκρίνουμε τις τιμές του kernel target alignment όλων των πυρήνων.

Έτσι, εφαρμόζοντας τη διαδικασία για το αρχικό επίπεδο επεξεργασίας γονιδίων, όπως ακριβώς αυτή αναλύεται παραπάνω, υπολογίζεται και σχεδιάζεται το kernel target alignment για κάθε πυρήνα, από την εξίσωση (3.5.2.3), καθώς και η ακρίβεια ταξινόμησης όλων των ταξινομητών με όλους τους πυρήνες, από την εξίσωση (3.5.1.1), για όλα τα υποσύνολα γονιδίων, και για τα 5 folds.

Μετράμε τη μέση ακρίβεια ταξινόμησης καθώς και το μέσο kernel target alignment για όλα τα υποσύνολα γονιδίων, και για τα 5 folds, για όλους τους συνδυασμούς ταξινομητή-πυρήνα. Τα αποτελέσματα συνοψίζονται στον πίνακα 3 (έχει τονιστεί με έντονα γράμματα η μεγαλύτερη τιμή ανά γραμμή):

Πίνακας 3: Μέσες τιμές για το kernel target alignment και το classification accuracy για όλους τους συνδυασμούς ταξινομητών και πυρήνων, στο 1ο επίπεδο αξιολόγησης δεδομένων.

Mean of the 5 folds	*Gaussian RBF*	*Linear*	*Polynomial 2nd degree*	*Polynomial 3rd degree*	*Quadratic*	*Svm Polynomial 3rd degree*
Kernel Target Alignment	0,22	**0,48**	0,28	0,20	0,28	0,20
Accuracy for knc	44,09	68,26	**77,98**	76,95	-	-
Accuracy for kmse	55,87	**96,61**	92,60	94,44	-	-
Accuracy for ks	93,11	**96,62**	50	90,45	-	-
Accuracy for svm	-	**96,62**	-	-	88,19	56,69

Τα συμπεράσματα που εξάγονται τόσο από αυτές τις τιμές, όσο και από τη γραφική παράσταση με το kernel target alignment όλων των πυρήνων είναι τα εξής:

- ✓ Το kernel target alignment έχει την υψηλότερη τιμή για τον γραμμικό πυρήνα, ενώ μετά ακολουθούν οι πυρήνες πολυωνυμικός δευτέρου βαθμού και τετραγωνικός, με σχεδόν παρόμοια τιμή για το ΚΤΑ, οι οποίοι αποδεικνύονται ελαφρώς καλύτεροι από τον Gaussian και τον πολυωνυμικό τρίτου βαθμού.

Η ερμηνεία που προκύπτει από αυτές τις τιμές είναι ότι ένας γραμμικός πυρήνας ευθυγραμμίζεται πολύ καλύτερα με το σύνολο δεδομένων που διαθέτουμε και άρα είναι ο καταλληλότερος για να εφαρμοστεί στους ταξινομητές. Γενικά είναι γνωστό ότι, σε σύνολα δεδομένων με πολύ λίγα δείγματα αλλά πολλά χαρακτηριστικά, όπως είναι το σύνολο δεδομένων αυτής της εργασίας, είναι καλύτεροι οι γραμμικοί πυρήνες γιατί οι μη γραμμικοί μπορεί να οδηγήσουν σε overfitting (το φαινόμενο όπου ο ταξινομητής προσαρμόζεται τόσο καλά στα δεδομένα εκπαίδευσης χάνοντας έτσι την ικανότητα γενίκευσης).

- ✓ Η ακρίβεια ενός support vector machine συμφωνεί απολύτως με τις τιμές του ΚΤΑ, οπότε εμφανίζει αντίστοιχα πολύ υψηλή ακρίβεια με τον γραμμικό πυρήνα, λίγο χαμηλότερη με τον τετραγωνικό και αρκετά χαμηλή με τον πολυωνυμικό πυρήνα τρίτου βαθμού.
- ✓ Η ακρίβεια της kernel nearest centroid μεθόδου είναι αρκετά υψηλή με τον πολυωνυμικό πυρήνα δευτέρου και τρίτου βαθμού, ενώ μειώνεται με τον γραμμικό πυρήνα και παίρνει πολύ χαμηλή τιμή με τον Gaussian πυρήνα.
- ✓ Η ακρίβεια της kernel minimum squared error μεθόδου είναι αρκετά χαμηλή με έναν Gaussian πυρήνα, όμως με έναν γραμμικό και πολυωνυμικό δευτέρου και τρίτου βαθμού είναι πολύ υψηλή.
- ✓ Η ακρίβεια της kernel subspace μεθόδου είναι πολύ υψηλή με τον Gaussian, γραμμικό και πολυωνυμικό τρίτου βαθμού πυρήνα, αλλά πολύ χαμηλή για τον πολυωνυμικό δευτέρου βαθμού.
- ✓ Παρατηρούμε ότι η απόδοση ενός ταξινομητή δεν εξαρτάται μόνο από το μηχανισμό του ίδιου του ταξινομητή (μορφή συνάρτησης απόφασης) αλλά και από τον πυρήνα που θα εφαρμοστεί σε αυτόν, καθώς η απόδοσή του μεταβάλλεται με τη χρήση διαφορετικών πυρήνων. Προσαρμόζοντας έναν πυρήνα [48], κατάλληλο, για να βελτιωθεί η ευθυγράμμιση με τα δείγματα του συνόλου εκπαίδευσης, τα οποία φέρουν ετικέτα, αυξάνεται σημαντικά η ευθυγράμμιση και με τα δείγματα του συνόλου ελέγχου, τα οποία καλούνται τα μοντέλα μάθησης να ταξινομήσουν, δίνοντας αρκετά βελτιωμένη ακρίβεια

ταξινόμησης. Οπότε, η επιλογή του κατάλληλου πυρήνα σε κάθε ταξινομητή είναι ζωτικής σημασίας για την ακρίβεια που θα σημειώσει στην ταξινόμηση.

- ✓ Παρατηρώντας τη γραφική παράσταση της εικόνας 34, βλέπουμε ότι το ΚΤΑ, για όλους τους πυρήνες, αυξάνεται στα τελευταία γονίδια, δηλαδή όταν έχουν αφαιρεθεί από το σύνολο δεδομένων αρκετά, μη σχετικά γονίδια, ο κάθε πυρήνας ευθυγραμμίζεται καλύτερα με τα δείγματα.

Καταλήγουμε στο ότι οι ιδανικοί συνδυασμοί που προκύπτουν είναι οι kernel nearest centroid μέθοδος με πολυωνυμικό πυρήνα δευτέρου βαθμού, kernel minimum squared error μέθοδος με γραμμικό πυρήνα, kernel subspace μέθοδος επίσης με γραμμικό πυρήνα και support vector machine πάλι με γραμμικό πυρήνα.

Επειδή όμως σε αυτή την εργασία θέλουμε να αναδείξουμε την εφαρμογή διαφορετικών πυρήνων σε διαφορετικούς ταξινομητές, και επειδή ο γραμμικός πυρήνας αποτελεί τη βέλτιστη επιλογή για τους περισσότερους των ταξινομητών, αποφασίσαμε να επιλέξουμε διαφορετικά ζευγάρια ταξινομητών-πυρήνων, προσπαθώντας πάλι οι επιλογές αυτές να είναι ανάμεσα στους καλύτερους συνδυασμούς. Στο υπόλοιπο της εργασίας χρησιμοποιούνται τα εξής ζευγάρια:

- ✓ <u>kernel nearest centroid μέθοδος με πολυωνυμικό πυρήνα δευτέρου βαθμού</u>
- ✓ <u>kernel minimum squared error μέθοδος με Gaussian πυρήνα</u>
- ✓ <u>kernel subspace μέθοδος με πολυωνυμικό πυρήνα τρίτου βαθμού</u>
- ✓ <u>support vector machine με γραμμικό πυρήνα</u>
- ✓ <u>support vector machine με τετραγωνικό πυρήνα</u>

Με αυτές τις επιλογές υπάρχει ποικιλία στη χρήση των πυρήνων στους ταξινομητές αφού κάθε πυρήνας εμφανίζεται μία φορά σε κάποιο ταξινομητή.

4.3.3 Επιλογή γονιδίων σύμφωνα με την απόδοση ταξινόμησης

Αφού έχουμε καταλήξει σε έναν πυρήνα για κάθε ταξινομητή, σκοπός τώρα είναι η εξαγωγή των αποτελεσμάτων για το σύνολο δεδομένων, όταν αυτό βρίσκεται στην αρχική του μορφή, όταν, δηλαδή, διαθέτει και τα 2467 γονίδια. Εφαρμόζεται πάλι η διαδικασία για το πρώτο επίπεδο αξιολόγησης δεδομένων, για όλα τα μοντέλα μάθησης και πρόβλεψης που αναφέρονται στην ενότητα 4.2, συμπεριλαμβανομένων και των επιλογών που έγιναν για τους πυρήνες των ατομικών ταξινομητών στην υποενότητα 4.3.2.

Πιο συγκεκριμένα, σε κάθε κύκλο επανάληψης του five fold cross validation, τα μοντέλα εκπαιδεύονται με το σύνολο δεδομένων εκπαίδευσης και στη συνέχεια ελέγχεται η απόδοσή τους πάνω σε διαφορετικά δείγματα, με το σύνολο δεδομένων ελέγχου. Σε κάθε έναν από τους πέντε κύκλους επανάληψης υπολογίζονται, για κάθε μοντέλο, ο μέσος όρος της ακρίβειας ταξινόμησης για όλα τα υποσύνολα γονιδίων που προκύπτουν από τον RFE-LNW αλγόριθμο, η μέγιστη ακρίβεια ταξινόμησης καθώς και ο ελάχιστος αριθμός γονιδίων του υποσυνόλου στο οποίο αυτή παρουσιάζεται. Στο τέλος και των πέντε επαναλήψεων υπολογίζεται η μέση τιμή των πέντε μέσων όρων της ακρίβειας ταξινόμησης, των πέντε μεγίστων τιμών ακρίβειας ταξινόμησης και των πέντε αριθμών ελαχίστων γονιδίων που αντιστοιχούν στις πέντε μέγιστες ακρίβειες.

Πίνακας 4: Η μέση ακρίβεια ταξινόμησης στο 1^{o} επίπεδο αξιολόγησης δεδομένων, για τους πέντε ατομικούς ταξινομητές και για το μείγμα αυτών που βασίζεται σε συγχώνευση με συνδυασμό ασαφών βαρών.

MEAN VALUES FOR 5 FOLDS	*MoE based on a fuzzy formulation*	*knc polynomial 2nd degree*	*kmse gaussian*	*ks polynomial 3rd degree*	*svm quadratic*	*svm linear*
MEAN CLASSIFICATION ACCURACY FOR ALL THE SUBSETS OF GENES	76,39	88,47	52,25	94,94	86,14	96
MAX ACCURACY	100	92,66	100	100	100	100
MIN NUMBER OF GENES WITH THE MAX ACCURACY	23	100	5	7	4	5

<u>Τα αποτελέσματα για τον πρώτο συνδυασμό ταξινομητών καθώς και για τους πέντε ατομικούς ταξινομητές που τον αποτελούν συνοψίζονται στον πίνακα 4.</u>

- ✓ Το μείγμα των ειδικών που βασίζεται σε συγχώνευση με συνδυασμό ασαφών βαρών επιτυγχάνει μία αρκετά καλή μέση ακρίβεια ταξινόμησης. Δεν είναι πιο υψηλή από τη μέση ακρίβεια του καλύτερου ταξινομητή, ούτε μεγαλύτερη από το μέσο όρο των μέσων ακριβειών και των πέντε ατομικών ταξινομητών, όμως δεν είναι πάντα αυτός ο στόχος όπως έχει αναφερθεί ήδη στην ενότητα 3.3. Μπορεί ένα μείγμα ειδικών να μην υπερέχει ως προς την απόδοση της ταξινόμησης που επιτυγχάνει σε σχέση με τους ατομικούς ταξινομητές, αλλά ως προς τη σταθερότητά του όταν εφαρμόζεται σε διαφορετικά σύνολα δεδομένων. Όλοι οι ατομικοί ταξινομητές, εκτός από τον kernel minimum squared error με Gaussian πυρήνα, παρουσιάζουν αρκετά υψηλή μέση τιμή ακρίβειας ταξινόμησης.
- ✓ Η μέση μέγιστη ακρίβεια για τα πέντε folds, την οποία σημειώνουν σχεδόν όλα τα μοντέλα είναι 100% και ο ελάχιστος αριθμός γονιδίων για τον οποίο αυτή επιτυγχάνεται, όπως φαίνεται από την τελευταία γραμμή του πίνακα, είναι πολύ μικρός. Μόνο ο kernel nearest centroid με πολυωνυμικό πυρήνα δευτέρου βαθμού επιτυγχάνει μέση μέγιστη ακρίβεια 92,66% για έναν σχετικά πιο μεγάλο αριθμό γονιδίων (100).

<u>Τα αποτελέσματα για το δεύτερο συνδυασμό ειδικών καθώς και για τους δύο ατομικούς ταξινομητές που τον αποτελούν συνοψίζονται στον πίνακα 5.</u>

- ✓ Στη δεύτερη περίπτωση εφαρμόζουμε συγχώνευση ταξινομητών οι οποίοι εκπαιδεύονται με έναν gradient descent learning algorithm, για τη δημιουργία του μείγματος ειδικών, και παρατηρούμε ότι και εδώ η μέση ακρίβεια ταξινόμησης του συνδυασμού είναι υψηλότερη από αυτή των δύο ατομικών ταξινομητών, χωρίς όμως να είναι υψηλή. Οι δύο ατομικοί ταξινομητές, όπως έχει προαναφερθεί, είναι δύο logistic sigmoid ταξινομητές οι οποίοι κάθε φορά ανανεώνουν τα βάρη τους βασιζόμενοι στο σφάλμα της απόφασης που έλαβε ο συνδυασμός τους και όχι στο σφάλμα της δικής τους απόφασης, κάτι το οποίο εξηγεί τη χαμηλή ακρίβεια που εμφανίζουν.

Πίνακας 5: Η μέση ακρίβεια ταξινόμησης στο 1ο επίπεδο αξιολόγησης δεδομένων, για τους δύο ατομικούς ταξινομητές και για το μείγμα αυτών που βασίζεται σε συγχώνευση με ένα gradient descent learning αλγόριθμο.

MEAN VALUES FOR 5 FOLDS	*MoE based on a gradient*	*1st expert*	*2nd expert*
MEAN CLASSIFICATION ACCURACY FOR ALL THE SUBSETS OF GENES	55,17	48,37	48,52
MAX ACCURACY	88,33	88,33	88,33
MIN NUMBER OF GENES WITH THE MAX ACCURACY	3	3	3

- ✓ Παρατηρούμε ότι ενώ η μέση ακρίβεια τόσο για το συνδυασμό όσο και για τους δύο ειδικούς είναι αρκετά χαμηλή, πάντα φτάνουν σε μία μέση μέγιστη ακρίβεια της τάξης του 88,33% για έναν πολύ μικρό αριθμό γονιδίων (3).

<u>Τα αποτελέσματα για τον τρίτο συνδυασμό ταξινομητών καθώς και για τους δύο ατομικούς ταξινομητές που τον αποτελούν συνοψίζονται στον πίνακα 6.</u>

- ✓ Στην τέταρτη περίπτωση εφαρμόζουμε πάλι το μείγμα των ειδικών που βασίζεται σε συγχώνευση με συνδυασμό ασαφών βαρών, αλλά αυτή τη φορά με δύο logistic sigmoid ταξινομητές να αποτελούν τους ειδικούς και όχι τους πέντε ατομικούς ταξινομητές πυρήνα που είδαμε στην πρώτη περίπτωση. Οι δύο αυτοί ταξινομητές, που έχουν προκύψει από έναν gradient descent learning αλγόριθμο, ανανεώνουν τα βάρη τους με βάση το δικό τους σφάλμα απόφασης (σε αντίθεση με την τρίτη περίπτωση). Παρατηρούμε ότι όλες οι μέσες τιμές απόδοσης είναι αρκετά υψηλές και ότι η μέση τιμή της ακρίβειας ταξινόμησης του συνδυασμού βρίσκεται ανάμεσα στις τιμές της μέσης ακρίβειας των δύο ατομικών ταξινομητών.

Πίνακας 6: Η μέση ακρίβεια ταξινόμησης στο 1^{ο} επίπεδο αξιολόγησης δεδομένων, για τους δύο ατομικούς ταξινομητές και για το μείγμα αυτών των ειδικών που βασίζεται σε ένα gradient descent learning αλγόριθμο και σε συνδυασμό ασαφών βαρών.

MEAN VALUES FOR 5 FOLDS	*MoE based on a fuzzy formulation for 2 logsigs*	*1st expert*	*2nd expert*
MEAN CLASSIFICATION ACCURACY FOR ALL THE SUBSETS OF GENES	86,42	91,61	83,79
MAX ACCURACY	100	100	100
MIN NUMBER OF GENES WITH THE MAX ACCURACY	4	4	4

- ✓ Η μέση μέγιστη ακρίβεια στην οποία φτάνουν τόσο το μείγμα ειδικών όσο και οι δύο ειδικοί είναι 100% σε όλα τα folds και ο ελάχιστος αριθμός γονιδίων, για τον οποίο αυτή μπορεί να επιτευχθεί, είναι τα 4 γονίδια.

<u>Τα αποτελέσματα για τον τέταρτο συνδυασμό ταξινομητών καθώς και για τους δύο ατομικούς ταξινομητές που τον αποτελούν συνοψίζονται στον πίνακα 7.</u>

- ✓ Το μείγμα ειδικών που βασίζεται στην επιλογή ταξινομητών με som διαχωρισμό του χώρου χαρακτηριστικών, επιτυγχάνει πολύ μεγαλύτερη μέση ακρίβεια ταξινόμησης από τους δύο ατομικούς ειδικούς. Αυτό μπορεί να εξηγηθεί από τον ίδιο τον ορισμό της επιλογής ταξινομητών, αφού οι δύο ατομικοί ταξινομητές <<βλέπουν>> κάθε φορά ένα μόνο κομμάτι του συνόλου δεδομένων εκπαίδευσης το οποίο, όπως φαίνεται, δεν είναι πάντα αντιπροσωπευτικό όλων των δειγμάτων ώστε να τους παρέχει μία ολοκληρωμένη εκπαίδευση. Έτσι η μέση ακρίβεια της πρόβλεψής τους δεν είναι πολύ υψηλή, όμως στο συνδυασμό τους, που αλληλοσυμπληρώνονται, παρουσιάζεται αρκετά υψηλή μέση ακρίβεια ταξινόμησης.

Πίνακας 7: Η μέση ακρίβεια ταξινόμησης στο 1ο επίπεδο αξιολόγησης δεδομένων, για τους δύο ατομικούς ταξινομητές και για το μείγμα αυτών που βασίζεται στην επιλογή ειδικών με som διαχωρισμό του χώρου χαρακτηριστικών.

MEAN VALUES FOR 5 FOLDS	*MoE based on som formulation*	*1st expert*	*2nd expert*
MEAN CLASSIFICATION ACCURACY FOR ALL THE SUBSETS OF GENES	88,87	62,68	62,64
MAX ACCURACY	100	96,66	96,66
MIN NUMBER OF GENES WITH THE MAX ACCURACY	10	540	801

- ✓ Η μέση μέγιστη ακρίβεια στην οποία φτάνει ο συνδυασμός των ειδικών είναι 100% και ο ελάχιστος αριθμός γονιδίων για τον οποίο αυτή μπορεί να επιτευχθεί είναι τα 10 γονίδια. Για τους δύο ατομικούς ταξινομητές παρατηρούμε ότι παρόλο που η μέση μέγιστη ακρίβειά τους είναι αρκετά υψηλή, επιτυγχάνεται για πολύ μεγάλο αριθμό γονιδίων, οπότε δεν τους χρησιμοποιούμε περαιτέρω στο υπόλοιπο της εργασίας, για την επιλογή γονιδίων.

Τα αποτελέσματα και για τους τέσσερις συνδυασμούς ταξινομητών συνοψίζονται στον πίνακα 8.

- ✓ Συγκρίνοντας τη μέση ακρίβεια ταξινόμησης και των τεσσάρων μειγμάτων ειδικών, οι οποίοι βέβαια, όπως αναλύθηκε παραπάνω, έχουν προκύψει από τους συνδυασμούς διαφορετικών ταξινομητών, παρατηρούμε ότι όλοι τους επιτυγχάνουν μία πολύ καλή μέση ακρίβεια ταξινόμησης που κυμαίνεται στο 76% - 88%, εκτός από το δεύτερο συνδυασμό που βασίζεται σε gradient descent, ο οποίος εμφανίζει χαμηλή μέση ακρίβεια. Κατατάσσοντάς τους, παρατηρούμε ότι πρώτος έρχεται ο συνδυασμός που βασίζεται στην επιλογή ταξινομητών με som διαχωρισμό του χώρου χαρακτηριστικών, και έχει προκύψει από τη συμπληρωματικότητα των επιμέρους τοπικών ειδικών, και μετά όλοι οι άλλοι συνδυασμοί που βασίζονται στη συγχώνευση των επιμέρους καθολικών ειδικών. Μέση μέγιστη ακρίβεια ίση με 100% επιτυγχάνουν όλοι οι συνδυασμοί για κάποιο ελάχιστο αριθμό γονιδίων που είναι πολύ μικρός, εκτός από το τρίτο μείγμα που βασίζεται σε gradient descent και που εμφανίζει μέση μέγιστη ακρίβεια, τη χαμηλότερη από όλους.

Πίνακας 8: Η μέση ακρίβεια ταξινόμησης στο 1^{ο} επίπεδο αξιολόγησης δεδομένων, για τους τέσσερις συνδυασμούς ειδικών.

MEAN VALUES FOR 5 FOLDS	*MoE based on a fuzzy formulation*	*MoE based on a gradient*	*MoE based on a fuzzy formulation for 2 logsigs*	*MoE based on som formulation*
MEAN CLASSIFICATION ACCURACY FOR ALL THE SUBSETS OF GENES	76,39	55,17	86,42	88,87
MAX ACCURACY	100	88,33	100	100
MIN NUMBER OF GENES WITH THE MAX ACCURACY	23	3	4	10

Αναλυτικά, όλες οι γραφικές παραστάσεις και των πέντε folds που απεικονίζουν τις τιμές της απόδοσης, για τους 5 ατομικούς ταξινομητές και για το πρώτο μείγμα αυτών, παρατίθενται στο παράρτημα Α.

Σαν ένα γενικό συμπέρασμα μπορούμε να πούμε ότι το μείγμα ειδικών που εμφανίζει την καλύτερη ακρίβεια ταξινόμησης, τόσο σε σύγκριση με τα υπόλοιπα μείγματα, όσο και σε σύγκριση με τους ατομικούς ταξινομητές που αυτό συνδυάζει, είναι το μείγμα που χρησιμοποιεί την επιλογή ταξινομητών και όχι κάποιο από τα μείγματα που χρησιμοποιούν τη συγχώνευση ταξινομητών. Με άλλα λόγια, το μείγμα που βασίζεται στην εκπαίδευση των επιμέρους ειδικών σε διαφορετικά υποσύνολα, τα οποία προκύπτουν με som διαχωρισμό του χώρου χαρακτηριστικών, αποδίδει πολύ καλύτερα από τους επιμέρους τοπικούς ειδικούς, αφού βασίζεται στη συμπληρωματικότητα τους, κάτι το οποίο δεν παρατηρείται πάντα για τα υπόλοιπα μείγματα που βασίζονται στη συγχώνευση των επιμέρους καθολικών ειδικών, όπου επικρατεί ανταγωνιστικότητα μεταξύ τους για τη δημιουργία της συνδυασμένης απόφασης.

Από τους πίνακες 4 έως και 7 που παρατέθηκαν παραπάνω, γίνεται αναφορά, στην τελευταία γραμμή τους, στον μέσο ελάχιστο αριθμό γονιδίων, από τα 2467, στον οποίο το κάθε μοντέλο παρουσιάζει τη μέση μέγιστη ακρίβεια ταξινόμησης. Αυτοί οι αριθμοί αποτελούν την αρχή για το πρώτο επίπεδο επιλογής γονιδίων που περιγράφεται στην επόμενη ενότητα 4.4 και συγκεντρώνονται στον πίνακα 9.

Πίνακας 9: Οι μέσοι ελάχιστοι αριθμοί γονιδίων A για όλους τους ατομικούς ταξινομητές και για όλα τα μείγματα συγκεντρωτικά, έτσι όπως αυτά προκύπτουν μετά το 1^{o} επίπεδο αξιολόγησης δεδομένων.

All the classifiers and the mixtures	Mean Value of the min number of genes with the max accuracy (A) for the 5 folds.
knc polynomial 2nd degree	100
kmse gaussian	5
ks polynomial 3rd degree	7
svm quadratic	4
svm linear	5
MoE based on a fuzzy formulation	23
MoE based on a gradient	3
MoE based on a fuzzy formulation for 2 logsigs	4
MoE based on som formulation	10

4.4 Πρώτο Επίπεδο Επιλογής Γονιδίων

Από το πρώτο επίπεδο αξιολόγησης δεδομένων, που περιγράφεται στην ενότητα 4.3, εξάγουμε για κάθε μοντέλο έναν αριθμό γονιδίων A (πίνακας 9), που αποφέρει μεγάλη ακρίβεια ταξινόμησης. Όμως δεν έχουμε καταλήξει σε συγκεκριμένα γονίδια, αφού το σύνολο των βέλτιστων γονιδίων αλλάζει για τα 5 folds και ο αριθμός A, για κάθε μοντέλο, είναι ένας μέσος όρος των 5 ελαχίστων αριθμών που προκύπτουν από τα 5 folds. Σε αυτό το πρώτο επίπεδο επιλογής γονιδίων, εξετάζουμε όλα τα γονίδια που έχουν επιλεγεί από τα 5 folds και τα αξιολογούμε ξανά εφαρμόζοντας την προηγούμενη διαδικασία πολλαπλές φορές, για να καταλήξουμε στο τελικό σετ των επιλεγμένων γονιδίων.

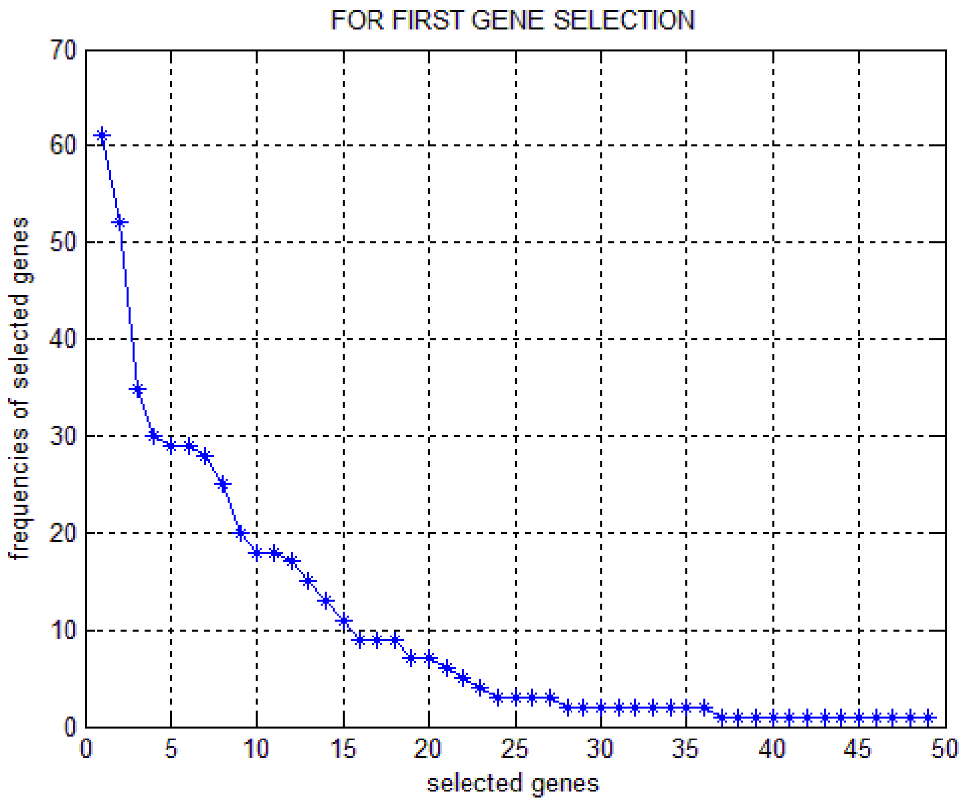
Στο πρώτο επίπεδο επιλογής γονιδίων, εκτελείται μία αρχική επιλογή γονιδίων, από τα 2467 που υπάρχουν στο σύνολο δεδομένων μας, η οποία βασίζεται σε δύο πληροφορίες. Αρχικά, στο μέσο αριθμό των ελαχίστων γονιδίων για τον οποίο το μοντέλο πέτυχε τη μεγαλύτερη μέση απόδοση ταξινόμησης, κατά το πρώτο επίπεδο αξιολόγησης δεδομένων της ενότητας 4.3, και έπειτα στις συχνότητες εμφάνισης που τα γονίδια αποκτούν όταν εφαρμοστεί ο αλγόριθμος RFE-LNW πολλαπλές φορές πάνω στο σύνολο δεδομένων μας. Επιλέγονται τα γονίδια με τις υψηλότερες συχνότητες σύμφωνα με ένα όριο τιμών (γόνατο γραφικής παράστασης).

<u>Πρώτη επιλογή γονιδίων σύμφωνα με τη συχνότητα εμφάνισής τους.</u>

Για να γίνει πιο κατανοητή η διαδικασία που ακολουθείται στο πρώτο επίπεδο επιλογής γονιδίων, ακολουθεί η αναλυτική περιγραφή της και η παρουσίαση των αποτελεσμάτων που προκύπτουν σε κάθε βήμα της, για το μοντέλο support vector machine με γραμμικό πυρήνα. Όμοια είναι η διαδικασία που έχει ακολουθηθεί και για τα υπόλοιπα μοντέλα μάθησης και πρόβλεψης που παρουσιάζονται στην ενότητα 4.2.

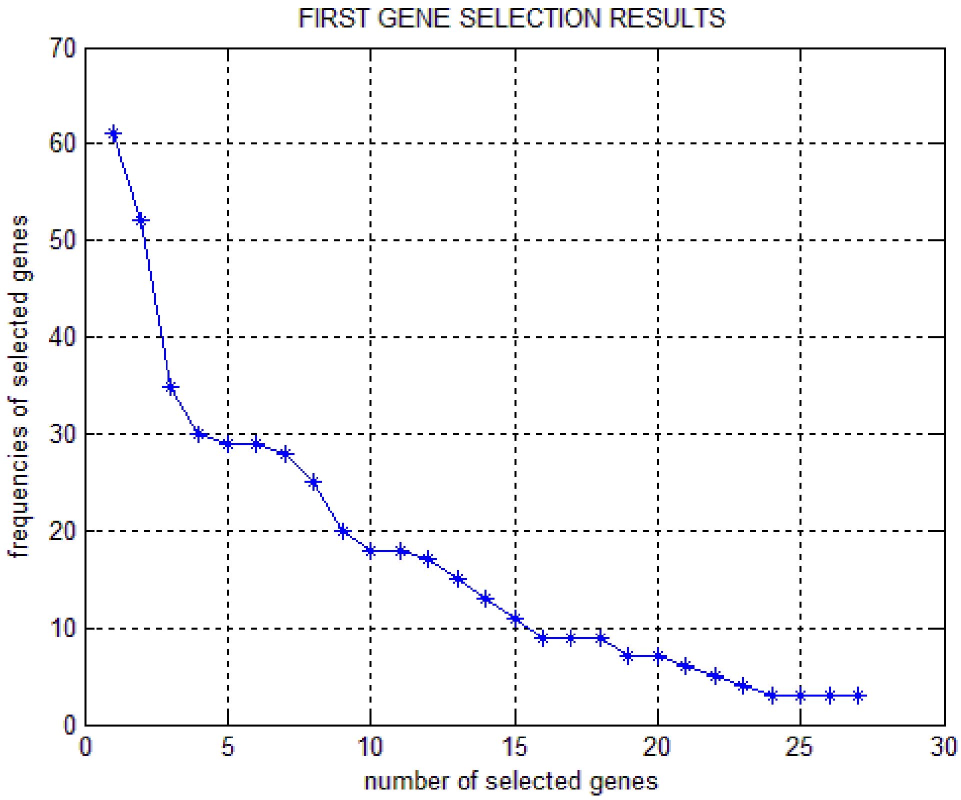
Από τον πίνακα 4 της υποενότητας 4.3.3, φαίνεται ότι ο μέσος αριθμός ελαχίστων γονιδίων, για τα οποία το γραμμικό support vector machine επιτυγχάνει τη μέση υψηλότερη απόδοση ταξινόμησης, μετά το τέλος του αρχικού επιπέδου διαλογής των γονιδίων, είναι τα 5 γονίδια (<u>*ή πιο γενικά, έστω A γονίδια*</u>). Εκτελούνται 20 επαναλήψεις των βημάτων 1 και 2, για το διαχωρισμό των δεδομένων με 5-fold cross validation και τη μείωση των γονιδίων με τον αλγόριθμο RFE-LNW αντίστοιχα, της διαδικασίας της ενότητας 4.3, με δύο διαφοροποιήσεις στο βήμα 2. Παραλείπεται από το βήμα 2 το σημείο της εκπαίδευσης και άρα και του ελέγχου του μοντέλου, αφού σε αυτή φάση δεν έχει ολοκληρωθεί ακόμα η επιλογή των γονιδίων.

Επιπλέον, η συνθήκη που εκτελείται σε κάθε fold, για τον αλγόριθμο RFE-LNW, δεν είναι πλέον «από 2467 έως και 1 γονίδια», όπως υποδεικνύει το βήμα 2 του πίνακα 1 στην υποενότητα 3.4.6, αλλά αλλάζει στη συνθήκη «από 2467 έως και 5 γονίδια». Δηλαδή, εκτελούνται (5 folds x 20 επαναλήψεις= 100) 100 runs του RFE-LNW, με σκοπό όχι την αναδρομική εξάλειψη όλων των γονιδίων όπως στο πρώτο επίπεδο αξιολόγησης δεδομένων της ενότητας 4.3, αλλά την αναδρομική εξάλειψη των (2467-5=2462) 2462 γονιδίων κάθε φορά. Τα 5 γονίδια που διατηρούνται και δεν εξαλείφονται σε κάθε επανάληψη, ελέγχονται πριν εισαχθούν σε μία καθολική λίστα συχνοτήτων. Αν τα γονίδια υπάρχουν ήδη στη λίστα από προηγούμενη επανάληψη, δεν εισάγονται ξανά, αλλά αυξάνεται η συχνότητά τους κατά 1, ενώ αν είναι καινούρια προστίθενται σε αυτήν με συχνότητα 1.



Εικόνα 35: Τα 49 γονίδια με τις συχνότητες τους, όπως αυτές προκύπτουν από το πρώτο επίπεδο επιλογής γονιδίων, για το γραμμικό support vector machine.

Στο τέλος όλων αυτών των 100 επαναλήψεων, προκύπτουν 49 γονίδια (*έστω B γονίδια*) με τις συχνότητές τους, οι οποίες απεικονίζονται στην εικόνα 35. Σημειώνουμε εδώ ότι, η μέγιστη τιμή που μπορεί να πάρει η συχνότητα ενός οποιουδήποτε γονιδίου είναι 100. Και αυτό γιατί εκτελούνται συνολικά 100 runs και μπορεί να τύχει κάποιο γονίδιο να είναι και τις 100 φορές μέσα στα 5 που κρατιούνται σε κάθε επανάληψη.



Εικόνα 36: Τα 27 γονίδια μαζί με τις συχνότητές τους, έτσι όπως διαμορφώνονται μετά από την ολοκλήρωση του πρώτου επιπέδου επιλογής γονιδίων, για το γραμμικό support vector machine.

Από τη γραφική παράσταση των συχνοτήτων των γονιδίων της εικόνας 35, είναι εύκολο να παρατηρηθεί ότι στα 27 γονίδια (*έστω C γονίδια*) δημιουργείται «γόνατο». Οπότε, από τα 49 γονίδια κρατιούνται μόνο τα 27 με τις υψηλότερες συχνότητες και απεικονίζονται σε μία νέα γραφική παράσταση που φαίνεται στην εικόνα 36. Αυτά τα 27 γονίδια (C γονίδια) αξιολογούνται ως προς την απόδοση ταξινόμησης που αποφέρουν στο δεύτερο επίπεδο αξιολόγησης δεδομένων που περιγράφεται στην ενότητα 4.5.

Τέλος, από τη νέα γραφική παράσταση των 27 γονιδίων της εικόνας 36, βρίσκουμε πάλι το «γόνατο» των συχνοτήτων, το οποίο παρατηρείται στα 20 γονίδια (*έστω D γονίδια*). Αυτός ο αριθμός γονιδίων που εξάγεται τελικά από το πρώτο επίπεδο επιλογής γονιδίων, αποτελεί το κριτήριο τερματισμού για το δεύτερο επίπεδο επιλογής γονιδίων, το οποίο περιγράφεται στην ενότητα 4.6.

Αναλυτικά, η διαδικασία για το πρώτο επίπεδο επιλογής γονιδίων, περιγράφεται στο διάγραμμα ροής της εικόνας 37.

Ο αριθμός των γονιδίων C που επιλέγονται ως τα πιο σημαντικά μετά το πρώτο επίπεδο επιλογής γονιδίων, και που θα αξιολογηθούν στο δεύτερο επίπεδο αξιολόγησης της ενότητας 4.5, ως προς την απόδοση ταξινόμησης που αποφέρουν, φαίνεται στον πίνακα 10, συγκεντρωτικά για όλα τα μοντέλα ταξινόμησης.

Πίνακας 10: Ο αριθμός των γονιδίων C που επιλέγονται ως τα πιο σημαντικά μετά το 1ο επίπεδο επιλογής γονιδίων, συγκεντρωτικά για όλα τα μοντέλα ταξινόμησης.

All the classifiers and the mixtures	Number of genes selected after the first selection level (C)
Knc polynomial 2nd degree	350
Kmse gaussian	23
Ks polynomial 3rd degree	30
Svm quadratic	25
Svm linear	27
MoE based on a fuzzy formulation	80
MoE based on a gradient	27
MoE based on a fuzzy formulation for 2 logsigs	26
MoE based on a som formulation	35

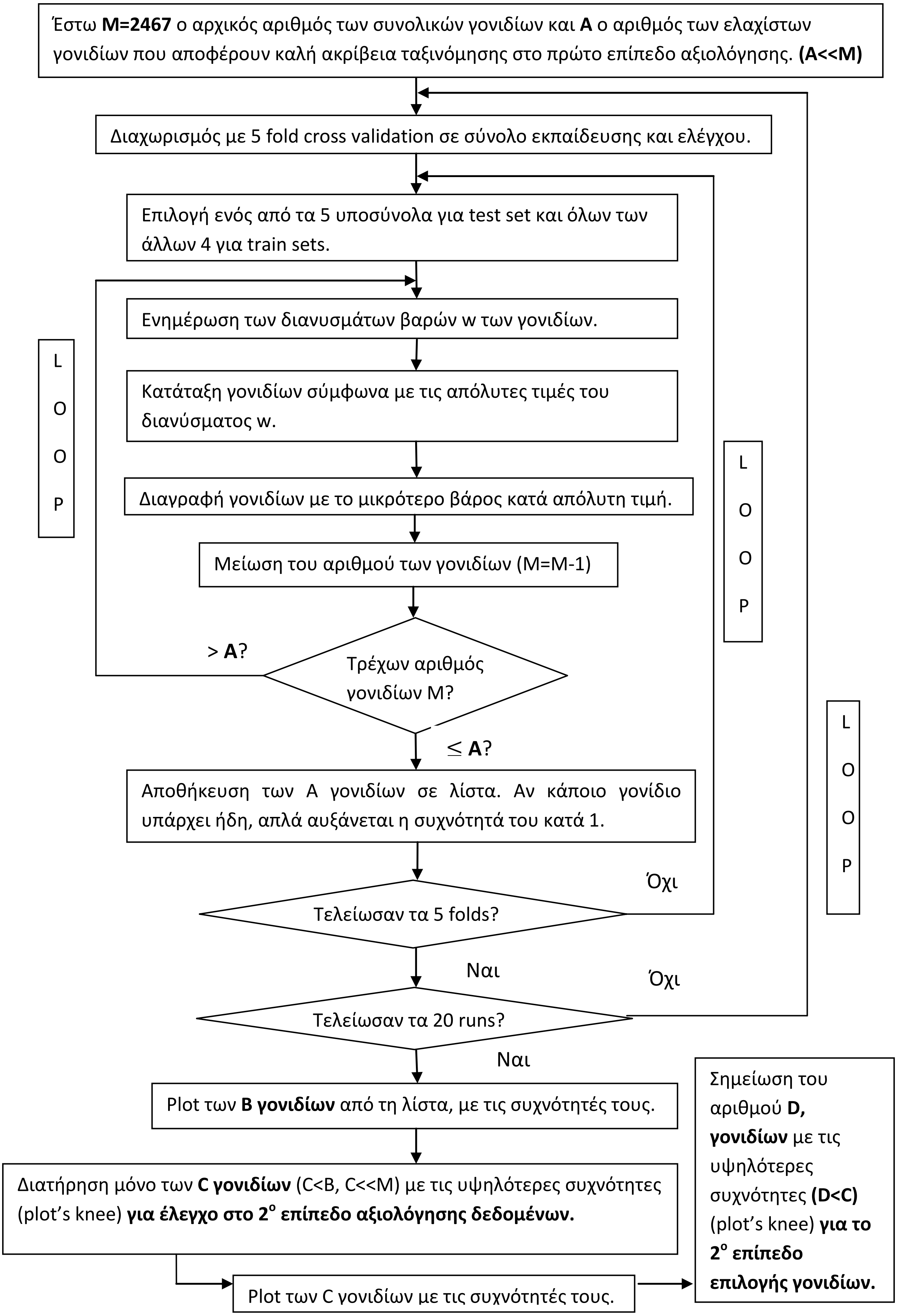
Ο αριθμός γονιδίων D που εισάγεται στο δεύτερο επίπεδο επιλογής γονιδίων της ενότητας 4.6, συγκεντρώνεται για το κάθε μοντέλο, στον πίνακα 11.

Πίνακας 11: Ο αριθμός των γονιδίων D που εισάγονται στο 2ο επίπεδο επιλογής γονιδίων, συγκεντρωτικά για όλα τα μοντέλα ταξινόμησης.

All the classifiers and the mixtures	Number of genes selected after the first selection level (D)
Knc polynomial 2nd degree	275
Kmse gaussian	20
Ks polynomial 3rd degree	22
Svm quadratic	20
Svm linear	20
MoE based on a fuzzy formulation	47
MoE based on a gradient	20
MoE based on a fuzzy formulation for 2 logsigs	20
MoE based on a som formulation	25

Από τον πίνακα 5 της ενότητας 4.3 με τη μέση απόδοση ταξινόμησης του δεύτερου μείγματος ειδικών βασισμένο σε συγχώνευση με ένα gradient descent learning αλγόριθμο, φαίνεται ότι ο ελάχιστος αριθμός γονιδίων (A) με τη μέση μέγιστη απόδοση κατά το πρώτο επίπεδο αξιολόγησης δεδομένων είναι ο ίδιος τόσο για το μείγμα όσο και για τους δύο ατομικούς ταξινομητές. Ομοίως, για τον πίνακα 6 της ενότητας 4.3 που αφορά στο τρίτο μείγμα ειδικών κάνουμε την ίδια παρατήρηση. Οπότε, για τους επιμέρους ταξινομητές του δεύτερου και του τρίτου μείγματος δεν προχωρήσαμε σε επιλογή γονιδίων, παρά μόνο για τα μείγματα αυτών θεωρώντας ότι αυτοί καταλήγουν στα ίδια επιλεγμένα γονίδια με το μείγμα (λόγω ίδιου αριθμού A).

Από τον πίνακα 7 της ενότητας 4.3 για το τέταρτο μείγμα ειδικών που βασίζεται σε επιλογή ειδικών με som διαχωρισμό του χώρου των χαρακτηριστικών, παρατηρούμε ότι ο ελάχιστος αριθμός γονιδίων (A) με τη μέση μέγιστη απόδοση κατά το πρώτο επίπεδο αξιολόγησης δεδομένων για τους δύο επιμέρους ειδικούς είναι πολύ μεγάλος. Ως εκ τούτου, δεν ασχοληθήκαμε περαιτέρω με αυτούς τους ταξινομητές, παρά μόνο με το μείγμα αυτών.

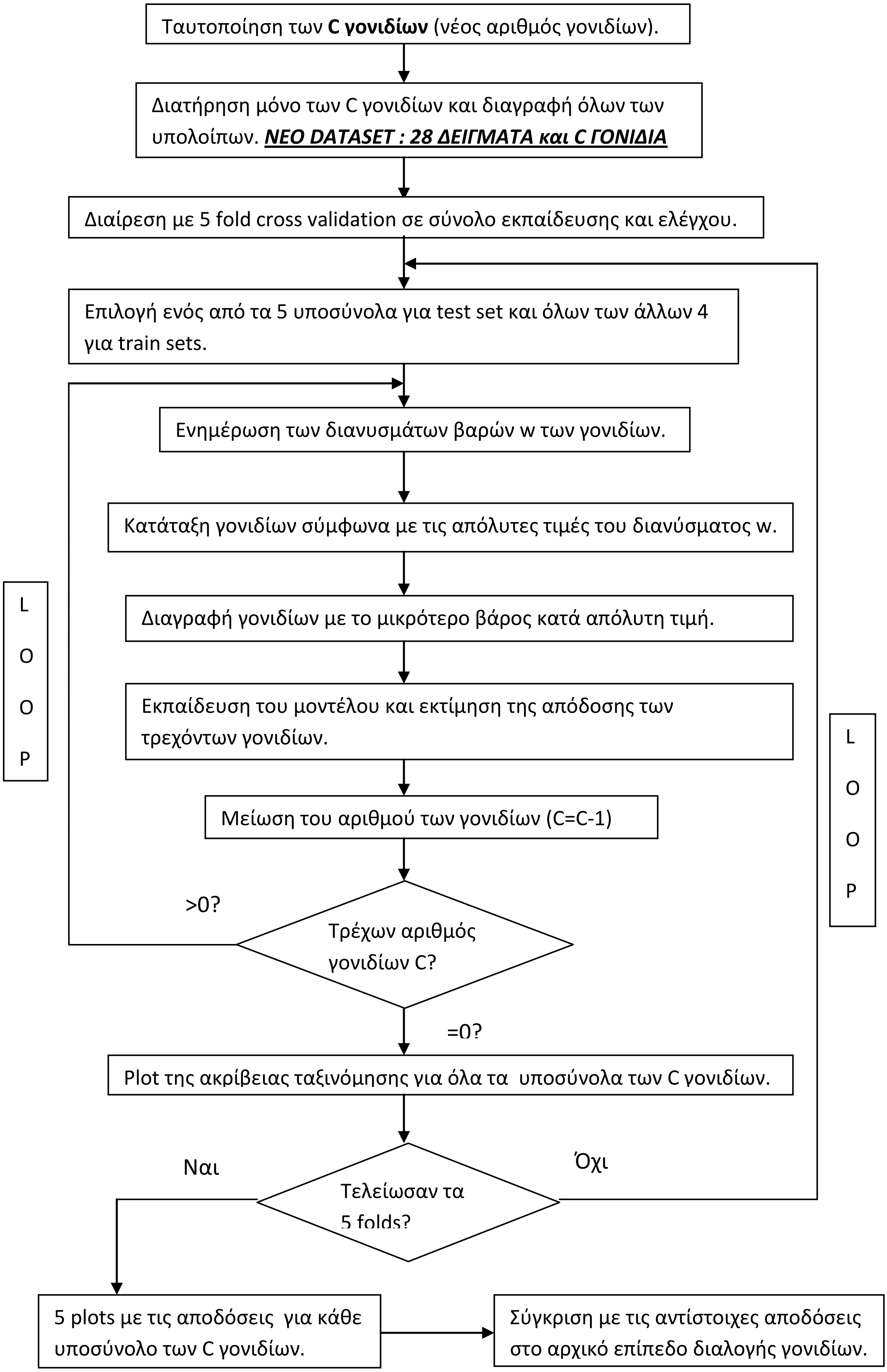


Εικόνα 37: Το πρώτο επίπεδο επιλογής γονιδίων της προτεινόμενης διαδικασίας επιλογής χαρακτηριστικών για την επιλογή γονιδίων.

4.5 Δεύτερο Επίπεδο Αξιολόγησης Δεδομένων

Για την πιο λεπτομερή εξέταση των δεδομένων για τα C επιλεγμένα γονίδια του πρώτου επιπέδου επιλογής γονιδίων της ενότητας 4.4, ακολουθούνται τα βήματα 1, 2 και 3 της διαδικασίας του πρώτου επιπέδου αξιολόγησης δεδομένων της ενότητας 4.3, ξεκινώντας όμως όχι από τα ολικά (2467) γονίδια αλλά από τα επιλεγμένα C γονίδια. Πιο συγκεκριμένα, εκτελείται ταυτοποίηση των C γονιδίων στο αρχικό σύνολο δεδομένων. Δηλαδή, από τα 2467 γονίδια που υπάρχουν στην αρχική μας βάση, διατηρούνται μόνο αυτά τα C, ενώ τα υπόλοιπα διαγράφονται. *Οπότε, δημιουργείται ένα νέο σύνολο δεδομένων με πολύ λιγότερα γονίδια από αυτά που είχε το αρχικό.*

Για το νέο σύνολο δεδομένων, εκτελούνται τα βήματα 1, 2 και 3 της διαδικασίας που περιγράφεται στην ενότητα 4.3 και παρουσιάζεται στην εικόνα 32, για τα C γονίδια, και με τη συνθήκη που εκτελείται σε κάθε fold, για τον αλγόριθμο RFE-LNW, να είναι «από C έως και 1 γονίδια», όπως υποδεικνύει το βήμα 2 του πίνακα 1 στην υποενότητα 3.4.6. Έτσι, το κάθε μοντέλο ταξινόμησης εκπαιδεύεται και έπειτα ελέγχεται η ικανότητα γενίκευσης του στο νέο τροποποιημένο σύνολο δεδομένων για όλους τους αριθμούς των υποσυνόλων των γονιδίων από C έως και 1. Τέλος, γίνεται σύγκριση της μέσης απόδοσής του στο δεύτερο επίπεδο αξιολόγησης δεδομένων σε σχέση με τη μέση απόδοσή του στο πρώτο επίπεδο αξιολόγησης δεδομένων. Αναλυτικά, η διαδικασία για την δεύτερη φάση της επιλογής γονιδίων, περιγράφεται στο διάγραμμα ροής της εικόνας 38.



Εικόνα 38: Το δεύτερο επίπεδο αξιολόγησης δεδομένων της διαδικασίας επιλογής χαρακτηριστικών για την επιλογή γονιδίων.

Αυτό που αλλάζει για κάθε μοντέλο, στην παραπάνω διαδικασία, είναι οι αριθμοί των σημαντικότερων C γονιδίων που επιλέγονται στο τέλος της διαδικασίας του πρώτου επιπέδου επιλογής γονιδίων της ενότητας 4.4, και που εισάγονται σε αυτό το δεύτερο αξιολόγησης δεδομένων με σκοπό την εκτίμηση της απόδοσης ταξινόμησης που αυτά αποφέρουν.

<u>Τα αποτελέσματα του δευτέρου επιπέδου αξιολόγησης δεδομένων, για τον πρώτο συνδυασμό ταξινομητών καθώς και για τους πέντε ατομικούς ταξινομητές που τον αποτελούν συνοψίζονται στον πίνακα 12.</u>

Αντλώντας πληροφορία από τον πίνακα 10 της ενότητας 4.4, ελέγχεται η μέση ακρίβεια ταξινόμησης για κάθε έναν από τους πέντε ατομικούς ταξινομητές, ξεχωριστά, για τα C γονίδια στα οποία έχει καταλήξει ο καθένας από αυτούς, και τα αποτελέσματα παρατίθενται στην πρώτη γραμμή του πίνακα 12. Στη δεύτερη γραμμή του, φαίνεται η μέση απόδοσή τους, καθώς η μέση απόδοση του συνδυασμού τους για τα γονίδια που έχουν προκύψει από αυτόν.

Πίνακας 12: <u>Η μέση ακρίβεια ταξινόμησης στο 2^o επίπεδο αξιολόγησης των C γονιδίων</u>, για τους πέντε ατομικούς ταξινομητές και για το μείγμα αυτών που βασίζεται σε συγχώνευση με συνδυασμό ασαφών βαρών.

MEAN CLASSIFICATION ACCURACY FOR THE 5 FOLDS	*MoE based on a fuzzy formulation*	*knc polynomial 2^{nd} degree*	*kmse gaussian*	*ks polynomial 3^{rd} degree*	*svm quadratic*	*svm linear*
FOR ALL THE SUBSETS OF C SELECTED GENES	-	65	94,52	92,87	95,72	95,88
FOR ALL THE SUBSETS OF C SELECTED GENES (MIXTURES')	71,84	60,02	91,8	96,61	86,1	97,74

- ✓ Παρατηρούμε ότι μετά το πρώτο επίπεδο επιλογής γονιδίων, κατά το οποίο ο αριθμός των γονιδίων μειώνεται αισθητά σε σχέση με το αρχικό επίπεδο, ο Gaussian kernel minimum squared error machine ταξινομητής έχει αυξήσει αισθητά τη μέση απόδοση ταξινόμησης. Αντίθετα, ο kernel nearest centroid με πολυωνυμικό πυρήνα δευτέρου βαθμού έχει μειώσει αισθητά την ακρίβεια ταξινόμησής του. Το svm με τετραγωνικό πυρήνα αύξησε λίγο την απόδοσή του, ενώ η ακρίβεια του kernel subspace με πολυωνυμικό πυρήνα τρίτου βαθμού και του γραμμικού svm φαίνεται να μένει ανεπηρέαστη από τη μείωση των γονιδίων, αφού παραμένει το ίδιο υψηλή. Η μέση απόδοση του συνδυασμού των ταξινομητών μειώνεται λίγο.

Ο έλεγχος για την απόδοση των C=27 γονιδίων που επιλέγονται στο πρώτο επίπεδο επιλογής γονιδίων, για το δεύτερο συνδυασμό ταξινομητών, καθώς και για τους δύο ειδικούς που τον αποτελούν, συνοψίζονται στον πίνακα 13.

Πίνακας 13: Η μέση ακρίβεια ταξινόμησης στο 2^o επίπεδο αξιολόγησης των C=27 γονιδίων, για τους δύο ειδικούς και για το μείγμα αυτών που βασίζεται στη συγχώνευση των ειδικών με ένα gradient descent learning αλγόριθμο .

MEAN CLASSIFICATION ACCURACY FOR THE 5 FOLDS	*MoE based on a gradient*	*1st expert*	*2nd expert*
FOR ALL THE SUBSETS OF C SELECTED GENES	73,46	73,40	73,31

- ✓ Παρατηρούμε ότι ο συνδυασμός των ειδικών που βασίζεται σε ένα gradient, αυξάνει κατά πολύ τη μέση απόδοσή του, όταν ο αριθμός γονιδίων μειώνεται δραστικά στα 27.

Ο έλεγχος για την απόδοση των C=26 γονιδίων που επιλέγονται στο πρώτο επίπεδο επιλογής γονιδίων, για τον τρίτο συνδυασμό ταξινομητών, καθώς και για τους δύο ειδικούς που τον αποτελούν, συνοψίζονται στον πίνακα 14.

Πίνακας 14: Η μέση ακρίβεια ταξινόμησης στο 2^o επίπεδο αξιολόγησης των C=26 γονιδίων, για τους δύο ειδικούς και για το μείγμα αυτών που βασίζεται σε ένα gradient descent learning αλγόριθμο και σε συνδυασμό ασαφών βαρών.

MEAN CLASSIFICATION ACCURACY FOR THE 5 FOLDS	*MoE based on a fuzzy formulation for 2 logsigs*	*1^{st} expert*	*2^{nd} expert*
FOR ALL THE SUBSETS OF C SELECTED GENES	92,15	93,97	92,90

- ✓ Παρατηρούμε ότι η μέση ακρίβεια του μείγματος είναι λίγο υψηλότερη από την αντίστοιχη στο αρχικό επίπεδο, ενώ των δύο ειδικών κυμαίνεται στα ίδια περίπου ποσοστά.

Ο έλεγχος για την απόδοση των C=35 γονιδίων που επιλέγονται στο πρώτο επίπεδο επιλογής γονιδίων, για τον τέταρτο συνδυασμό ταξινομητών καθώς και για τους δύο ειδικούς που τον αποτελούν, συνοψίζονται στον πίνακα 15.

Πίνακας 15: Η μέση ακρίβεια ταξινόμησης στο 2^o επίπεδο αξιολόγησης των C=35 γονιδίων, για τους δύο ειδικούς και για το μείγμα αυτών που βασίζεται στην επιλογή ταξινομητών με som διαχωρισμό του χώρου χαρακτηριστικών.

MEAN CLASSIFICATION ACCURACY FOR THE 5 FOLDS	*MoE based on som formulation*	*1^{st} expert*	*2^{nd} expert*
FOR ALL THE SUBSETS OF C SELECTED GENES	97,67	50,56	51,35

- ✓ Παρατηρούμε ότι ο συνδυασμός των ειδικών που βασίζεται σε μία som διαμόρφωση, αυξάνει κατά πολύ τη μέση απόδοσή του, όταν ο αριθμός γονιδίων μειώνεται δραστικά στα 35.

Σαν μία γενική παρατήρηση, μπορούμε να πούμε ότι η μέση ακρίβεια όλων των ταξινομητών αλλά και των συνδυασμών τους, εμφανίζεται αυξημένη μετά το πρώτο επίπεδο επιλογής γονιδίων όπου ο αριθμός των γονιδίων έχει υποστεί μία πρώτη μείωση σε σχέση με το αρχικό αριθμό τους, με μόνη εξαίρεση την kernel nearest centroid μέθοδο με πολυωνυμικό πυρήνα δευτέρου βαθμού. Αξιοσημείωτο είναι το γεγονός ότι αυτός ο ταξινομητής μείωσε αρκετά τη μέση απόδοσή του, όταν τα γονίδια μειώθηκαν στα 350, ενώ ο ταξινομητής kernel minimum squared error με Gaussian πυρήνα αύξησε κατά πολύ τη μέση απόδοσή του όταν ο αριθμός των γονιδίων μειώθηκε στα 23.

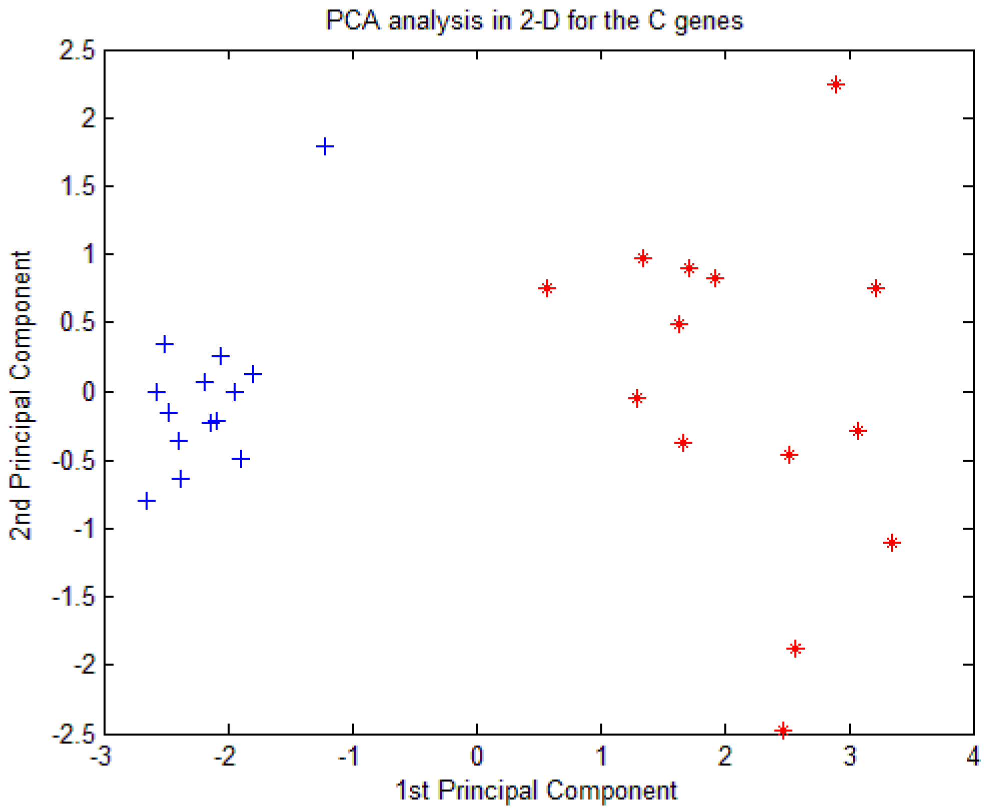
Σαν επιπλέον κριτήριο επιβεβαίωσης της ποιότητας των αποτελεσμάτων μαζί με τη μέση ακρίβεια ταξινόμησης μετράμε και το μέσο kernel target alignment και για τα 5 folds, για όλους τους συνδυασμούς ταξινομητή-πυρήνα, πάνω στο νέο σύνολο δεδομένων των C γονιδίων. Τα αποτελέσματα συνοψίζονται στον πίνακα 16.

- ✓ Παρατηρούμε ότι το kernel target alignment για το νέο σύνολο δεδομένων με τα 80 γονίδια, που σχηματίστηκε μετά το πρώτο επίπεδο επιλογής για πρώτο μείγμα ειδικών, έχει αυξηθεί αισθητά σε σχέση με το αρχικό επίπεδο, για όλους τους πυρήνες. Δηλαδή, όλοι οι πυρήνες με πρώτο πάλι τον γραμμικό, ευθυγραμμίζονται πολύ καλύτερα στο σύνολο δεδομένων με τα μειωμένα γονίδια από ότι στο αρχικό, που περιλαμβάνει πολλά μη σχετικά γονίδια.
- ✓ Η μέση ακρίβεια ταξινόμησης παρουσιάζει αλλαγές μόνο για τον kernel nearest centroid ταξινομητή και για τον kernel minimum squared error, σε σχέση με την αντίστοιχη στο αρχικό επίπεδο. Ο πρώτος μειώνει την απόδοσή του, για πυρήνα πολυωνυμικό δευτέρου και τρίτου βαθμού, ενώ ο δεύτερος αυξάνει την απόδοσή του για τον γκαουσιανό πυρήνα. Οι υπόλοιπες μέσες αποδόσεις δεν παρουσιάζουν σημαντικές αλλαγές.

Πίνακας 16: Μέσες τιμές για το kernel target alignment και το classification accuracy για όλους τους συνδυασμούς ταξινομητών και πυρήνων, στο 2ο επίπεδο αξιολόγησης των C=80 γονιδίων.

Mean of the 5 folds	*Gaussian RBF*	*Linear*	*Polynomial 2nd degree*	*Polynomial 3rd degree*	*Quadratic*	*Svm Polynomial 3rd degree*
Kernel Target Alignment	0,39	**0,83**	0,50	0,43	0,5	0,43
Accuracy for knc	47,69	56,61	58,52	**65,85**	-	-
Accuracy for kmse	91,98	91,81	93,55	**96**	-	-
Accuracy for ks	**96,88**	90	50	95	-	-
Accuracy for svm	-	**96,33**	-	-	87,41	87,11

Επίσης, εφαρμόζοντας την PCA ανάλυση στο νέο σύνολο δεδομένων με τα C=80 γονίδια που προέκυψαν για το πρώτο μείγμα ειδικών, κάνουμε μία οπτική επιβεβαίωση των αποτελεσμάτων για το νέο μειωμένο σύνολο δεδομένων. Παρατηρούμε ότι τα δείγματα της κλάσης elutriation βρίσκονται συγκεντρωμένα σε μία μικρότερη περιοχή, σε σχέση με εκείνα της κλάσης heat-cold-dtt, τα οποία βρίσκονται πιο συγκεντρωμένα στο χώρο, από ότι στο αρχικό επίπεδο με τα 2467 γονίδια, αλλά σε μία μεγαλύτερη περιοχή. Η διαχωρισιμότητα των δύο κλάσεων γίνεται σαφώς καλύτερη από το αρχικό επίπεδο.



Εικόνα 39: Αποτέλεσμα της PCA ανάλυσης στο νέο σύνολο δεδομένων με τα C=80 γονίδια. Με τους μπλε σταυρούς παρουσιάζεται η κλάση elutriation, ενώ με τα κόκκινα αστέρια η κλάση heat-cold-dtt.

4.6 Δεύτερο Επίπεδο Επιλογής Γονιδίων

Σε αυτή την ενότητα παρουσιάζεται το δεύτερο επίπεδο της επιλογής χαρακτηριστικών της μεθόδου επιλογής των γονιδίων, στο τέλος του οποίου εξάγεται ο αριθμός F των σημαντικότερων γονιδίων από όλα τα 2467 γονίδια του αρχικού συνόλου δεδομένων μας. Αυτά τα F γονιδία αποτελούν τη γονιδιακή υπογραφή για το κάθε μοντέλο ταξινόμησης και η απόδοση ταξινόμησης αυτών αξιολογείται στο τελικό επίπεδο αξιολόγησης των γονιδιακών υπογραφών της επόμενης ενότητας 4.7.

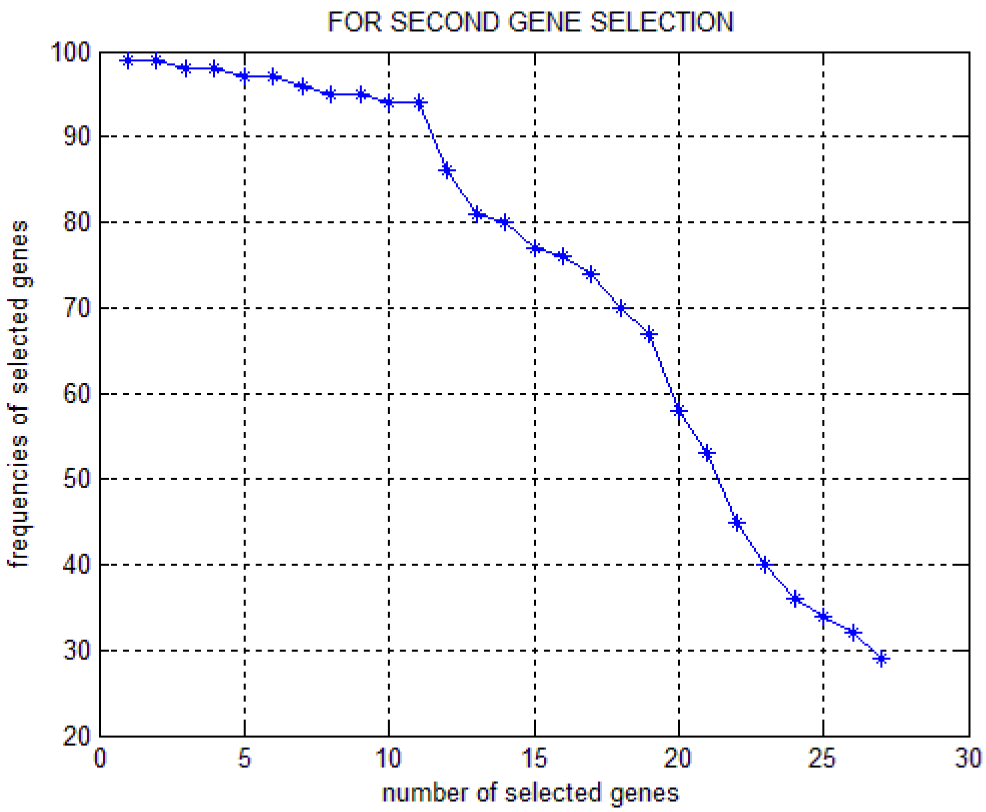
Στο δεύτερο επίπεδο επιλογής γονιδίων, αρχικά βασιζόμαστε στον αριθμό γονιδίων D του πίνακα 11 της ενότητας 4.4, που αποτελεί τον αριθμό των γονιδίων με τις υψηλότερες συχνότητες από όλα τα C γονίδια που υπάρχουν στο νέο (μειωμένο) σύνολο δεδομένων του δευτέρου επιπέδου αξιολόγησης δεδομένων που περιγράφεται στην ενότητα 4.5. Έπειτα, αυτή η φάση λαμβάνει υπόψη της τις συχνότητες εμφάνισης που τα D γονίδια αποκτούν, όταν εφαρμοστεί ο αλγόριθμος RFE-LNW πολλαπλές φορές πάνω στο νέο σύνολο δεδομένων, και επιλέγει τα F γονίδια με τις υψηλότερες συχνότητες σύμφωνα με ένα όριο τιμών.

<u>Δεύτερη επιλογή γονιδίων σύμφωνα με τη συχνότητα εμφάνισής τους.</u>

Για να γίνει πιο κατανοητή η διαδικασία που ακολουθείται στο δεύτερο επίπεδο επιλογής γονιδίων, ακολουθεί η αναλυτική περιγραφή της και η παρουσίαση των αποτελεσμάτων που προκύπτουν σε κάθε βήμα της, για το μοντέλο support vector machine με γραμμικό πυρήνα, ως συνέχεια των δύο φάσεων του πρώτου επιπέδου. Όμοια είναι η διαδικασία που έχει ακολουθηθεί και για τα υπόλοιπα μοντέλα μάθησης και πρόβλεψης που παρουσιάζονται στην ενότητα 4.2.

Εκτελούνται 20 επαναλήψεις των βημάτων 1 και 2, της διαδικασίας της ενότητας 4.3, με δύο διαφοροποιήσεις στο βήμα 2. Παραλείπεται από το βήμα 2 το σημείο της εκπαίδευσης και άρα και του ελέγχου του μοντέλου, αφού σε αυτή φάση δεν έχει ολοκληρωθεί ακόμα η επιλογή των γονιδίων.

Επιπλέον, η συνθήκη που εκτελείται σε κάθε fold, για τον αλγόριθμο RFE-LNW, δεν είναι πλέον «από 2467 έως και 1 γονίδια», όπως υποδεικνύει το βήμα 2 του πίνακα 1 στην υποενότητα 3.4.6, αλλά αλλάζει στη συνθήκη «από 27 έως και 20 γονίδια» (ή πιο γενικά «από C έως και D»). Εκτελούνται δηλαδή (5 folds x 20 επαναλή



Εικόνα 40: Τα 27 γονίδια με τις συχνότητες τους, όπως αυτές προκύπτουν από το δεύτερο επίπεδο επιλογής γονιδίων, για το γραμμικό support vector machine.

ψεις= 100) 100 runs του RFE-LNW, όχι για όλα τα γονίδια όπως στο αρχικό επίπεδο επεξεργασίας των γονιδίων, αλλά για 27-20=7 γονίδια κάθε φορά. Τα 7 γονίδια που διατηρούνται και δεν εξαλείφονται σε κάθε επανάληψη, ελέγχονται πριν εισαχθούν σε μία καθολική λίστα συχνοτήτων. Αν τα γονίδια υπάρχουν ήδη στη λίστα από προηγούμενη επανάληψη, δεν εισάγονται ξανά, αλλά αυξάνεται η συχνότητά τους κατά 1, ενώ αν είναι καινούρια προστίθενται σε αυτήν με συχνότητα 1.

Στο τέλος όλων αυτών των 100 επαναλήψεων, προκύπτουν 27 γονίδια (*έστω Ε γονίδια*) με τις συχνότητές τους, οι οποίες απεικονίζονται στην εικόνα 40. Σημειώνουμε εδώ ότι, η μέγιστη τιμή που μπορεί να πάρει η συχνότητα ενός οποιουδήποτε γονιδίου είναι 100. Και αυτό γιατί εκτελούνται συνολικά 100 runs και μπορεί να τύχει κάποιο γονίδιο να είναι και τις 100 φορές μέσα στα 7 που κρατιούνται σε κάθε επανάληψη. Από τη γραφική παράσταση των συχνοτήτων των γονιδίων είναι εύκολο να παρατηρηθεί ότι στα 11 γονίδια (*έστω F γονίδια*) δημιουργείται «γόνατο».

Οπότε, από τα 27 γονίδια διατηρούνται μόνο τα 11 με τις υψηλότερες συχνότητες. Αυτός ο αριθμός γονιδίων που εξάγεται τελικά από το δεύτερο επίπεδο επιλογής γονιδίων, σχηματίζει το δεύτερο και τελικό σύνολο δεδομένων μέσω του οποίου θα γίνει έλεγχος της απόδοσης του μοντέλου στο τελικό επίπεδο αξιολόγησης των γονιδιακών υπογραφών.

Αυτό που αλλάζει για κάθε μοντέλο είναι οι αριθμοί C και D των γονιδίων που αντλούνται από τους πίνακες 10 και 11 της ενότητας 4.4, οπότε και ο αριθμός F των γονιδίων που επιλέγονται ως τα σημαντικότερα γονίδια από όλα τα 2467, στο τέλος της διαδικασίας αυτού του δεύτερου επιπέδου επιλογής.

Ο αριθμός των γονιδίων F που επιλέγονται ως τα πιο σημαντικά μετά το δεύτερο επίπεδο επιλογής γονιδίων, και που θα αξιολογηθούν στο τελικό επίπεδο αξιολόγησης γονιδιακών υπογραφών της ενότητας 4.7, ως προς την απόδοση ταξινόμησης που αποφέρουν, φαίνεται στον πίνακα 17, συγκεντρωτικά για όλα τα μοντέλα ταξινόμησης.

Πίνακας 17: Ο αριθμός των γονιδίων F που επιλέγονται ως τα πιο σημαντικά μετά το 2^o επίπεδο επιλογής γονιδίων, συγκεντρωτικά για όλα τα μοντέλα ταξινόμησης.

All the classifiers and the mixtures	Number of genes selected after the second selection level (F)
Knc polynomial 2nd degree	44
Kmse gaussian	13
Ks polynomial 3rd degree	13
Svm quadratic	13
Svm linear	11
MoE based on a fuzzy formulation	9
MoE based on a gradient	13
MoE based on a fuzzy formulation for 2 logsigs	13
MoE based on a som formulation	13

4.7 Τελικό Επίπεδο Αξιολόγησης Γονιδιακών Υπογραφών

Στο τελικό επίπεδο αξιολόγησης των γονιδιακών υπογραφών, εκτελείται έλεγχος ως προς την απόδοση της ταξινόμησης που σημειώνουν τα μοντέλα, με τα γονίδια που έχουν επιλεγεί στο δεύτερο επίπεδο επιλογής γονιδίων. Το σύνολο δεδομένων αλλάζει πάλι αφού τα γονίδια έχουν μειωθεί περισσότερο, και στη συνέχεια εφαρμόζονται, στο νέο αυτό σύνολο δεδομένων, πολλαπλές επαναλήψεις μόνο του βήματος 1 της διαδικασίας που περιγράφεται στην ενότητα 4.3, για το διαχωρισμό του συνόλου δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου. Δηλαδή, στο τελικό επίπεδο αξιολόγησης των γονιδίων, έχοντας καταλήξει στον αριθμό των πιο σημαντικών F γονιδίων για κάθε μοντέλο ταξινόμησης, εκπαιδεύουμε και ελέγχουμε τα μοντέλα, πολλαπλές φορές, γι’ αυτά τα γονίδια, θέλοντας να συγκρίνουμε την απόδοση που επιτυγχάνουν σε σχέση με αυτή που σημειώνουν για τα αρχικά 2467 γονίδια.

Για το support vector machine με γραμμικό πυρήνα παρατηρούμε από τον πίνακα 17 ότι μετά το δεύτερο επίπεδο επιλογής γονιδίων καταλήγει σε F=11 σημαντικά γονίδια. Εκτελείται ταυτοποίηση των 11 γονιδίων στο σύνολο δεδομένων με τα C γονίδια του πρώτου επιπέδου επιλογής γονιδίων. Δηλαδή, από τα 27 γονίδια που υπάρχουν σε αυτό το σύνολο, διατηρούνται μόνο αυτά τα 11, ενώ τα υπόλοιπα διαγράφονται. *Οπότε, δημιουργείται νέο σύνολο δεδομένων με πολύ λιγότερα γονίδια από αυτά που είχε το αρχικό αλλά και αυτό που δημιουργήθηκε μετά το πρώτο επίπεδο επιλογής.*

Για το νέο σύνολο δεδομένων, εκτελείται 10 φορές μόνο το βήμα 1 της διαδικασίας που περιγράφεται στην ενότητα 4.3 για το διαχωρισμό του συνόλου δεδομένων με cross validation, για τα 11 γονίδια, καθώς δεν εφαρμόζουμε σε αυτά τον αλγόριθμο RFE-LNW, αφού δε θέλουμε να μετρήσουμε τη μέση ακρίβεια της ταξινόμησης για όλα τα υποσύνολα γονιδίων από 11 έως και 1, αλλά μόνο για το σύνολο των 11 συγκεκριμένων γονιδίων. Έτσι, το γραμμικό support vector machine εκπαιδεύεται και ελέγχεται η ικανότητα γενίκευσης του, μετά από (5folds x 10 επαναλήψεις=50) 50 επαναλήψεις, στο νέο τροποποιημένο σύνολο δεδομένων με τα 11 γονίδια. Τέλος, γίνεται σύγκριση της μέσης απόδοσης του support vector machine στο δεύτερο επίπεδο επιλογής γονιδίων σε σχέση με τη μέση απόδοσή του στο πρώτο επίπεδο αξιολόγησης δεδομένων.

Αναλυτικά, τα πέντε στάδια της διαδικασία επιλογής γονιδίων, έτσι όπως έχουν παρουσιαστεί μέχρι στιγμής στις ενότητες 4.3 έως και 4.7, περιγράφονται στο διάγραμμα ροής της εικόνας 42.

<u>Τα αποτελέσματα του τελικού επιπέδου αξιολόγησης γονιδιακών υπογραφών, για τον πρώτο συνδυασμό ταξινομητών καθώς και για τους πέντε ατομικούς ταξινομητές που τον αποτελούν συνοψίζονται στον πίνακα 18.</u>

Αντλώντας πληροφορία από τον πίνακα 17, ελέγχεται η μέση ακρίβεια ταξινόμησης για κάθε έναν από τους πέντε ατομικούς ταξινομητές, ξεχωριστά, για τα γονίδια στα οποία έχει καταλήξει ο καθένας από αυτούς, και τα αποτελέσματα παρατίθενται στην πρώτη γραμμή του πίνακα 18. Στη δεύτερη γραμμή του, φαίνεται η μέση απόδοσή τους, καθώς η μέση απόδοση του συνδυασμού τους, για τα γονίδια που έχουν προκύψει από αυτόν. Η τρίτη γραμμή περιέχει την ίδια την πληροφορία, αλλά χωρίς τον kernel nearest centroid ταξινομητή, ο οποίος αφαιρέθηκε σκοπίμως σε αυτή την περίπτωση από το μείγμα των ειδικών, επειδή η μέση απόδοσή του είναι χαμηλή.

- ✓ Παρατηρούμε ότι μετά το δεύτερο επίπεδο επιλογής γονιδίων κατά το οποίο ο αριθμός των γονιδίων μειώνεται περισσότερο σε σχέση με το πρώτο επίπεδο, η μέση ακρίβεια ταξινόμησης για τα 5 folds, για όλα τα μοντέλα, εκτός του kernel nearest centroid με πολυωνυμικό πυρήνα δευτέρου βαθμού, είναι 100%, τόσο όταν αυτή υπολογίζεται με βάση τα F γονίδια στα οποία έχει καταλήξει το καθένα από αυτά, όσο και για τα F γονίδια του συνδυασμού τους. Η μέση απόδοση του μείγματος και του kernel nearest centroid είναι αρκετά χαμηλότερη στα λίγα γονίδια από την αντίστοιχη στο αρχικό επίπεδο επεξεργασίας των γονιδίων, κάτι το οποίο μας οδηγεί στο συμπέρασμα ότι αυτός ο ταξινομητής δεν ταξινομεί καλά όταν ο αριθμός των γονιδίων είναι μικρός.

Πίνακας 18: Η μέση ακρίβεια ταξινόμησης στο τελικό επίπεδο αξιολόγησης των F γονιδίων, για τους πέντε ατομικούς ταξινομητές και για το μείγμα αυτών που βασίζεται σε συγχώνευση με συνδυασμό ασαφών βαρών.

MEAN CLASSIFICATION ACCURACY FOR THE 5 FOLDS	*MoE based on a fuzzy formulation*	*knc polynomial 2^{nd} degree*	*kmse gaussian*	*ks polynomial 3^{rd} degree*	*svm quadratic*	*svm linear*
FOR THE F SELECTED GENES	-	50	100	100	100	100
FOR THE F SELECTED GENES of the MIXTURES'	64	54	100	96,66	100	100
FOR THE F SELECTED GENES of the MIXTURES' (without knc)	100	-	100	96,66	100	100

Ο έλεγχος για την απόδοση των F=13 γονιδίων που επιλέγονται στο δεύτερο επίπεδο επιλογής γονιδίων, για το δεύτερο συνδυασμό ταξινομητών, καθώς και για τους δύο ειδικούς που τον αποτελούν, συνοψίζονται στον πίνακα 19.

Πίνακας 19: Η μέση ακρίβεια ταξινόμησης στο τελικό επίπεδο αξιολόγησης των F γονιδίων, για τους δύο ειδικούς και για το μείγμα αυτών που βασίζεται σε ένα gradient descent learning αλγόριθμο.

MEAN CLASSIFICATION ACCURACY FOR THE 5 FOLDS	*MoE based on a gradient*	*1^{st} expert*	*2^{nd} expert*
FOR THE F SELECTED GENES	85	85	85

- ✓ Παρατηρούμε ότι ο συνδυασμός των ειδικών που βασίζεται σε ένα gradient descent learning αλγόριθμο, αυξάνει κατά πολύ τη μέση απόδοσή του, όταν ο αριθμός γονιδίων μειώνεται δραστικά στα 13.

Ο έλεγχος για την απόδοση των F=13 γονιδίων που επιλέγονται στο δεύτερο επίπεδο επιλογής γονιδίων, για τον τρίτο συνδυασμό ταξινομητών, καθώς και για τους δύο ειδικούς που τον αποτελούν, συνοψίζονται στον πίνακα 20.

Πίνακας 20: Η μέση ακρίβεια ταξινόμησης στο τελικό επίπεδο αξιολόγησης των F γονιδίων, για τους δύο ειδικούς και για το μείγμα αυτών που βασίζεται σε ένα gradient descent learning αλγόριθμο και σε συνδυασμό ασαφών βαρών.

MEAN CLASSIFICATION ACCURACY FOR THE 5 FOLDS	*MoE based on a fuzzy formulation for 2 logsigs*	*1^{st} expert*	*2^{nd} expert*
FOR THE F SELECTED GENES	100	100	100

- ✓ Παρατηρούμε ότι η μέση ακρίβεια του μείγματος είναι λίγο υψηλότερη από την αντίστοιχη στο αρχικό επίπεδο, ενώ των δύο ειδικών κυμαίνεται στα ίδια περίπου ποσοστά.

Ο έλεγχος για την απόδοση των F=13 γονιδίων που επιλέγονται στο δεύτερο επίπεδο επιλογής γονιδίων, για τον τέταρτο συνδυασμό ταξινομητών καθώς και για τους δύο ειδικούς που τον αποτελούν, συνοψίζονται στον πίνακα 21.

- ✓ Παρατηρούμε ότι ο συνδυασμός των ειδικών που βασίζεται σε μία som διαμόρφωση, αυξάνει τη μέση απόδοσή του στο 100%, όταν ο αριθμός γονιδίων μειώνεται δραστικά στα 13.

Πίνακας 21: <u>Η μέση ακρίβεια ταξινόμησης στο τελικό επίπεδο αξιολόγησης των F γονιδίων</u>, για τους δύο ειδικούς και για το μείγμα αυτών που βασίζεται στην επιλογή ταξινομητών με som διαχωρισμό του χώρου χαρακτηριστικών.

MEAN CLASSIFICATION ACCURACY FOR THE 5 FOLDS	*MoE based on som formulation*	*1st expert*	*2nd expert*
FOR THE F SELECTED GENES	100	50	50

Σαν μία γενική παρατήρηση, μπορούμε να πούμε ότι η μέση ακρίβεια όλων των ταξινομητών αλλά και των συνδυασμών τους εμφανίζεται αυξημένη μετά και από το δεύτερο επίπεδο επιλογής γονιδίων όπου ο αριθμός των γονιδίων έχει υποστεί μία σημαντική μείωση σε σχέση με το αρχικό αριθμό τους. Εξαίρεση σε αυτό το συμπέρασμα αποτελεί η kernel nearest centroid μέθοδο με πολυωνυμικό πυρήνα δευτέρου βαθμού, που μειώνει αρκετά τη μέση απόδοσή της για μικρό αριθμό γονιδίων, και λαμπρή επιβεβαίωση η μέθοδος kernel minimum squared error με γκαουσιανό πυρήνα, που αυξάνει πολύ τη μέση ακρίβεια ταξινόμησης για το σύνολο των F επίλεκτων γονιδίων.

4.7.1 Ποιοτικά μέτρα των αποτελεσμάτων

Μετράμε τη μέση ακρίβεια ταξινόμησης καθώς και το μέσο kernel target alignment και για τα 5 folds, για όλους τους συνδυασμούς ταξινομητή-πυρήνα. Τα αποτελέσματα συνοψίζονται στον παρακάτω πίνακα (έχει τονιστεί με έντονα γράμματα η μεγαλύτερη τιμή ανά γραμμή):

Συγκρίνοντας τις τιμές του πίνακα 22 με τις αντίστοιχες του πίνακα 3, καταλήγουμε στα εξής συμπεράσματα:

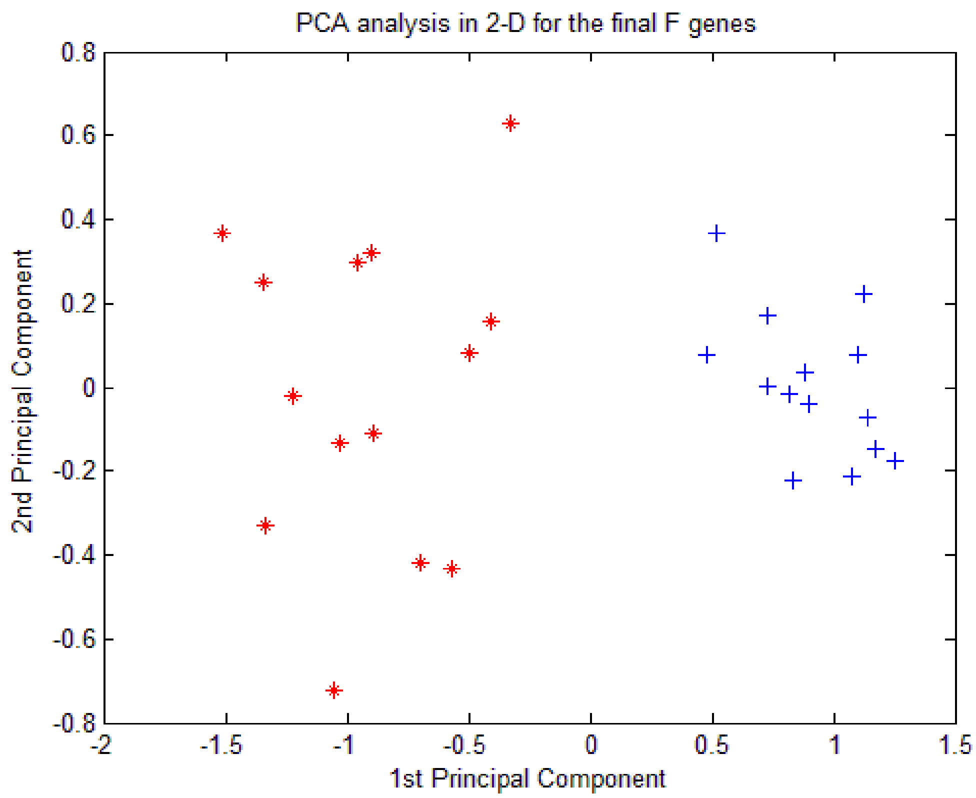
- ✓ Παρατηρούμε ότι το kernel target alignment για το νέο σύνολο δεδομένων με τα 9 γονίδια, που σχηματίστηκε μετά το δεύτερο επίπεδο επιλογής γονιδίων για πρώτο μείγμα ειδικών, έχει αυξηθεί αισθητά σε σχέση με το αρχικό επίπεδο, για όλους τους πυρήνες. Δηλαδή, όλοι οι πυρήνες με πρώτο πάλι τον γραμμικό, ευθυγραμμίζονται πολύ καλύτερα στο σύνολο δεδομένων με τα μειωμένα γονίδια από ότι στο αρχικό, που περιλαμβάνει πολλά μη σχετικά γονίδια.

Πίνακας 22: Μέσες τιμές για το kernel target alignment και το classification accuracy για όλους τους συνδυασμούς ταξινομητών και πυρήνων, στο τελικό επίπεδο αξιολόγησης *των F=9 γονιδίων.*

Mean of the 5 folds	*Gaussian RBF*	*Linear*	*Polynomial 2^{nd} degree*	*Polynomial 3^{rd} degree*	*Quadratic*	*Svm Polynomial 3^{rd} degree*
Kernel Target Alignment	0,23	**0,72**	0,35	0,41	0,62	0,54
Accuracy for knc	58,52	57,41	58,52	**58,89**	-	-
Accuracy for kmse	83,52	**90,56**	82,96	83,71	-	-
Accuracy for ks	**87,41**	60,93	50	82,78	-	-
Accuracy for svm	-	**93,15**	-	-	88,89	89,63

- ✓ Η μέση ακρίβεια ταξινόμησης παρουσιάζει αλλαγές για τον kernel nearest centroid ταξινομητή, αφού εμφανίζεται μειωμένη για όλους τους πυρήνες, κάτι το οποίο υποδεικνύει ότι αυτό το μοντέλο δεν ταξινομεί με ακρίβεια όταν ο αριθμός των γονιδίων είναι πολύ μικρός. Η μέση ακρίβεια ταξινόμησης παρουσιάζεται αυξημένη για τον kernel minimum squared error με Gaussian πυρήνα σε σχέση με την αντίστοιχη στο αρχικό επίπεδο επεξεργασίας, και μειωμένη για τον kernel subspace με γραμμικό πυρήνα. Οι υπόλοιπες μέσες αποδόσεις δεν παρουσιάζουν σημαντικές αλλαγές.

Θέλοντας να κάνουμε οπτική επιβεβαίωση των αποτελεσμάτων μας, εφαρμόζουμε την PCA ανάλυση στο νέο σύνολο δεδομένων με τα 9 γονίδια που προέκυψαν για το πρώτο μείγμα ειδικών. Παρατηρούμε από την εικόνα 41, ότι τόσο τα δείγματα της κλάσης elutriation, όσο και εκείνα της κλάσης heat-cold-dtt βρίσκονται συγκεντρωμένα μεταξύ τους στο χώρο σε σχέση με τη θέση που είχαν στο αρχικό επίπεδο με τα 2467 γονίδια. Έτσι, η διαχωρισιμότητα των δύο κλάσεων βελτιώνεται σημαντικά μετά από την εφαρμογή της διαδικασίας επιλογής των γονιδίων.



Εικόνα 41: Αποτέλεσμα της PCA ανάλυσης στο νέο σύνολο δεδομένων με τα F=9 γονίδια. Με τους μπλε σταυρούς παρουσιάζεται η κλάση elutriation, ενώ με τα κόκκινα αστέρια η κλάση heat-cold-dtt.

ΠΡΩΤΟ ΕΠΙΠΕΔΟ ΑΞΙΟΛΟΓΗΣΗΣ ΔΕΔΟΜΕΝΩΝ:

1) ***Αρχική Dataset_1 : 28 δείγματα και 2467 γονίδια***. Διαίρεση με 5 fold cross validation σε σύνολο εκπαίδευσης και ελέγχου.
2) Εφαρμογή RFE-LNW **από 2467 μέχρι 1 γονίδια,** για την εκπαίδευση και τον έλεγχο απόδοσης του μοντέλου.
3) 5 plots, που αναπαριστούν την ακρίβεια ταξινόμησης, για κάθε υποσύνολο των **2467 γονιδίων**.
4) Από κάθε ένα από τα 5 plots που προκύπτουν, επιλογή του **ελαχίστου αριθμού γονιδίων** όπου παρατηρείται η **μέγιστη απόδοση ταξινόμησης**.
5) Εύρεση του μέσου όρου των 5 αριθμών. Έστω αριθμός **Α γονιδίων** που θα χρησιμοποιηθεί ως κριτήριο τερματισμού στο πρώτο επίπεδο επιλογής των γονιδίων. **(A<<2467)**

⇩

ΠΡΩΤΟ ΕΠΙΠΕΔΟ ΕΠΙΛΟΓΗΣ ΓΟΝΙΔΙΩΝ :

1) Εφαρμογή πολλαπλών επαναλήψεων (5 folds x 20 times = 100 runs) στα βήματα 1 και 2 του αρχικού επιπέδου, αλλά με νέα συνθήκη, **από 2467 μέχρι Α γονίδια**, για το βήμα 2.
2) Σε κάθε επανάληψη σώζονται τα Α γονίδια, που έχουν επιζήσει, σε μία λίστα. Αν κάποιο γονίδιο υπάρχει ήδη στη λίστα, απλά αυξάνεται η συχνότητα του κατά 1. (συχνότητα σύμφωνα με τον αριθμό εμφανίσεων).
3) Στο τέλος όλων των επαναλήψεων, πλοτάρονται τα **Β** γονίδια από τη λίστα με τις συχνότητές τους. **(B>A, B<2467)**
4) Εύρεση του γόνατου στο plot. Από τα Β γονίδια διατηρούνται μόνο τα **C γονίδια** με τις υψηλότερες συχνότητες για να αξιολογηθούν στο δεύτερο επίπεδο αξιολόγησης δεδομένων. **(C<B, C<<2467)**
5) Plot των **C** γονιδίων με τις συχνότητές τους.
6) Εύρεση του γόνατου στο plot. Από τα C γονίδια διατηρούνται μόνο τα **D γονίδια** με τις υψηλότερες συχνότητες για να χρησιμοποιηθεί ως κριτήριο τερματισμού στο δεύτερου επίπεδο επιλογής γονιδίων. **(D<C)**

⇩

<u>ΔΕΥΤΕΡΟ ΕΠΙΠΕΔΟ ΑΞΙΟΛΟΓΗΣΗΣ ΔΕΔΟΜΕΝΩΝ :</u>

1) Ταυτοποίηση αυτών των C γονιδίων στη Dataset_1. Διατήρηση μόνο αυτών των C γονιδίων και αφαίρεση όλων των άλλων.
2) ***<u>Νέα Dataset 2 : 28 δείγματα και C γονίδια. (C<<2467).</u>*** Διαίρεση με 5 fold cross validation σε σύνολο εκπαίδευσης και ελέγχου.
3) Εφαρμογή RFE-LNW **από C μέχρι 0 γονίδια,** για την εκπαίδευση και τον έλεγχο απόδοσης του μοντέλου.
4) 5 plots, που αναπαριστούν την ακρίβεια ταξινόμησης, για κάθε υποσύνολο των **C γονιδίων**.
5) Σύγκριση της μέσης απόδοσης ταξινόμησης με την αντίστοιχη μέση ακρίβεια από τα αρχικό επίπεδο.

<u>ΔΕΥΤΕΡΟ ΕΠΙΠΕΔΟ ΕΠΙΛΟΓΗΣ ΓΟΝΙΔΙΩΝ:</u>

1) Εφαρμογή πολλαπλών επαναλήψεων (5 folds x 20 times = 100 runs) στα βήματα 1 και 2 του αρχικού επιπέδου, αλλά με νέα συνθήκη, **από C μέχρι D γονίδια**, για το βήμα 2.
2) Σε κάθε επανάληψη σώζονται τα **D** γονίδια, που έχουν επιζήσει, σε μία λίστα. Αν κάποιο γονίδιο υπάρχει ήδη στη λίστα, απλά αυξάνεται η συχνότητα του κατά 1. (συχνότητα σύμφωνα με τον αριθμό εμφανίσεων).
3) Στο τέλος όλων των επαναλήψεων, πλοτάρονται τα **E** γονίδια από τη λίστα με τις συχνότητές τους. **(E>D ,E< C)**
4) Από τα E γονίδια γίνεται επιλογή αυτών με συχνότητα μεγαλύτερη από 90. Έστω **F genes.(F<<2467)**

<u>ΤΕΛΙΚΟ ΕΠΙΠΕΔΟ ΑΞΙΟΛΟΓΗΣΗΣ ΓΟΝΙΔΙΑΚΩΝ ΥΠΟΓΡΑΦΩΝ :</u>

1) Ταυτοποίηση αυτών των F γονιδίων στη Dataset_2. Διατήρηση μόνο αυτών των F γονιδίων και αφαίρεση όλων των άλλων.
2) ***<u>Νέα Dataset 3 : 28 δείγματα και F γονίδια. (F<<C , F<<<2467).</u>*** Διαίρεση με 5 fold cross validation σε σύνολο εκπαίδευσης και ελέγχου.
3) Εφαρμογή πολλαπλών επαναλήψεων (5 folds x 10 times = 50 runs) για την εκπαίδευση και τον έλεγχο απόδοσης του μοντέλου, για τα **F γονίδια**.
4) Σύγκριση της μέσης απόδοσης ταξινόμησης με την αντίστοιχη μέση ακρίβεια από τα αρχικό επίπεδο
5) **Ορισμός των F γονιδίων ως gene markers (τα σημαντικότερα από όλα τα 2467 αρχικά γονίδια).**

Εικόνα 42: Επανάληψη της εικόνας 30 με λεπτομέρειες. Τα 5 βήματα της προτεινόμενης μεθοδολογίας της επιλογής χαρακτηριστικών για την επιλογή γονιδίων.

ΚΕΦΑΛΑΙΟ 5: Αποτελέσματα και Αξιολόγηση

Σε αυτό το κεφάλαιο παρατίθενται όλα τα στατιστικά και βιολογικά αποτελέσματα που αποφέρει η προτεινόμενη μεθοδολογία του κεφαλαίου 4. Πραγματοποιείται αξιολόγηση, σύγκριση και ερμηνεία των αποτελεσμάτων της εργασίας, στην προσπάθειά μας να εκτιμήσουμε βιολογικά την επιρροή των αλγοριθμικών παραμέτρων στη διαδικασία επιλογής βιοδεικτών στο σύνολο δεδομένων του γονιδιώματος του *Saccharomyces cerevisiae*.

Ο στόχος της διαδικασίας επιλογής γονιδίων [3] είναι η εξαγωγή ενός μικρού συνόλου από πληροφοριακά γονίδια τα οποία εκφράζονται διαφορετικά στις δύο κλάσεις ενδιαφέροντος. Πληροφοριακά γονίδια θεωρούνται εκείνα που τα επίπεδα έκφρασής τους είναι χρήσιμα για την εκπαίδευση ενός μοντέλου το οποίο θα μπορεί να γενικεύσει, δηλαδή να προβλέψει σωστά την κλάση νέων δειγμάτων. Γενικά, σε αυτή την εργασία χρησιμοποιήσαμε κατάλληλα στατιστικά, όπως την αναδρομική εξάλειψη χαρακτηριστικών, για να παράγουμε μία λίστα από ταξινομημένα γονίδια, η οποία τοποθετεί, από τα περισσότερο διαφορετικά εκφρασμένα γονίδια, εκείνα που εμφανίζονται συχνότερα, στις κορυφαίες θέσεις της λίστας. Διαδοχικά επίπεδα έκφρασης των πιο σημαντικών γονιδίων χρησιμοποιούνται για την εκπαίδευση και τον έλεγχο των μοντέλων χρησιμοποιώντας τα δείγματα των 2 κλάσεων ενδιαφέροντος.

Αρχικά, η ακρίβεια πρόβλεψης ταξινόμησης του μοντέλου για τα 2467 γονίδια, χρησιμοποιείται ως μέτρο της αξίας πρόγνωσης των γονιδίων, δηλαδή επιλέγεται ένας αρχικός αριθμός γονιδίων με τον οποίο ξεκινά η διαδικασία της επιλογής χαρακτηριστικών, σύμφωνα με την ικανότητα ταξινόμησης που αυτά επιδεικνύουν. Στόχος είναι η εκπαίδευση ταξινομητών οι οποίοι είναι ικανοί να γενικεύουν, δηλαδή να προβλέπουν την κλάση των νέων δειγμάτων που δεν ανήκουν στο σύνολο δεδομένων εκπαίδευσης, χρησιμοποιώντας τα επίπεδα έκφρασης λίγων μόνο γονιδίων, σχετικών με το εκάστοτε θέμα της ταξινόμησης. Οπότε, έχει νόημα η χρήση της ακρίβειας πρόβλεψης ως μέτρο της σχετικότητας ενός γονιδίου.

Στη συνέχεια, η προτεινόμενη μεθοδολογία μας παρέχει μία αξιόπιστη μέθοδο για την εκχώρηση στατιστικής σημαντικότητας στα γονίδια, και την ανάδειξη εκείνων μόνο, που αποδεικνύονται τα πιο σχετικά από το αρχικό σύνολο γονιδίων. Η σχετικότητα ενός γονιδίου διαμορφώνεται από το «πόσο διαφορετικά» εκφράζεται στις δύο κλάσεις ενδιαφέροντος και παράλληλα από το «πόσο συχνά» εμφανίζεται σε σχέση με τα υπόλοιπα γονίδια. Έτσι, αν μπορεί [16] να προσδιοριστεί με μεγάλη βεβαιότητα ένας αριθμός γονιδίων με πολύ διαφορετικές εκφράσεις στις δύο κλάσεις της ταξινόμησης, τότε με μεγάλη πιθανότητα αυτό το σύνολο γονιδίων θα εκφέρει «καλή» προγνωστική δύναμη.

Μετά το τέλος του δεύτερου επιπέδου επιλογής γονιδίων, προκύπτουν 9 διαφορετικές υπογραφές, μία για κάθε μοντέλο που εξετάστηκε. Αυτές οι υπογραφές συγκρίνονται και εξετάζονται οι διαφορές τους σε στατιστικό επίπεδο, με βάση την πιθανή επικάλυψή τους, και επίσης στο επίπεδο της βιολογικής περιγραφής των επιλεγμένων γονιδίων.

5.1 Στατιστικά Αποτελέσματα

Στους πίνακες 23 και 24 παρατίθενται οι κωδικοί των γονιδίων από τα οποία αποτελούνται αυτές οι γονιδιακές υπογραφές. Αναλυτική ονοματολογία γι’ αυτά τα γονίδια, δηλαδή ονόματα και σύμβολα, παρατίθεται στο παράρτημα Β.

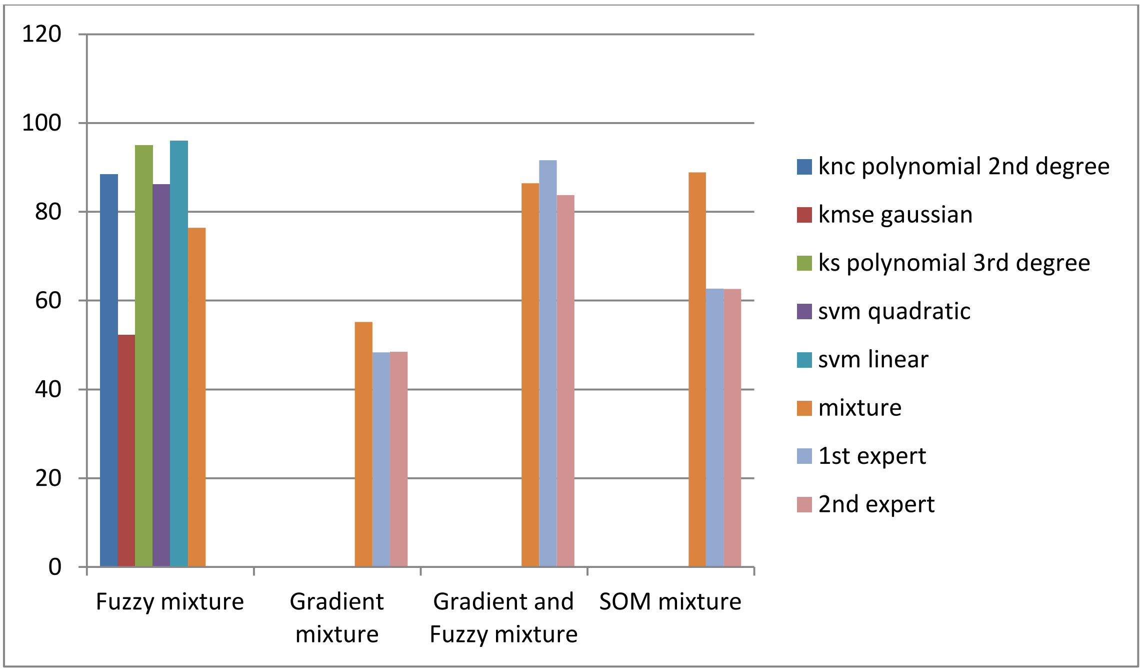
Πίνακας 23: Η γονιδιακή υπογραφή των 5 ατομικών ταξινομητών και του συνδυασμού τους.

SVM LINEAR 11 genes	***SVM QUADRATIC 13 genes***	***KS POLYNOMIAL 3^{rd} 13 genes***	***KMSE GAUSSIAN 13 genes***	***KNC POLYNOMIAL 2^{nd} 44 genes***		***MoE BASED ON A FUZZY SET FORMULATION 9 genes***
104	104	7	7	7	1092	30
475	475	30	30	104	1126	120
1066	934	104	104	136	1138	475
1213	1066	475	475	144	1213	1066
1262	1213	1066	1066	231	1238	1213
1283	1262	1213	1213	235	1250	1262
1307	1283	1262	1262	307	1260	1307
1723	1307	1283	1283	327	1266	2116
2116	1723	1307	1307	396	1414	2466
2364	1819	1723	1723	475	1819	
2466	2116	2116	2116	556	1916	
	2364	2364	2364	576	1992	
	2466	2466	2466	584	2062	
				616	2101	
				624	2116	
				644	2167	
				833	2231	
				934	2338	
				938	2359	
				952	2364	
				1065	2427	
				1074	2466	

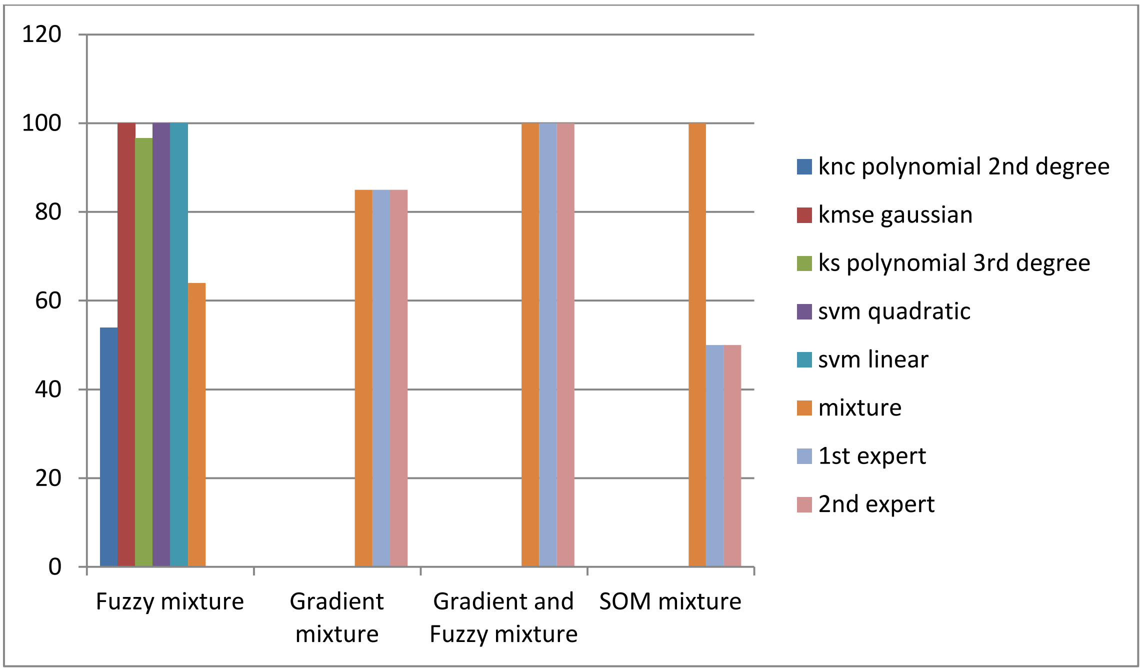
Πίνακας 24: Οι γονιδιακές υπογραφές των 4 διαφορετικών μειγμάτων ειδικών.

MoE BASED ON A FUZZY SET FORMULATION 9 genes	***MoE BASED ON SOM FORMULATION 13 genes***	***MoE BASED ON A GRADIENT 13 genes***	***MoE BASED ON 2 LOGSIGS 13 genes***
30	30	7	104
120	104	30	475
475	136	104	625
1066	475	475	934
1213	1066	1066	1066
1262	1213	1213	1213
1307	1262	1262	1262
2116	1283	1283	1283
2466	1307	1307	1307
	1723	1723	1723
	2116	2116	2116
	2364	2364	2364
	2466	2466	2466

Από τους πίνακες 4 έως 7 της ενότητας 4.3 και από τους 18 έως 21 της 4.7, που παρουσιάζουν τη μέση ακρίβεια ταξινόμησης για κάθε μοντέλο στο πρώτο επίπεδο και στο τελικό επίπεδο αξιολόγησης γονιδίων αντίστοιχα, προκύπτουν τα επόμενα δύο ραβδογράμματα εκατοστιαίας βάσης των εικόνων 43 και 44.



Εικόνα 43: Η μέση ακρίβεια όλων των ταξινομητών και μειγμάτων αυτών, για τα 2467 γονίδια.



Εικόνα 44: Η μέση ακρίβεια όλων των ταξινομητών και μειγμάτων αυτών, για τα F επίλεκτα γονίδια.

Συγκρίνοντας τα αποτελέσματα, καταλήγουμε στις εξής παρατηρήσεις:

- ✓ Το κάθε μοντέλο καταλήγει σε διαφορετικό αριθμό γονιδίων, όχι όμως και σε εντελώς διαφορετικές γονιδιακές υπογραφές μεταξύ τους, αφού πέρα από τις διαφορές στα επιλεγμένα γονίδια, υπάρχουν και κάποια που είναι κοινά.

- ✓ Η ακρίβεια ταξινόμησης που επιτυγχάνεται από τα μοντέλα, μέσω των γονιδίων που αποτελούν τις υπογραφές, είναι σημαντικά καλύτερη από εκείνη που σημειώνεται μέσω των γονιδίων του αρχικού συνόλου με τα 2467 γονίδια.
- ✓ Το κάθε μοντέλο καταλήγει σε μία γονιδιακή υπογραφή, μέσω της οποίας επιτυγχάνει διαφορετικά ποσοστά επιτυχίας ταξινόμησης, από τα υπόλοιπα μοντέλα.
- ✓ Το μοντέλο μείγματος ειδικών μπορεί να επιτύχει υψηλότερη ακρίβεια ταξινόμησης, στην περίπτωση που για τον συνδυασμό εφαρμόζεται η μέθοδος της επιλογής ταξινομητών, όπως είναι το τέταρτο μείγμα που βασίζεται στην εκπαίδευση των ειδικών σε διαφορετικά υποσύνολα του χώρου των χαρακτηριστικών που διαχωρίζονται με som.

Τα ποσοστά των κοινών γονιδίων ανάμεσα στους 5 ατομικούς ταξινομητές και στο μείγμα αυτών, καθώς και ανάμεσα στα 4 μείγματα, έτσι όπως αυτά προκύπτουν από τους πίνακες 23 και 24, φαίνονται στους πίνακες 25 και 26.

Πίνακας 25: Τα ποσοστά ομοιότητας των 5 ατομικών ταξινομητών και του συνδυασμού τους.

Ποσοστά % ομοιότητας	***SVM LINEAR***	***SVM QUADRATIC***	***KS POLYNOMIAL 3rd***	***KMSE GAUSSIAN***	***KNC POLYNOMIAL 2nd***
SVM LINEAR		100%	100%	100%	54%
SVM QUADRATIC			84,62%	84,62%	61,54%
KS POLYNOMIAL 3rd				100%	53,85%
KMSE GAUSSIAN					53,85%
MoE BASED ON A FUZZY SET FORMULATION	77%	77%	88,88%	88,88%	44,44%

Πίνακας 26: Τα ποσοστά ομοιότητας των 4 διαφορετικών συνδυασμών ειδικών.

Ποσοστά % ομοιότητας	***MoE BASED ON A FUZZY SET FORMULATION***	***MoE BASED ON SOM FORMULATION***	***MoE BASED ON A GRADIENT***	***MoE BASED ON 2 LOGSIGS***
MoE BASED ON A FUZZY SET FORMULATION		88,88%	88,88%	77,77%
MoE BASED ON SOM FORMULATION			92,31%	84,62%
MoE BASED ON A GRADIENT				84,62

Το ποσοστό ομοιότητας ανάμεσα στα σύνολα γονιδίων δύο μοντέλων έχει υπολογιστεί με βάση πάντα το μικρότερο σύνολο γονιδίων. Δηλαδή, αν συγκρίνοντας δύο διαφορετικά σε αριθμό σύνολα, διαπιστώσουμε ότι όλα τα στοιχεία του μικρότερου συνόλου συμπεριλαμβάνονται στο μεγαλύτερο σύνολο, τότε θεωρούμε ότι το ποσοστό ομοιότητάς τους είναι 100%.

Παρατηρώντας τους πίνακες 25 και 26, διαπιστώνουμε ότι τα ποσοστά ομοιότητας των γονιδίων είναι υψηλά σε πολλές από τις περιπτώσεις. Ωστόσο, υπάρχουν και συνδυασμοί με μεγάλες αποκλίσεις. Οι υπογραφές των ατομικών ταξινομητών εμφανίζουν αρκετά κοινά γονίδια τόσο μεταξύ τους, όσο και με το μείγμα τους. Ο ταξινομητής που διαφοροποιείται περισσότερο είναι ο kernel nearest centroid με πολυωνυμικό πυρήνα δευτέρου βαθμού, κάτι το οποίο είναι λογικό αν σκεφτούμε ότι η γονιδιακή του υπογραφή αποτελείται από πολλά περισσότερα γονίδια από ότι των υπολοίπων μοντέλων. Επίσης, τα ποσοστά ομοιότητας των υπογραφών των διαφορετικών μειγμάτων είναι υψηλά. Αυτό δείχνει ότι έχουμε μία σχετικά καθαρή εικόνα για το σύνολο των γονιδίων που είναι τα σημαντικότερα από όλα τα 2467 του αρχικού συνόλου δεδομένων.

5.2 Βιολογικά Αποτελέσματα

Γονίδια που μπορεί να εμφανίζονται διαφορετικά από τη στατιστική τους αξιολόγηση, μπορεί να παρουσιάζουν κάποια βιολογική σχέση. Δηλαδή, μπορεί δύο ή περισσότερες γονιδιακές υπογραφές [51] που εμφανίζουν μικρή επικάλυψη μεταξύ τους, να μοιράζονται σημαντική βιολογική επικάλυψη σε συναφείς περιγραφές ή βιολογικές διεργασίες.

Εργαζόμαστε προς αυτή την κατεύθυνση, χρησιμοποιώντας το σύστημα ταξινόμησης FunCat [52, 53], το οποίο είναι ένα ιεραρχικά δομημένο, ανεξάρτητο από οργανισμό, ευέλικτο και προεκτάσιμο ελεγχόμενο σύστημα, που επιτρέπει τη λειτουργική περιγραφή των πρωτεϊνών από κάθε οργανισμό, όπως προκαρυωτικά γονιδιώματα, ευκαρυωτικά μονοκύτταρα γονιδιώματα, φυτά και ζώα. Το FunCat, λαμβάνοντας υπόψη το ευρύ και εξαιρετικά ποικίλο φάσμα των γνωστών πρωτεϊνικών συναρτήσεων, αποτελείται από 28 κύριες λειτουργικές κατηγορίες που καλύπτουν γενικά πεδία όπως κυτταρική μεταφορά, μεταβολισμό και κυτταρική επικοινωνία/μετάδοση σήματος.

Παρουσιάζει μία ιεραρχική δομή, που μοιάζει με δέντρο, με τη χρήση έως και έξι επιπέδων αυξανόμενης εξειδίκευσης και γενικά, η έκδοση 2, που χρησιμοποιήθηκε σε αυτή την εργασία, περιλαμβάνει 1307 λειτουργικές κατηγορίες, για τα ευκαρυωτικά γονιδιώματα του Saccharomyces cerevisiae. Οι περιγραφές του FunCat χρησιμοποιήθηκαν για να γίνει σάρωση των σημαντικά εμπλουτισμένων κατηγοριών μέσα στο σύνολο των διαφορετικά εκφρασμένων γονιδίων. Σημαντικός εμπλουτισμός καθορίστηκε από την p-value με κατώφλι (cutoff) το 0.05.

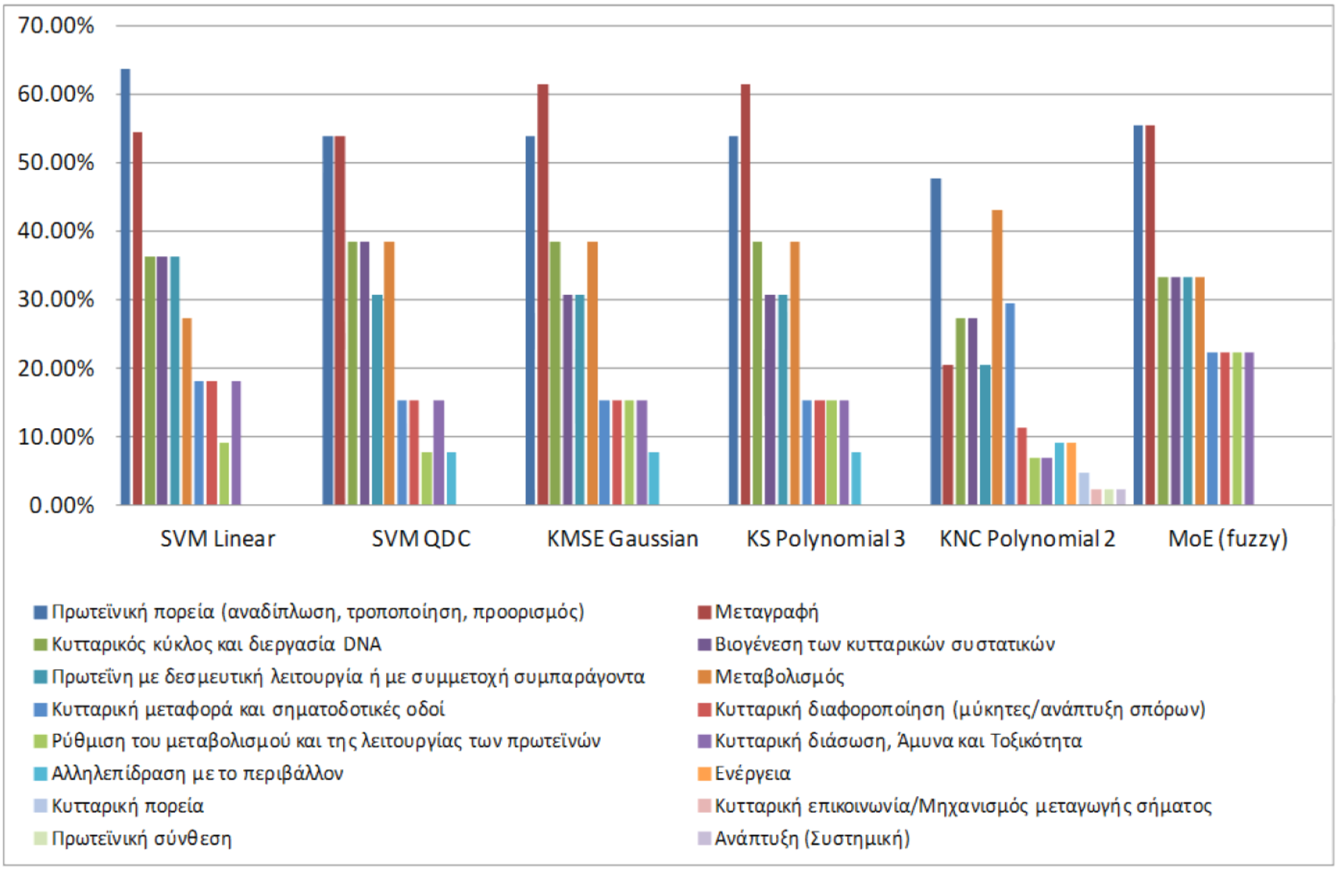
Εξετάζοντας κάθε γονιδιακή υπογραφή, με το εν λόγω σύστημα, προσδιορίστηκε ένας αριθμός από ενδιαφέρουσες βιολογικές διεργασίες στις οποίες συμμετέχουν τα γονίδια που τις αποτελούν. Αυτές οι διεργασίες φαίνονται στα επόμενα σχήματα.

Παρατηρούμε ότι τα 11, 13, 13, 13, 44 και 9 γονίδια των πέντε ατομικών ταξινομητών και του μείγματος αυτών αντίστοιχα, μπορεί να μην είναι τα ίδια ακριβώς, όμως φαίνεται ότι όλα συμμετέχουν στις ίδιες 16 βιολογικές διεργασίες (εικόνες 45 και 46). Οδηγώντας τη βιολογική ανάλυση ένα επίπεδο πιο βαθιά, μπορούμε να δούμε, επίσης, ότι συμμετέχουν και σε 12 σημαντικές ($p<0,05$)

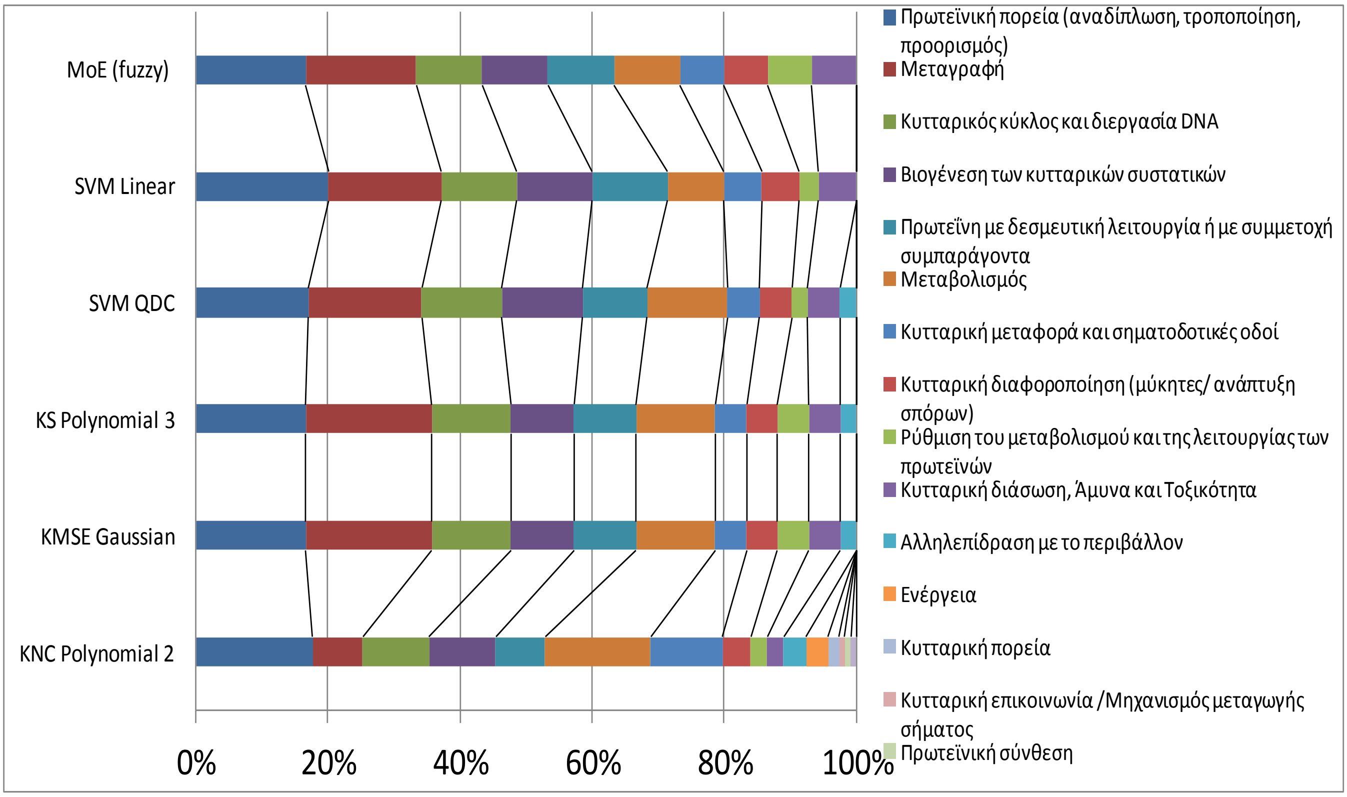
βιολογικές διεργασίες (εικόνες 47-49). Το ίδιο συμβαίνει και με τις γονιδιακές υπογραφές των τεσσάρων μειγμάτων, τα γονίδια των οποίων βρίσκουμε ότι μοιράζονται σε 12 και 10 κύριες βιολογικές διεργασίες (εικόνες 50-53), καθώς και σε 9 σημαντικές ($p<0,05$) βιολογικές διεργασίες (εικόνα 54).

Είναι σημαντικό να αναφερθεί ότι οι παρατηρούμενες βιολογικές διεργασίες, όπως φαίνεται στις εικόνες 45 έως 54, ανακλούν τις αναμενόμενες κυτταρικές διεργασίες σύμφωνα με τα δεδομένα που εξετάστηκαν (κυτταρικός κύκλος, απόκριση σε ακραίες περιβαλλοντικές συνθήκες). Για παράδειγμα, τα γονίδια τόσο των ατομικών ειδικών όσο και των μειγμάτων ειδικών μεθόδων συμμετέχουν σε ένα μεγάλο ποσοστό στις κατηγορίες του MIPS FunCat "κυτταρικός κύκλος και επεξεργασία DNA", διεργασία που ανακλά τα γεγονότα που συμβαίνουν κατά τη διάρκεια του κυτταρικού κύκλου, και "μεταγραφή", υποστηρίζοντας προηγούμενα ευρήματα σχετικά με την υψηλή περιοδική μεταγραφική δραστηριότητα κατά τη διάρκεια του κυτταρικού κύκλου του *Saccharomyces cerevisae* [21]. Επιπλέον, η κυτταρική διαδικασία "πρωτεϊνική πορεία" (protein fate) περιλαμβάνει τη διαδικασία της αναδίπλωσης των πρωτεϊνών (protein folding), που αποτελεί την πιο σημαντική διαδικασία απόκρισης στο θερμικό σοκ, αντικατοπτρίζοντας έτσι τις υπό μελέτη περιβαλλοντικές αλλαγές (π.χ. θερμικό σοκ).

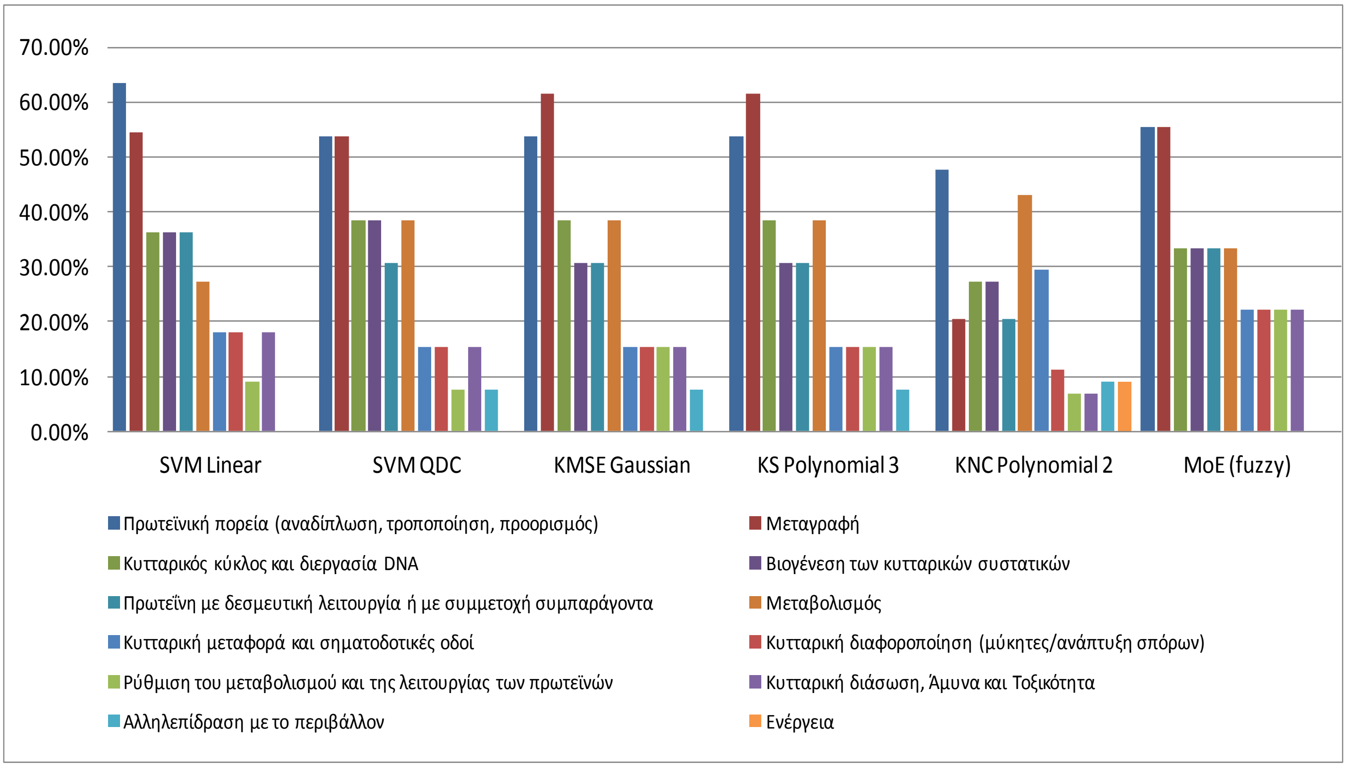
Σε αυτό το σημείο μπορεί να τονιστεί το εξής συμπέρασμα: *Ο τρόπος της εκχώρησης βαρών στα γονίδια και το κριτήριο της επιλογής τους με σκοπό την αναδρομική εξάλειψή τους από τον αλγόριθμο FRE-LNW της ενότητας 3.4, για την ανάδειξη εκείνων μόνο που εκφράζονται διαφορετικά στις δύο κλάσεις ενδιαφέροντος, είναι ο ίδιος. Αυτό που διαφοροποιείται στην διαδικασία ανάδειξης των σημαντικότερων γονιδίων και οδηγεί στο σχηματισμό 9 διαφορετικών γονιδιακών υπογραφών είναι οι διαφορετικοί αλγοριθμικοί παράμετροι (διαφορετικά μοντέλα ταξινόμησης) που εισάγονται στον τρόπο αξιολόγησης των υποσυνόλων των γονιδίων, και ο τρόπος επιλογής του τελικού αριθμού γονιδίων (διαφορετική μέγιστη απόδοση κατά το πρώτο επίπεδο αξιολόγησης για κάθε μοντέλο). Σαν απόρροια αυτών των παραγόντων, εμφανίζονται κάποιες διαφορές σε επίπεδο γονιδίων. Όμως υπάρχει πολύ μεγαλύτερη επικάλυψη σε επίπεδο βιολογικών διεργασιών, γεγονός που καθιστά απαραίτητη τη βιολογική εξερεύνηση των αποτελεσμάτων.*



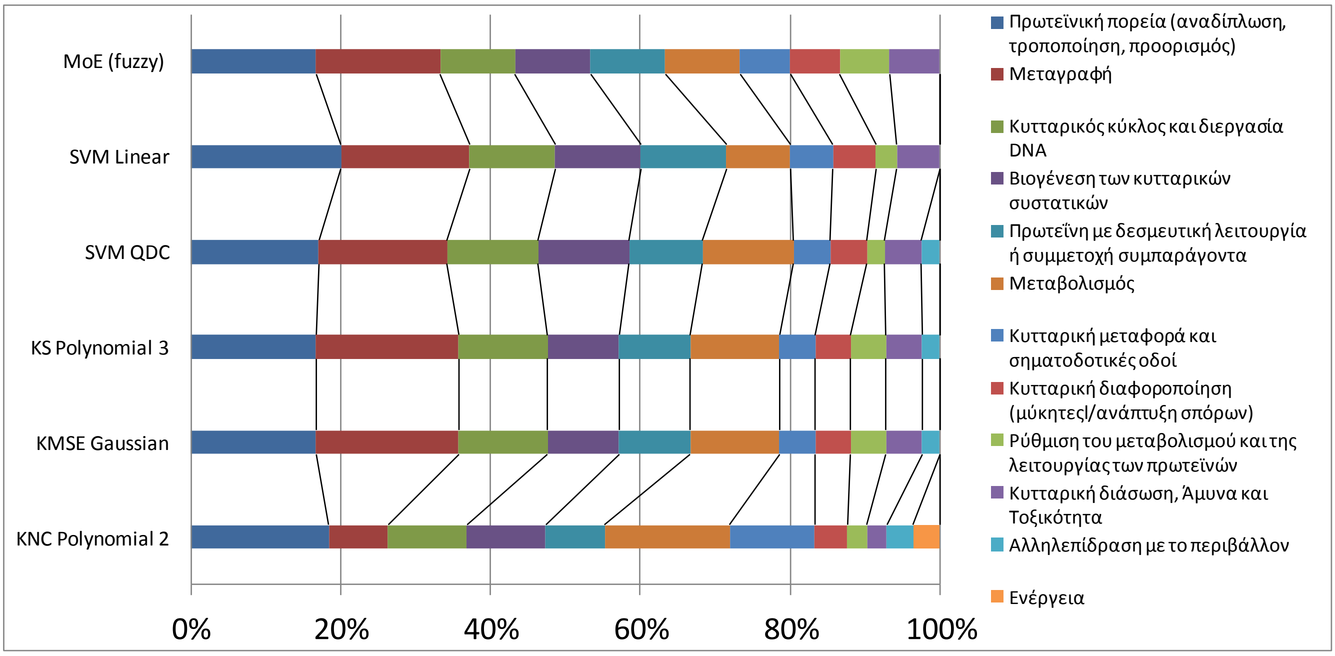
Εικόνα 45: (1η αναπαράσταση με απλά ραβδογράμματα) Συγκριτικά αποτελέσματα 16 βιολογικών διεργασιών για τους 5 ατομικούς ταξινομητές και του fuzzy συνδυασμού τους.



Εικόνα 46: (2η αναπαράσταση με ραβδογράμματα σε εκατοστιαία κλίμακα) Συγκριτικά αποτελέσματα 16 βιολογικών διεργασιών για τους 5 ατομικούς ταξινομητές και του fuzzy συνδυασμού τους.



Εικόνα 47: (1η αναπαράσταση με απλά ραβδογράμματα) Συγκριτικά αποτελέσματα 12 βιολογικών διεργασιών για τους 5 ατομικούς ταξινομητές και του fuzzy συνδυασμού τους.



Εικόνα 48: (2η αναπαράσταση με ραβδογράμματα σε εκατοστιαία κλίμακα) Συγκριτικά αποτελέσματα 12 βιολογικών διεργασιών για τους 5 ατομικούς ταξινομητές και του fuzzy συνδυασμού τους.

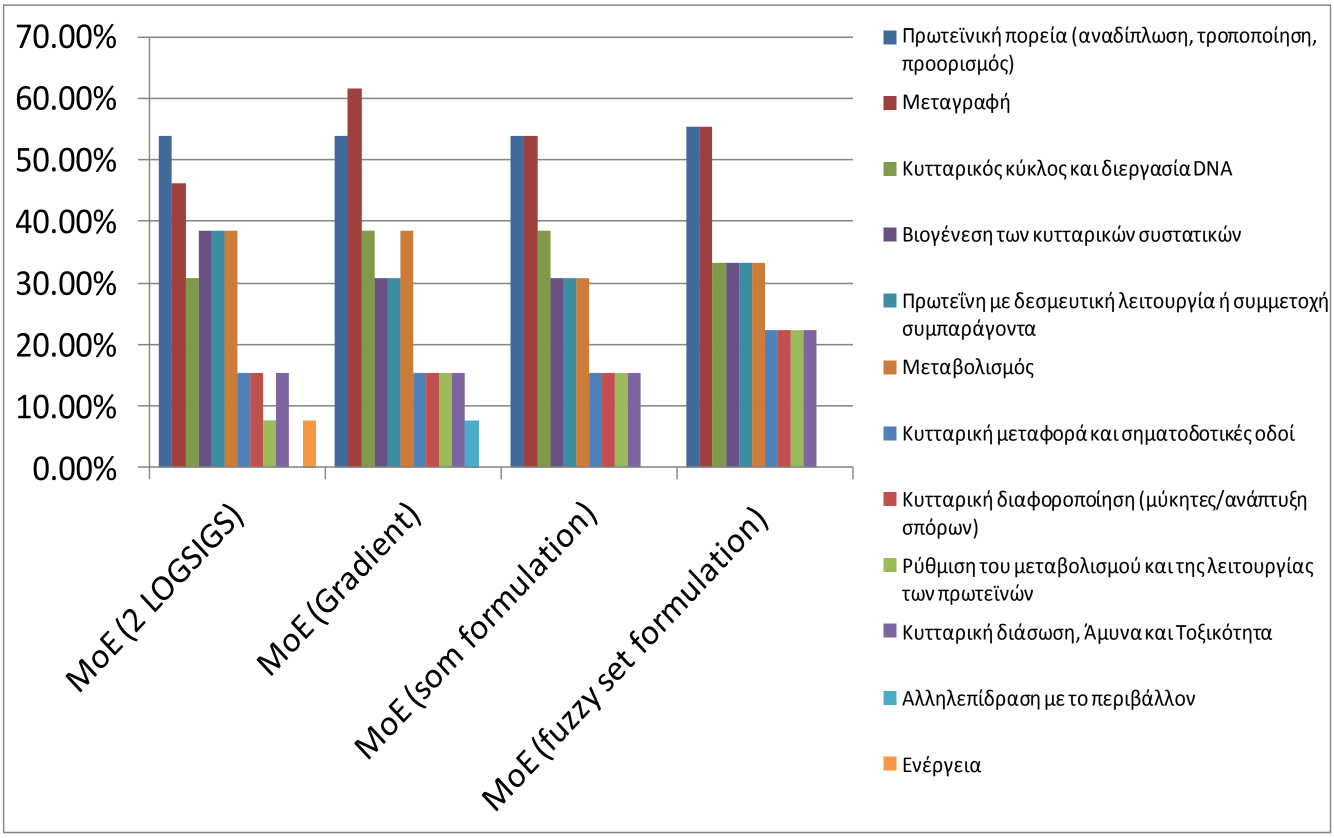
	Πρωτεϊνική πορεία (αναδίπλωση, τροποποίηση, προορισμός)	Μεταγραφή	Κυτταρικός κύκλος και διεργασία DNA	Βιογένεση των κυτταρικών συστατικών	Μεταβολισμός	Κυτταρική μεταφορά και σηματοδοτικές οδοί	Ρύθμιση του μεταβολισμού και της λειτουργίας των πρωτεϊνών	Πρωτεΐνη με δεσμευτική λειτουργία ή συμμετοχή συμπαράγοντα	Κυτταρική διαφοροποίηση (μύκητες/ανάπτυξη σπόρων)	Κυτταρική διάσωση, Άμυνα και Τοξικότητα	Αλληλεπίδραση με το κυτταρικό περιβάλλον	Ενέργεια
SVM Linear	63.6	54.5	9.09	18.1	9.09	0.00%	0.00%	27.2	18.1	0.00%	0	0
SVM QDC	53.8	53.8	38.4	38.4	22.99	0.00%	0.00%	23	15.3	0.00%	0.00%	0.00%
KMSE Gaussian	53.8	61.5	38.4	15.3	30.69	0.00%	15.3	23	15.3	0.00%	0.00%	0.00%
KS Polynomial 3	53.8	61.5	38.4	15.3	30.69	0.00%	15.3	23	15.3	0.00%	0.00%	0.00%
KNC Polynomial 2	47.7	20.4	27.2	27.2	43.1	29.5	0.00%	2.27	0.00%	0.00%	9.09	6.81
MoE (fuzzy)	55.5	55.5	0.00%	0.00%	22.2	0.00%	22.2	0.00%	22.2	11.1	0.00%	0.00%

Εικόνα 49: Συγκριτικά αποτελέσματα 12 βιολογικών διεργασιών για τους 5 ατομικούς ταξινομητές και το fuzzy συνδυασμό τους με χρήση του p-value.

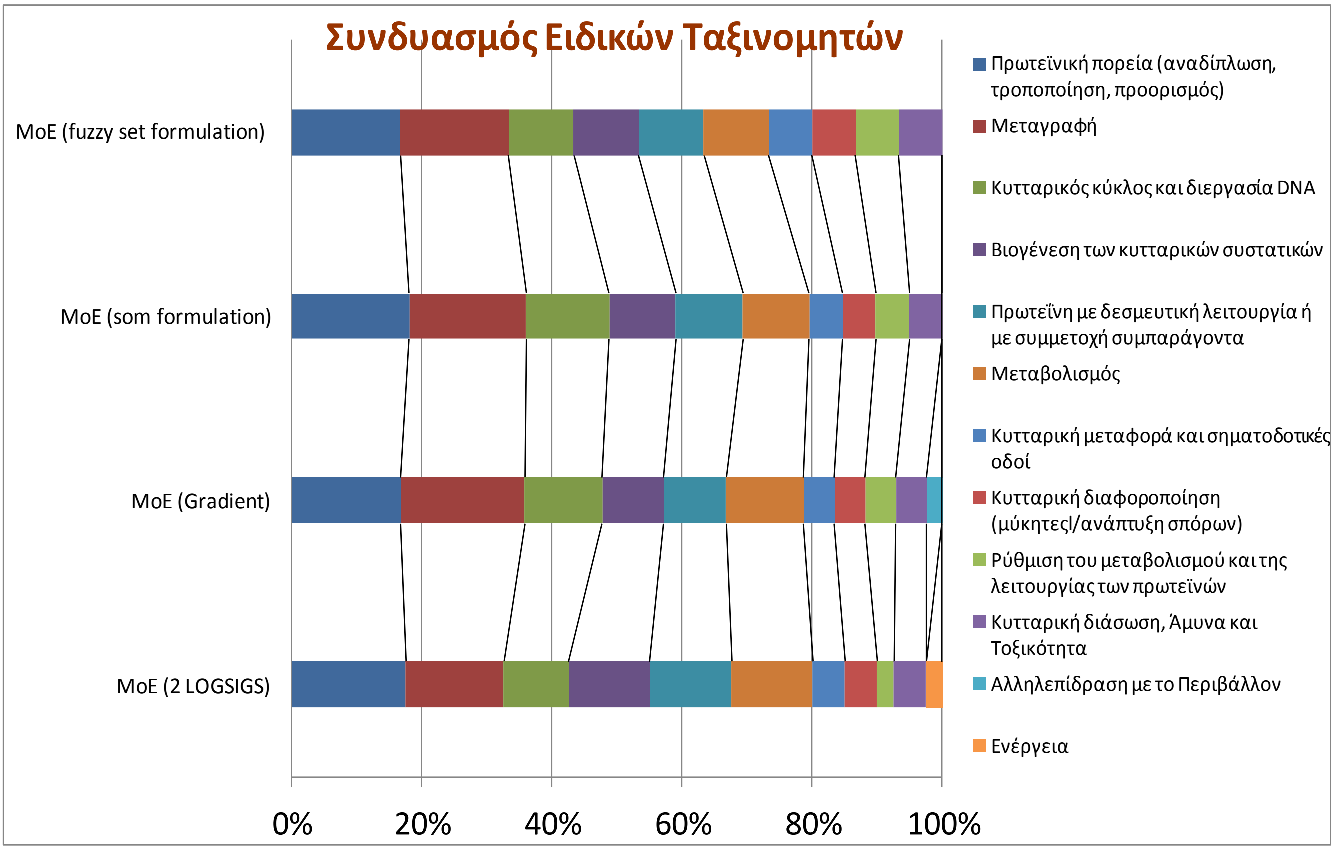
	στατιστικά σημαντικές ΒΔ	**στατιστικά μη σημαντικές ΒΔ**
SVM Linear	24.39	75.61
SVM QDC	22.47	77.53
KMSE Gaussian	23.26	76.74
KS Polynomial 3	23.26	76.74
KNC Polynomial 2	**26.77**	73.23
MoE (fuzzy)	22.86	77.14

(πράσινο)	p<0.05 στο επίπεδο-1
(γαλάζιο)	p<0.05 στο επίπεδο-2 έως το επίπεδο-5
(καφέ)	p<0.00005 στο επίπεδο-1 και το επίπεδο-3; p<0.0005 στο επίπεδο-2
(κόκκινο)	p<0.05 στο επίπεδο-1, p<0.0005 στο επίπεδο-3
(μπλε)	p<0.0005 στο επίπεδο-5

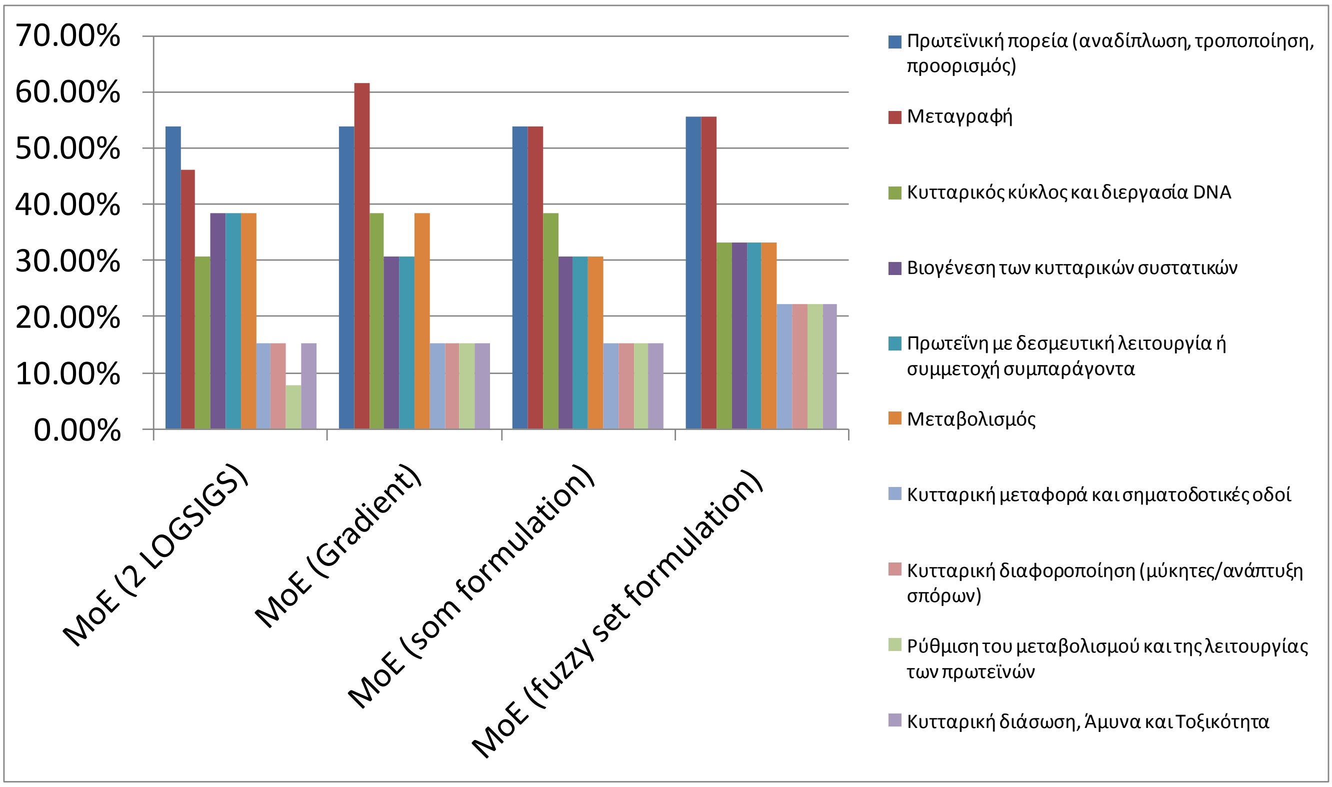
Από την εικόνα 48, φαίνεται ότι οι σημαντικές διεργασίες δε διαφοροποιούνται από τις αρχικές. Η τιμή p δηλώνει τη σημασία του αριθμού των εμφανίσεων για κάθε κατηγορία (βιολογική διεργασία) στο σύνολο των γονιδίων του κάθε ταξινομητή, λαμβάνοντας υπόψη τον αριθμό των συμπτώσεων για ολόκληρο το γονιδίωμα του *Saccharomyces cerevisiae* (6130 γονίδια). Οι κατηγορίες (βιολογικές διεργασίες) αξιολογούνται ως πιο σημαντικές όσο μικρότερη είναι η τιμή του p. Αυτές οι βιολογικές διεργασίες συμφωνούν με τις αρχικές βιολογικές διεργασίες και είναι οι αναμενόμενες, όπως αναφέρθηκε παραπάνω, αφού ανακλούν τις υπό μελέτη συνθήκες του συνόλου δεδομένων του σακχαρομύκητα (κυτταρικός κύκλος, απόκριση σε περιβαλλοντικά σοκ).



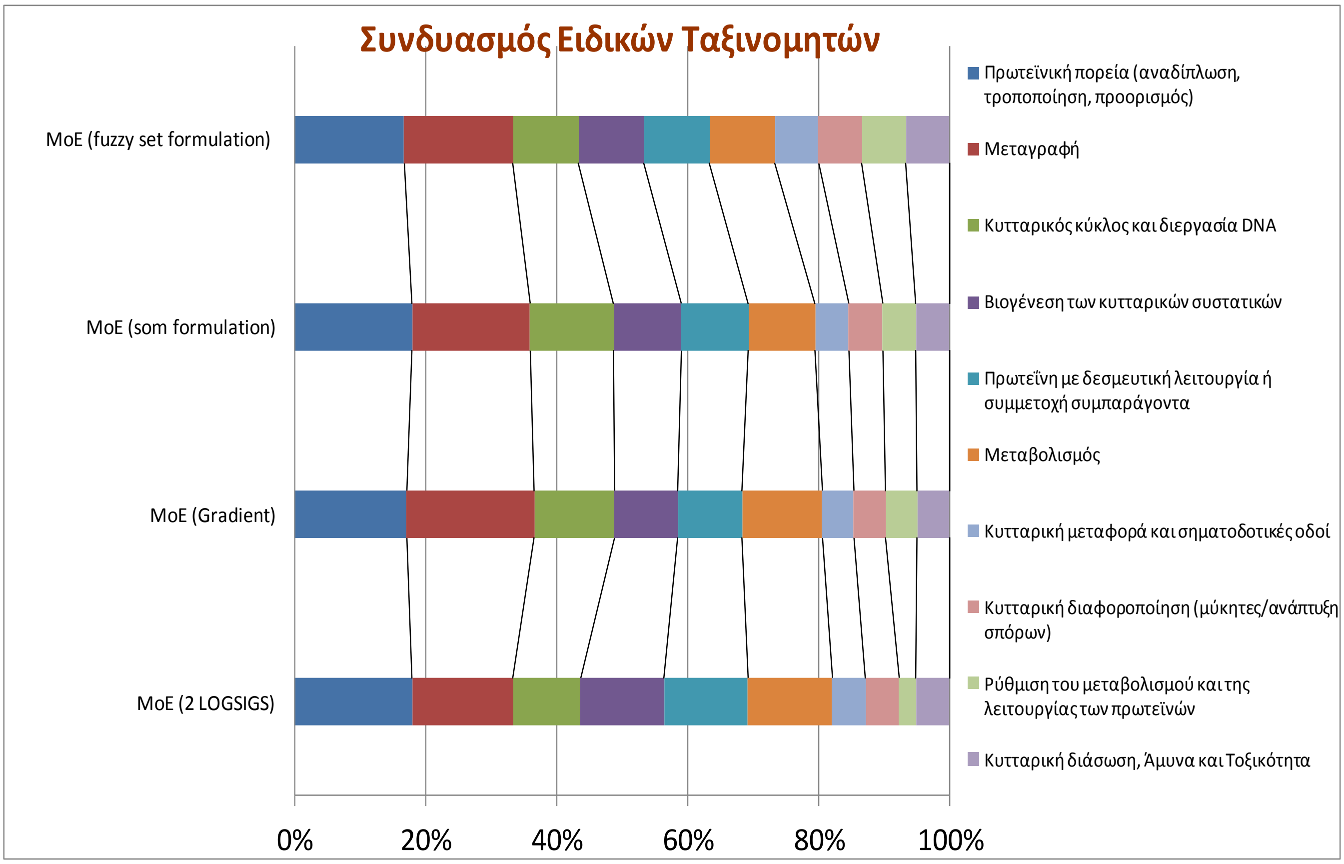
Εικόνα 50: (1η αναπαράσταση με απλά ραβδογράμματα) Συγκριτικά αποτελέσματα 12 βιολογικών διεργασιών για τους 4 διαφορετικούς συνδυασμούς ειδικών.



Εικόνα 51: (2η αναπαράσταση με ραβδογράμματα σε εκατοστιαία κλίμακα) Συγκριτικά αποτελέσματα 12 βιολογικών διεργασιών για τους 4 διαφορετικούς συνδυασμούς ειδικών.



Εικόνα 52: (1η αναπαράσταση με απλά ραβδογράμματα) Συγκριτικά αποτελέσματα 10 βιολογικών διεργασιών για τους 4 διαφορετικούς συνδυασμούς ειδικών.



Εικόνα 53: (2η αναπαράσταση με ραβδογράμματα σε εκατοστιαία κλίμακα) Συγκριτικά αποτελέσματα 10 βιολογικών διεργασιών για τους 4 διαφορετικούς συνδυασμούς ειδικών.

	Πρωτεϊνική πορεία (αναδίπλωση, τροποποίηση, προορισμός)	Μεταγραφή	Κυτταρικός κύκλος και διεργασία DNA	Βιογένεση των κυτταρικών συστατικών	Μεταβολισμός	Πρωτεΐνη με δεσμευτική λειτουργία ή συμμετοχή συμπαράγοντα	Κυτταρική διαφοροποίηση (μύκητες/ανάπτυξη σπόρων)	Ρύθμιση του μεταβολισμού και της λειτουργίας των πρωτεϊνών	Κυτταρική διάσωση, Άμυνα και Τοξικότητα
MoE (2 LOGSIGS)	53.8	46.1	7.69	38.4	15.38	23	15.3	0.00%	0.00%
MoE (Gradient)	53.8	61.5	38.4	15.3	30.69	23	15.3	15.3	0.00%
MoE (som formulation)	53.8	53.8	38.4	15.3	22.99	23	15.3	15.3	0.00%
MoE (fuzzy set formulation)	55.5	55.5	0.00%	0.00%	22.2	0.00%	22.2	22.2	11.1

Εικόνα 54: Συγκριτικά αποτελέσματα 9 βιολογικών διεργασιών για τους 4 διαφορετικούς συνδυασμούς ειδικών, με χρήση του p-value.

	στατιστικά σημαντικές ΒΔ	**στατιστικά μη σημαντικές ΒΔ**
MoE (2 LOGSIGS)	21.74	78.26
MoE (Gradient)	23.25	76.74
MoE (som formulation)	**26.5**	73.5
MoE (fuzzy set formulation)	22.86	77.14

(πράσινο)	$p<0.05$ στο επίπεδο-1
(γαλάζιο)	$p<0.05$ στο επίπεδο-2 έως το επίπεδο-5
(κόκκινο)	$p<0.05$ στο επίπεδο-1, $p<0.0001$ στο επίπεδο-2 και το επίπεδο-4

ΚΕΦΑΛΑΙΟ 6: Συζήτηση και Μελλοντικές Προεκτάσεις

Σε αυτό το κεφάλαιο συνοψίζονται τα σημαντικότερα συμπεράσματα που προκύπτουν από τα αποτελέσματα του κεφαλαίου 5, έτσι όπως αυτά διαμορφώθηκαν ακολουθώντας την προτεινόμενη μεθοδολογία του κεφαλαίου 4. Επίσης, δίνονται ιδέες για μελλοντική έρευνα σχετική με το αντικείμενο της παρούσας εργασίας.

6.1 Συμπεράσματα της Εργασίας

Σε αυτή την εργασία εξετάζουμε την επιρροή των αλγοριθμικών παραμέτρων στην προγνωστική δύναμη μίας γονιδιακής υπογραφής καθώς και στη διαδικασία επιλογής της ίδιας της υπογραφής. Προς αυτή την κατεύθυνση, προτείνουμε την ακρίβεια πρόβλεψης, όπως επίσης και τη συχνότητα εμφάνισης των γονιδίων ως μέτρα για τη στατιστική σημαντικότητά τους, σε συνδυασμό με μία διαδικασία αναδρομικής εξάλειψης χαρακτηριστικών. Καταλήγουμε σε 9 γονιδιακές υπογραφές με τα σχετικότερα γονίδια. Εξετάσαμε τόσο το αρχικό σύνολο γονιδίων όσο και τα γονίδια των τελικών υπογραφών με διαφορετικούς ατομικούς ταξινομητές αλλά και με διαφορετικούς συνδυασμούς αυτών. Τα ευρήματα αυτής της εργασίας συνοψίζονται ως εξής:

- ✓ Παρατηρούμε ότι η απόδοση ενός ταξινομητή δεν εξαρτάται μόνο από το μηχανισμό του ίδιου του ταξινομητή (μορφή συνάρτησης απόφασης) αλλά και από τον πυρήνα που θα εφαρμοστεί σε αυτόν, καθώς η απόδοσή του μεταβάλλεται με τη χρήση διαφορετικών πυρήνων. Προσαρμόζοντας έναν πυρήνα [48], κατάλληλο, για να βελτιωθεί η ευθυγράμμιση με τα δείγματα του συνόλου εκπαίδευσης, τα οποία φέρουν ετικέτα, αυξάνεται σημαντικά η ευθυγράμμιση και με τα δείγματα του συνόλου ελέγχου, τα οποία τα μοντέλα μάθησης καλούνται να ταξινομήσουν, δίνοντας αρκετά βελτιωμένη ακρίβεια ταξινόμησης. Οπότε, η επιλογή του κατάλληλου πυρήνα σε κάθε ταξινομητή είναι ένα θέμα ζωτικής σημασίας για την ακρίβεια που αυτός θα σημειώσει κατά την ταξινόμηση.

- ✓ Η σωστή επιλογή ενός μικρού αριθμού από σημαντικά γονίδια, προσφέρει καλή προγνωστική ακρίβεια σε όλα τα μοντέλα που εξετάστηκαν σε αυτή την εργασία, με εξαίρεση τον kernel nearest centroid ταξινομητή του οποίου η απόδοση μειώνεται για τα λίγα γονίδια της γονιδιακής του υπογραφής. Συμπεραίνουμε ότι, τα μοντέλα ταξινομούν καλύτερα όταν από το σύνολο δεδομένων έχουν αφαιρεθεί αρκετά μη σχετικά γονίδια, εκτός από τον παραπάνω ταξινομητή, ο οποίος φαίνεται ότι αποδίδει καλύτερα όταν τα γονίδια είναι περισσότερα.
- ✓ Τα μείγματα ειδικών σημειώνουν είτε την ίδια περίπου απόδοση με εκείνη των ατομικών ταξινομητών, είτε ακόμα καλύτερη. Το μείγμα των ειδικών που βασίζεται στην επιλογή ταξινομητών με som διαχωρισμό του χώρου χαρακτηριστικών, αποδίδει πολύ καλύτερα από τους επιμέρους τοπικούς ειδικούς, αφού βασίζεται στη συμπληρωματικότητα των αποφάσεών τους, κάτι το οποίο δεν παρατηρείται πάντα για τα υπόλοιπα τρία μείγματα που βασίζονται στη συγχώνευση των επιμέρους καθολικών ειδικών, όπου επικρατεί ανταγωνιστικότητα μεταξύ τους για τη δημιουργία της συνδυασμένης απόφασης.
- ✓ Από τα αποτελέσματα γίνεται φανερό ότι διαφορετικά συστήματα επιλογής οδηγούν σε ποικίλα μεγέθη γονιδιακής υπογραφής, με διαφορές στα επιλεγμένα γονίδια. Ωστόσο, όταν περιγράφουμε τα γονίδια κάθε υπογραφής [51] φαίνεται ότι αυτά επικαλούνται τις ίδιες βιολογικές διεργασίες, με μικρές πιθανές διαφορές στη σχετική συχνότητα συμμετοχής.

Η μεθοδολογία που αναπτύχθηκε σε αυτή την εργασία ανέδειξε τη ζωτικής σημασίας χρήση του κατάλληλου πυρήνα σε κάθε ταξινομητή για την επίτευξη της καλύτερης δυνατής απόδοσης από εκείνον καθώς και τη χρησιμότητα των μειγμάτων ειδικών ως μοντέλα, των οποίων η απόδοση είναι συγκρίσιμη ή και καλύτερη από την αντίστοιχη των επιμέρους ταξινομητών.

Επιπλέον η επιλογή ενός συνόλου σημαντικών γονιδίων αυξάνει σημαντικά την ακρίβεια ταξινόμησης των μοντέλων μειώνοντας την αβεβαιότητα που εισάγεται από τις παραμέτρους αυτών. Διαφορετικοί αλγόριθμοι που λειτουργούν πάνω στο ίδιο σύνολο δεδομένων μπορεί να οδηγούν σε διαφορετικές γονιδιακές υπογραφές με ελάχιστη ή καθόλου επικάλυψη μεταξύ τους και επίσης, οι διαφορετικές παράμετροι του ίδιου αλγορίθμου μπορεί να επηρεάζουν τα αποτελέσματα της εκτίμησης των δεικτών.

Από αυτά τα αποτελέσματα, επιβεβαιώνεται η ανάγκη για βιολογική εξερεύνηση των αποτελεσμάτων που σχετίζονται με το γονότυπο και τη βιολογική υποστήριξη των μέτρων ομοιότητας και/ή συναρτήσεων απόστασης που χρησιμοποιούνται στις στατιστικές προσεγγίσεις, προκειμένου να μειωθεί η αβεβαιότητα στη λήψη αποφάσεων που περιλαμβάνουν τα γονιδιωματικά δεδομένα.

6.2 Μελλοντικές Επεκτάσεις

Μελλοντική εργασία, πάνω σε αυτό το θέμα, θα μπορούσε να επικεντρωθεί στην εφαρμογή όλων των παραπάνω μεθόδων και μοντέλων μάθησης και πρόβλεψης, σε διαφορετικά σύνολα δεδομένων με στόχο να εκτιμηθεί αν αποδίδουν το ίδιο καλά σε διαφορετικά δεδομένα. Αυτός είναι και ένας τρόπος να ελεχθούν τα μείγματα ειδικών ως προς την ευρωστία τους, δηλαδή να συγκριθεί η απόδοσή τους πάνω σε διαφορετικά δεδομένα, για να διαπιστωθεί αν η ακρίβειά τους κυμαίνεται στα ίδια περίπου ποσοστά, για να μπορέσουν έτσι να θεωρηθούν αξιόπιστα ως μοντέλα.

Επίσης, θα μπορούσαν τα μείγματα των ειδικών αυτής της εργασίας να εφαρμοστούν με διαφορετικούς συνδυασμούς ατομικών ταξινομητών, ως προς τον αριθμό και τη φύση των τελευταίων, για περαιτέρω ανάλυση της λειτουργίας τους.

Επιπρόσθετα, θα μπορούσαν να εξεταστούν και άλλα μοντέλα μείγματος ειδικών, όπως η μέθοδος που βασίζεται σε expectation maximization αλγόριθμο.

Τέλος, σημαντικό θα είναι να εξετάσουμε και διαφορετικούς τρόπους επιλογής γονιδίων πέραν του RFE-LNW, καθώς επίσης και τη μίξη διαφορετικών μεθόδων επιλογής, όπου περιμένουμε πολύ μικρή επικάλυψη γονιδιακών υπογραφών για να εξετάσουμε αν η μεθοδολογία μίξης μπορεί στατιστικά ή βιολογικά να οδηγήσει σε αυξημένη επικάλυψη των αποτελεσμάτων.

Βιβλιογραφία

[1] Gerald Schaefer, Tomoharu Nakashima. "Data Mining of Gene Expression Data by Fuzzy and Hybrid Fuzzy Methods", *IEEE Transactions on Information Technology in Biomedicine, Vol. 14, No. 1,* pp. 23-29, January 2010.

[2] Soumya Raychaudhuri, Patrick D. Sutphin, Jeffrey T. Chang and Russ B. Altman. "Basic microarray analysis: grouping and feature reduction", *Trends in Biotechnology, Vol. 19, No. 5,* pp. 189-193, 5 May 2011.

[3] Rosalia Maglietta, Annarita D'Addabbo,Ada Piepoli,Francesco Perri, Sabino Liuni,Graziano Pesole,Nicola Ancona. "Selection of relevant genes in cancer diagnosis based on their prediction accuracy", *Artificial Intelligence in Medicine,* pp. 29-44, 6 June 2006.

[4] Gabrys, Dymitr Ruta, Bogdan. "An Overview of Classifier Fusion Methods", *Computing and Information Systems, Vol. 7,* pp. 1-10, 2000.

[5] R. Bailey. "About.com Biology", The New York Times Company, [Online]. Available: http://biology.about.com/od/geneticsglossary/g/Genes.htm. [Accessed 8 May 2012].

[6] "http://ec.asm.org/content/7/8.cover-expansion".

[7] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. "Cluster analysis and display of genome-wide expression patterns", *The National Academy of Sciences, pp*. 14863–14868, December 1998.

[8] Alvis Brazma, Pascal Hingamp,John Quackenbush,Gavin Sherlock,Paul Spellman,Chris Stoeckert,John Aach,Wilhelm Ansorge,Catherine A. Ball,Helen C. Causton,Terry Gaasterland,Patrick Glenisson,Frank C.P. Holstege,Irene F. Kim. "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data", *Nature Publishing Group,* pp. 365-371, December 2001.

[9] "Dynamics of Cellular Regulation", Computational Systems Biology, 22 July 2008, [Online]. Available: http://www.csb.ethz.ch/research/dynamic. [Accessed 9 May 2012].

[10] "Budding Yeast: Saccharomyces cerevisiae", 18 May 2011, [Online]. Available: http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/Y/Yeast.html. [Accessed 9 May 2012].

[11] "Modeling the Budding Yeast Cell Cycle", 13 April 2005, [Online]. Available: http://mpf.biol.vt.edu/research/budding_yeast_model/pp/intro.php. [Accessed 9 May 2012].

[12] Satoshi Niijima, Satoru Kuhara. "Gene subset selection in kernel-included feature space", *Pattern Recognition Letters, Vol. 27,* pp. 1884-1892, 27 June 2006.

[13] Gotoh, Xiaosheng Wang,Osamu. "A Robust Gene Selection Method for Microarray-based Cancer Classifiacation", *Cancer Informatics,* pp. 15-30, 2010.

[14] Bin Tang, Malcolm I. Heywood, Michael Shepherd. "Input Partitioning to Mixture of Experts".

[15] LESKI, Jacek. "A FUZZY IF-THEN RULE-BASED NONLINEAR CLASSIFIER", *Int. J. Math. Comput. Sci.,Vol. 13, No. 2,* pp. 215-223, 2003.

[16] K Dobbin, R. Simon. "Sample size planning for developing classifiers using high-dimensional DNA microarray data", *Biostatistics, Vol. 8, No. 1,* pp. 101-117, 2007.

[17] "Θέματα Γονιδιωματικής Πρωτεομικής και Βιοτεχνολογίας", [Online]. Available: http://www.bio-optics.ece.ntua.gr/postgrand_notes/2%CE%B7_%CE%B4%CE%B9%CE%B1%CE%BB%CE%B5%CE%BE%CE%B7_%CF%85%CE%BB%CE%B9%CE%BA%CF%8C.pdf]. [Accessed 21 May 2012].

[18] Cherry, J. M. Ball,C. Weng, S. Juvik, G. Schmidt, R. Adler. "Nature (London) 387", pp. 67–73, 1997.

[19] Søren Brunak, Rasmus Wernersson,Thomas S. Jensen & Ulrik de Lichtenberg. "State-of-the-art in Saccharomyces cerevisiae synchronization protocols", *DIAMONDS Deliverable 1-D1.1.*

[20] Audrey P. Gasch, Paul T. Spellman,Camilla M. Kao. "Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes", *Molecular Biology of the Cell, Vol. 11,* pp. 4241-4257, December 2000.

[21] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R.,Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein and Bruce Futcher. "Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization", *Molecular Biology of the Cell, Vol. 9,* pp. 3273–3297, December 1998.

[22] "Pattern Recognition", [Online]. Available: http://en.wikipedia.org/wiki/Pattern_recognition. [Accessed 3 April 2012].

[23] J.P. Marques de Sa, Pattern recognition, Concepts, Methods and Applications, Springer, 2001.

[24] "Statistical classification", [Online]. Available: http://en.wikipedia.org/wiki/Classification. [Accessed 2 April 2012].

[25] "Cluster analysis", [Online]. Available: http://en.wikipedia.org/wiki/Cluster_analysis. [Accessed 2 April 2012].

[26] Klaous-Robert Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, Bernhard Scholkopf. "An Introduction to Kernel-Based Learning Algorithms", *IEEE Transactions On Neural Networks, Vol. 12, No. 2.,* March 2001.

[27] Minseok Seo, Sejong Oh. "Derivation of an artificial gene to improve classification accuracy upon gene selection", *Computational Biology and Chemistry,* pp. 1-12, 20 November 2011.

[28] "Linear and nonlinear classifiers", [Online]. Available: http://cs.joensuu.fi/pages/whamalai/expert/lecture6.htm. [Accessed 6 April 2012].

[29] "Fundamentals of Statistics", [Online]. Available: http://www.statistics4u.com/fundstat_eng/cc_data_structure.html. [Accessed 6 April 2012].

[30] "Nonlinear SVMs", [Online]. Available: http://nlp.stanford.edu/IR-book/html/htmledition/nonlinear-svms-1.html. [Accessed 6 April 2012].

[31] Michael I.Jordan. "The Kernel Trick", *Advanced Topics in Learning & Decision Making*.

[32] Asa Ben-Hur, Jason Weston. "A User's Guide to Support Vector Machines".

[33] Suzi Fei. "The "Kernel Trick", A Computational Biology Example using Support Vector Machines" 13 January 2009, [Online]. Available: http://web.cecs.pdx.edu/~mm/MachineLearningWinter2009/slides/Jan14KernelSlides.pdf, [Accessed 8 April 2012].

[34] A. Perolini. "Genetic Algorithms and Kernel Matrix-based Criteria Combined Approach to Perform Feature and Model Selection for Support Vector Machines", *Word Academy of Science, Engineering and Technology,* 2012.

[35] [Online]. Available: http://iccle.googlecode.com/svn/trunk/share/img/. [Accessed 10 May 2012].

[36] Baudat G., Anouar F. "Kernel-based Methods and Function Approximation", *Electronics International,* pp. 1244-1249, 2001.

[37] Trevor Hastie, Robert Tibshirani, Jerome H. Friedman. The elements of statistical learning, data mining, inference, and prediction, New York: Springer, 2001.

[38] Gutierrez-Osuna, Ricardo, Intelligent Sensor Systems, "Lecture 13: Validation", [Online]. Available: http://research.cs.tamu.edu/prism/lectures/iss/iss_l13.pdf. [Accessed 2 May 2012].

[39] "Feature Selection", [Online]. Available: http://en.wikipedia.org/wiki/Feature_selection. [Accessed 11 April 2012].

[40] Andreas G.K. Janecek, Wilfried N. Ganstener,Michael A. Demel,Gerhard F. Ecker. "On the relationship Between Feature Selection and Classification Accuracy", *JMRL: Workshop and Conference Proceedings, Vol. 4,* pp. 90-105.

[41] Michalis E. Blazadonakis, Michalis Zervakis. "Wrapper filtering criteria via linear neuron and kernel approaches", *Computers in Biology and Medicine,* pp. 894-912, 16 May 2008.

[42] "Support Vector Machine", 28 February 2012, [Online]. Available: http://en.wikipedia.org/wiki/Support_vector_machine. [Accessed 18 April 2012].

[43] "SVM - Support Vector Machines, Introduction to Support Vector Machine (SVM) Models", [Online]. Available: http://www.dtreg.com/svm.htm. [Accessed 18 April 2012].

[44] Te-Ming Huang, Vojislav Kecman, Ivica Kopriva. "Comparison between RFE-SVMs and the Nearest Shrunken Centroid Method", *Kernel Based Algorithms for Mining Huge Data Sets*, Springer.

[45] "A SVM-based server for rice blast prediction dedicated to the farming community", [Online]. Available: http://www.imtech.res.in/raghava/rbpred/algorithm.html. [Accessed 5 May 2012].

[46] Masulli, Giorgio Valentini and Francesco. "Ensembles of Learning Machines".

[47] Michael I. Jordan, Robert A. Jacobs. "Modular and hierarchical learning systems", *The Handbook of Brain Theory and Neural Networks*, Cambridge, MIT Press.

[48] Nello Cristianini, John Shawe-Taylor,Andre Elisseeff,Jaz Kandola. "On Kernel-Target Alignment".

[49] Michael Hsing, Kendall Grant Byler and Artem Cherkasov. "The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks", *BMC Systems Biology,* 16 September 2008.

[50] Lam, Xudong Xie and Kin-Man. "Gabor-Based Kernel PCA With Doubly Nonlinear Mapping for Face Recognition With a Single Face Image", *IEEE Transactions on Image Processing, Vol. 15, No. 9,* pp. 2481-2492, September 2006.

[51] J-Z. Xu, C-W. Wong. "Hunting for robust gene signature from cancer profiling data: Sources of variability, different interpretations, and recent methodological developments", *Cancer Letters, Vol. 296,* pp. 9-16, 2010.

[52] Igor V. Tetko1, Igor V. Rodchenkov, Mathias C. Walter, Thomas Rattei and Hans-Werner Mewes. "Beyond the ‘best’ match: machine learning annotation of protein sequences by integration of different sources of information", *Bioinformatics Original Paper, Vol. 24,*

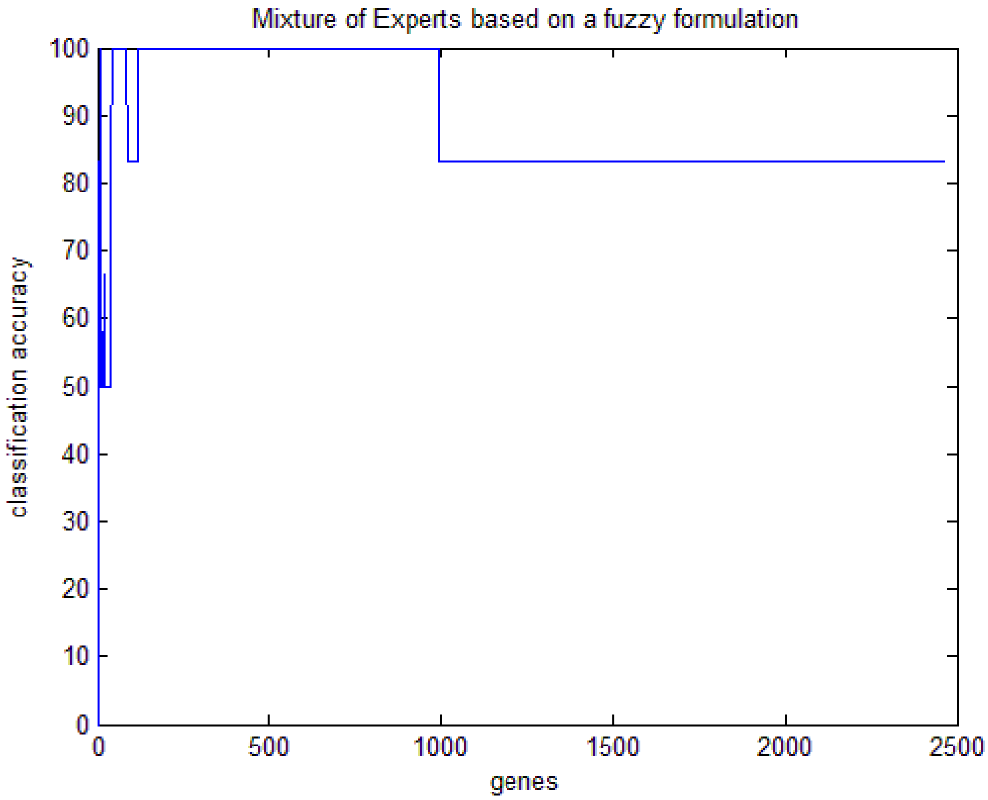
No. 5, pp. 621–628, 3 January 2008.

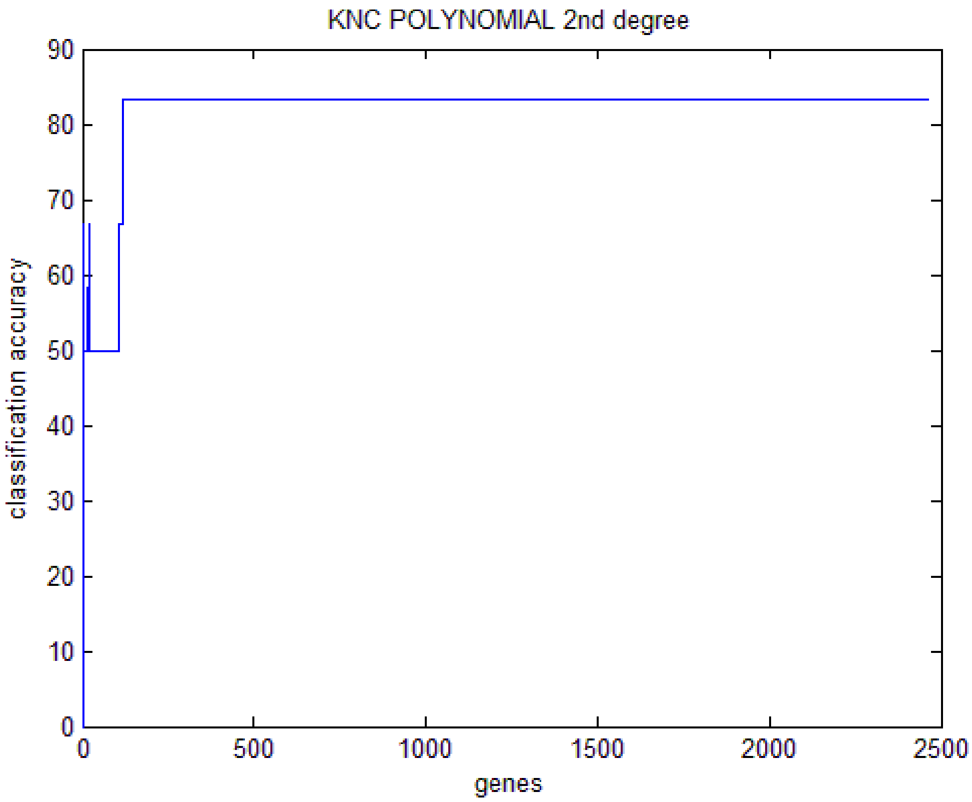
[53] Steffen Priebe a, Jörg Linde, Daniela Albrecht, Reinhard Guthke, Axel A. Brakhage. "FungiFun: A web-based application for functional categorization of fungal genes and proteins", *Fungal Genetics and Biology,Vol. 48,* pp. 353–358, 1 November 2010.

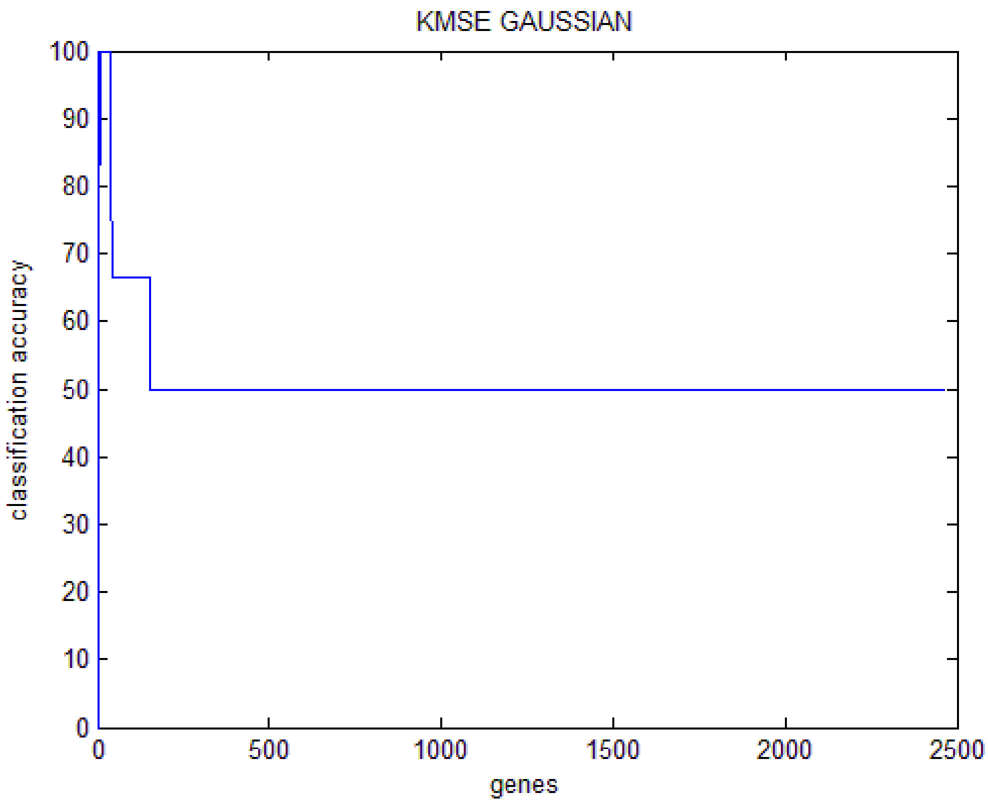
Παράρτημα Α

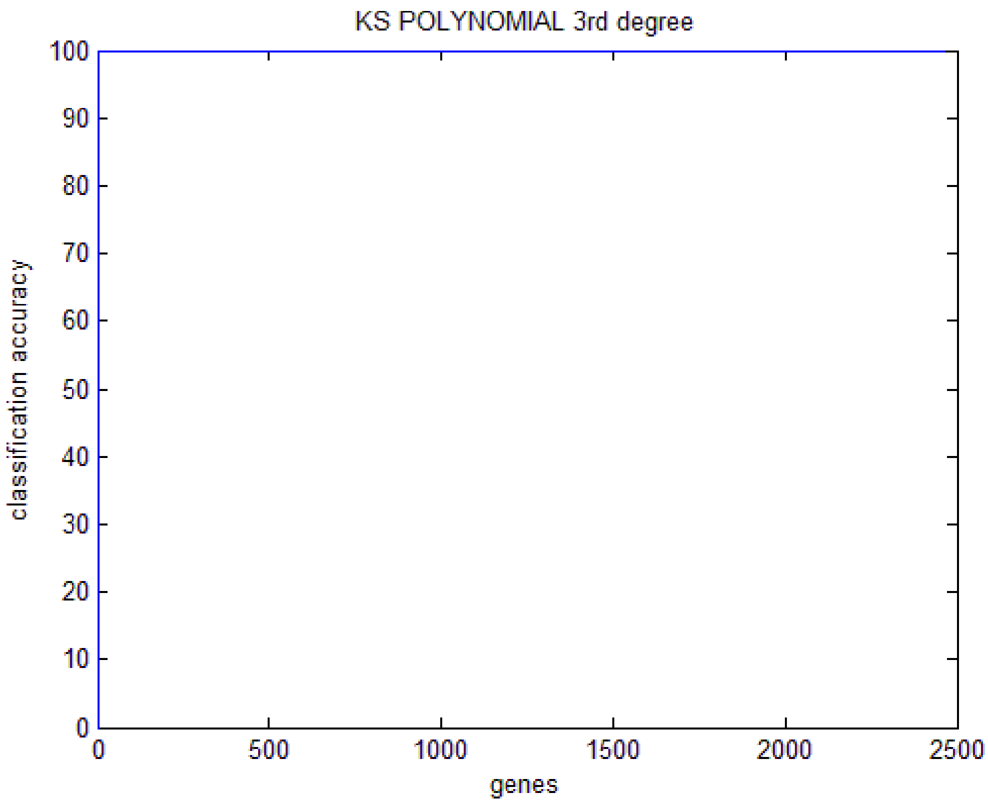
Ακολουθούν οι γραφικές παραστάσεις που διεξήχθησαν κατά το πρώτο επίπεδο αξιολόγησης δεδομένων, σύμφωνα με τη μεθοδολογία που περιγράφεται αναλυτικά στην ενότητα 4.3. Εκτελούνται 5 folds, στο κάθε ένα από τα οποία παρουσιάζεται η ακρίβεια ταξινόμησης για τους πέντε ατομικούς ταξινομητές kernel nearest centroid με πολυωνυμικό πυρήνα δευτέρου βαθμού, kernel minimum squared error με Gaussian πυρήνα, kernel subspace με πυρήνα πολυωνυμικό τρίτου βαθμού, SVM με τετραγωνικό και SVM με γραμμικό πυρήνα, καθώς και για το μείγμα αυτών που βασίζεται στη συγχώνευσή τους με ένα συνδυασμό ασαφών βαρών.

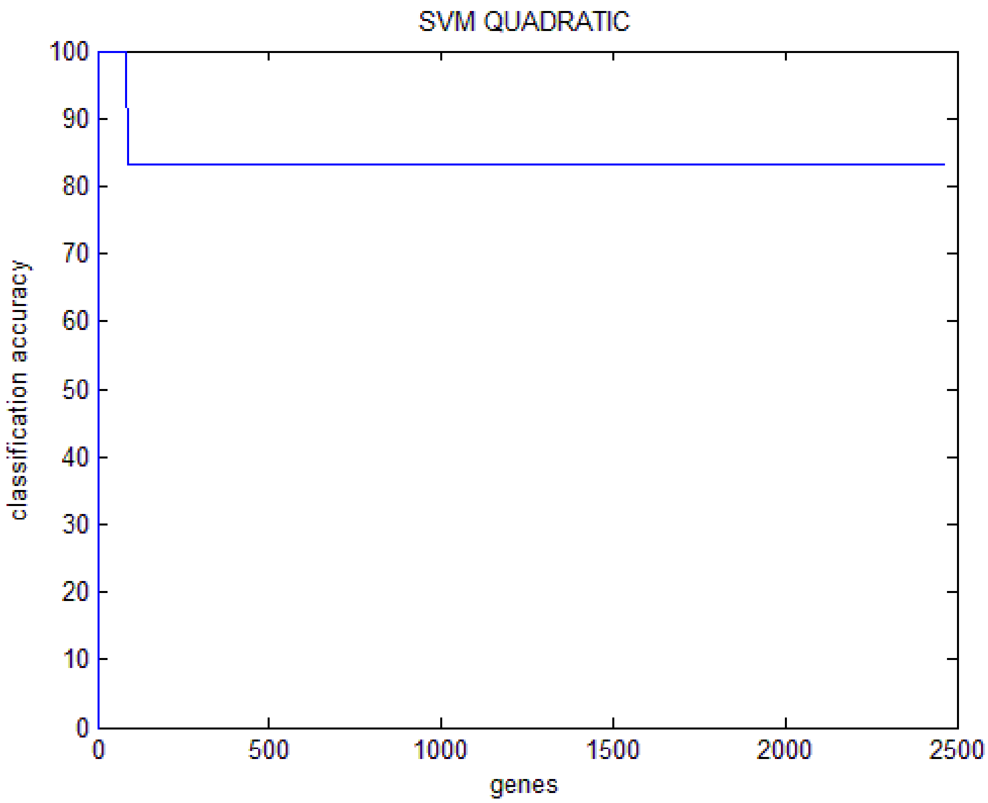
1st FOLD

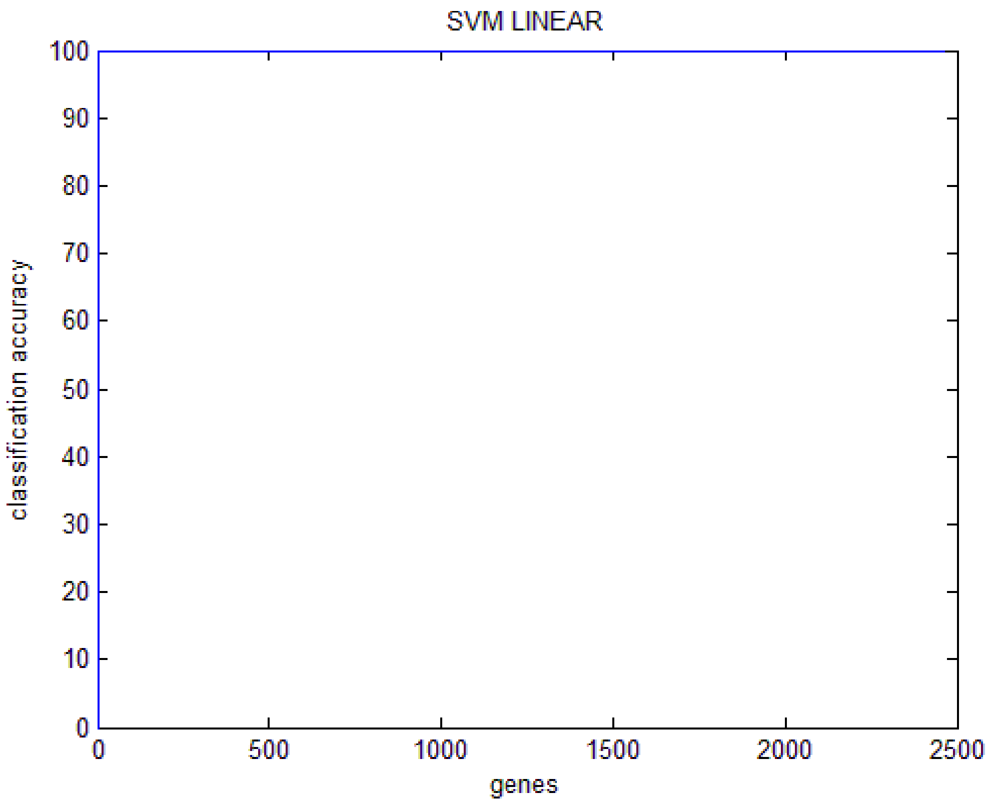

Mixture of Experts based on a fuzzy formulation
classification accuracy
genes
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500
2000
2500


KNC POLYNOMIAL 2nd degree
classification accuracy
genes
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500
2000
2500

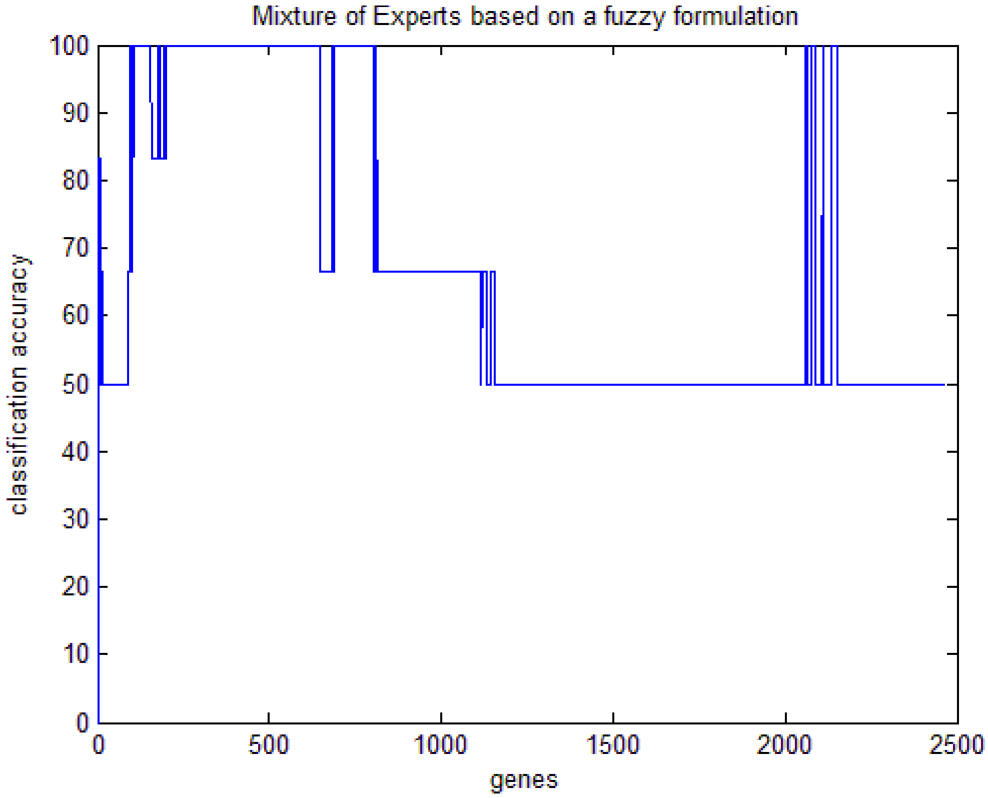

KMSE GAUSSIAN
classification accuracy
genes
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500
2000
2500


KS POLYNOMIAL 3rd degree
classification accuracy
genes
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500
2000
2500

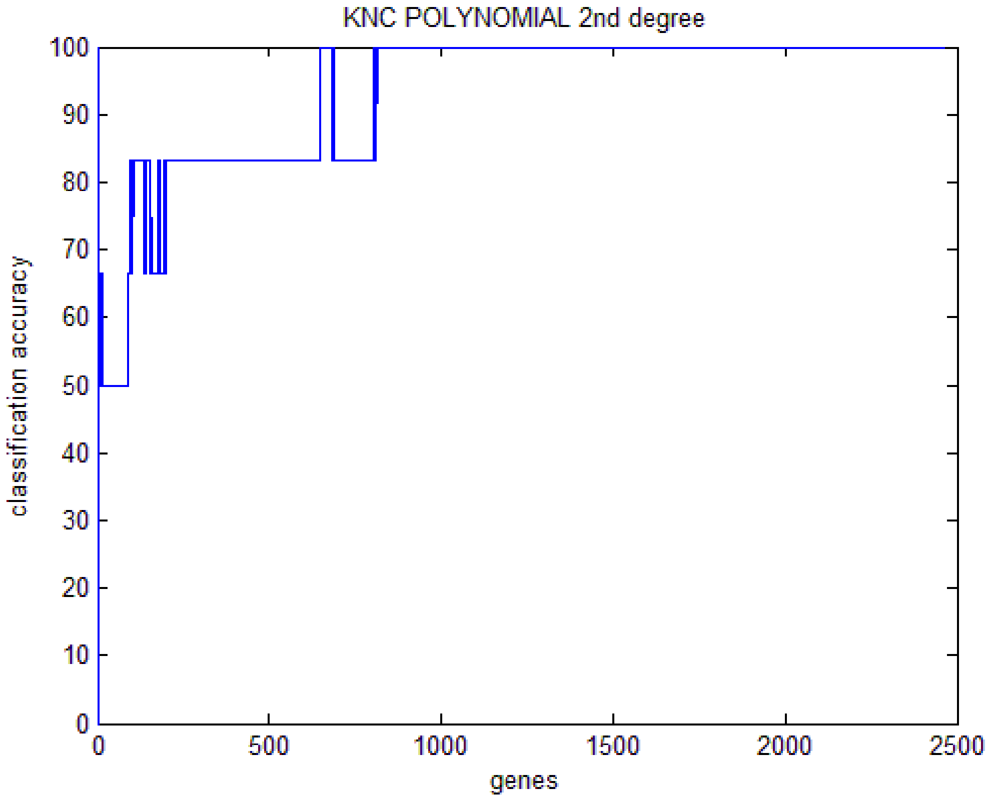

SVM QUADRATIC
100
90
80
70
60
50
40
30
20
10
0
classification accuracy
0
500
1000
1500
2000
2500
genes


SVM LINEAR
100
90
80
70
60
50
40
30
20
10
0
classification accuracy
0
500
1000
1500
2000
2500
genes

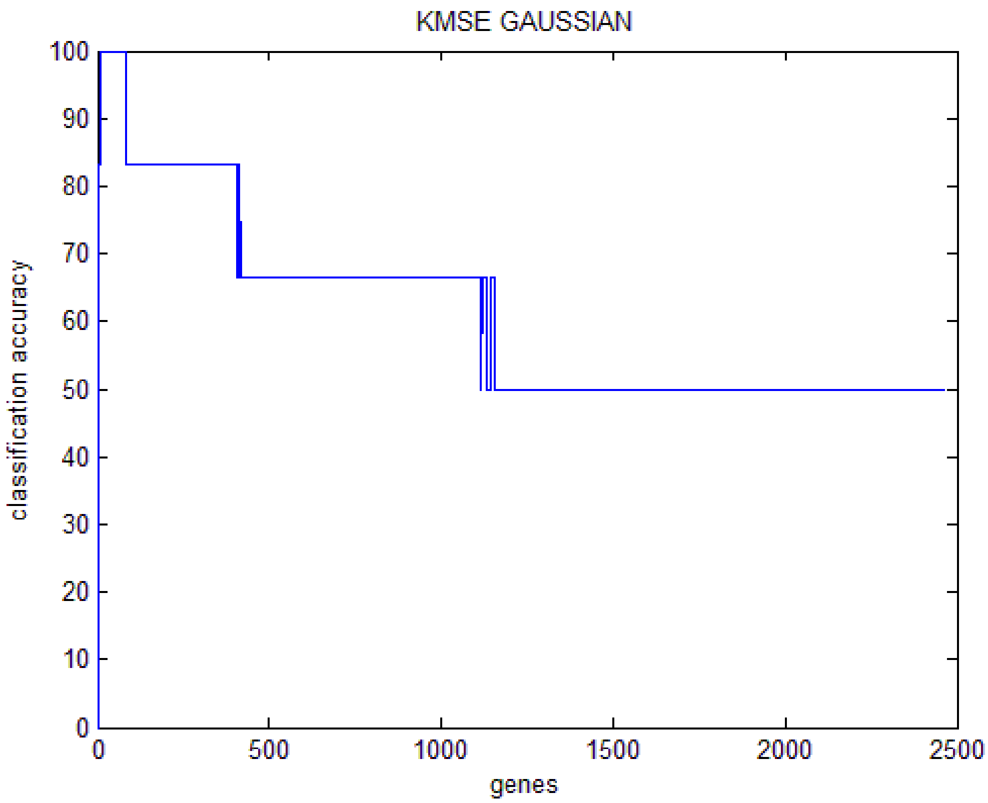
2^{nd} FOLD

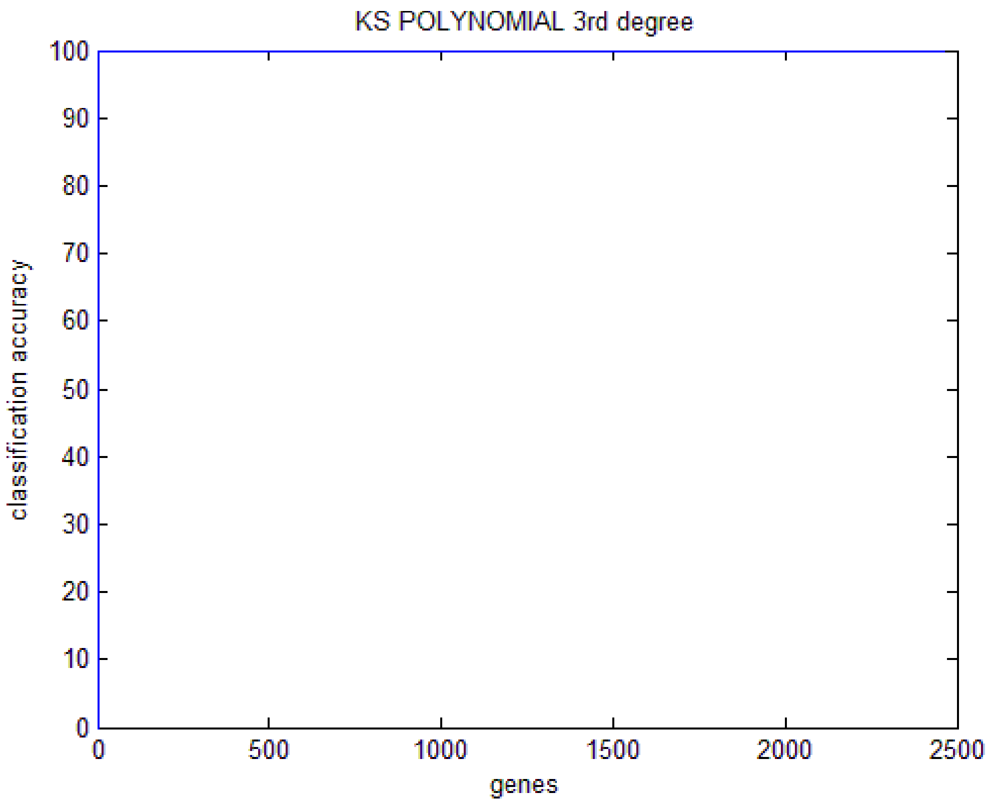


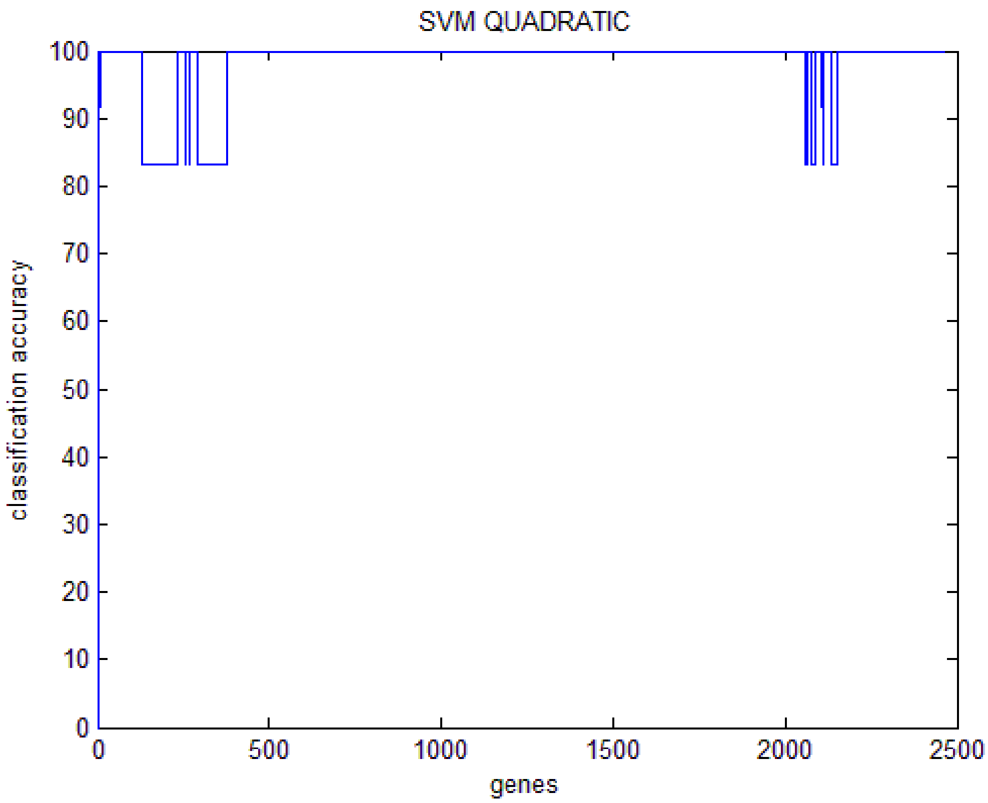
Mixture of Experts based on a fuzzy formulation
classification accuracy
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500
2000
2500
genes

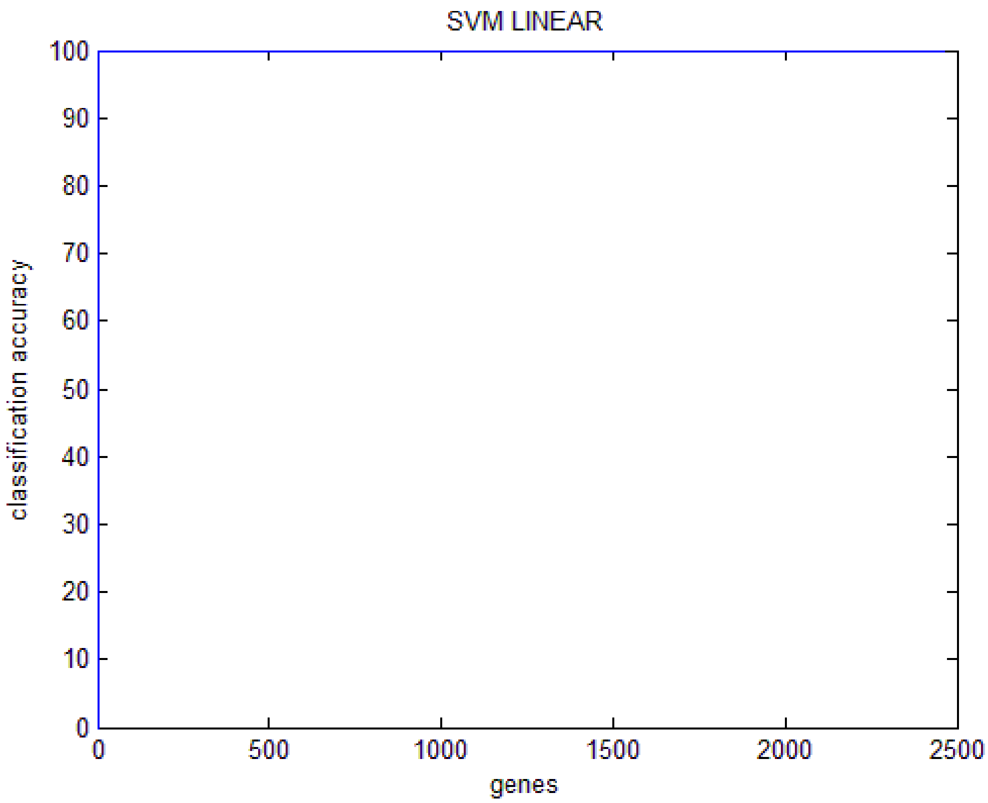


KNC POLYNOMIAL 2nd degree
classification accuracy
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500
2000
2500
genes

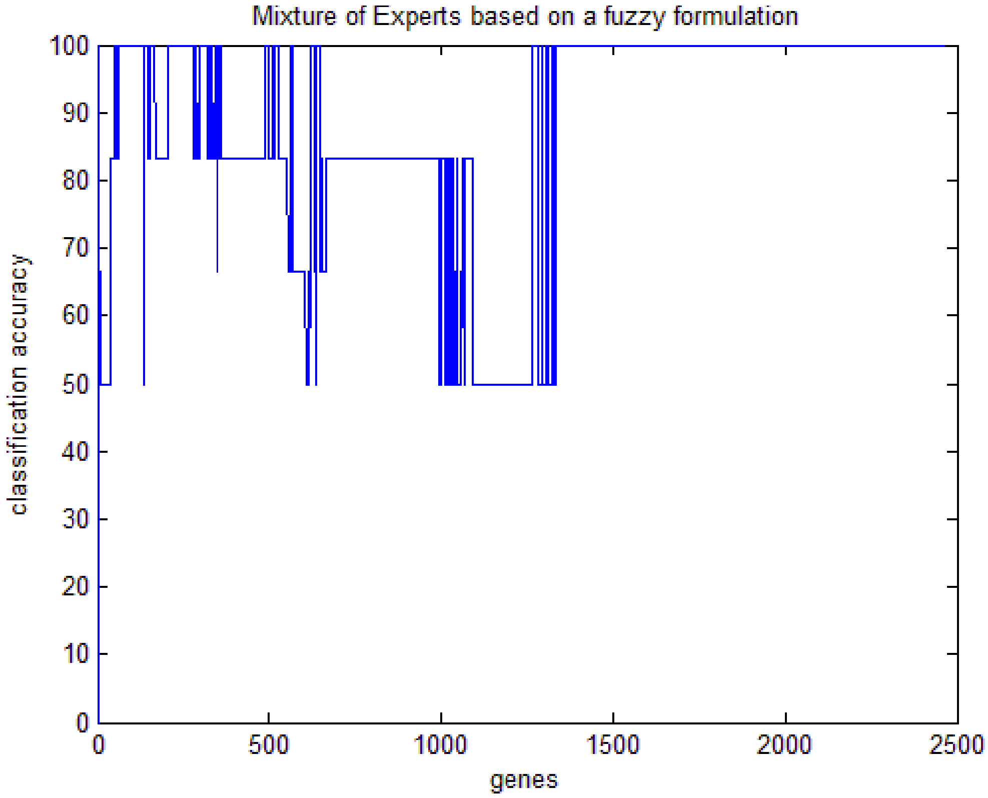

KMSE GAUSSIAN
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
500
1000
1500
2000
2500

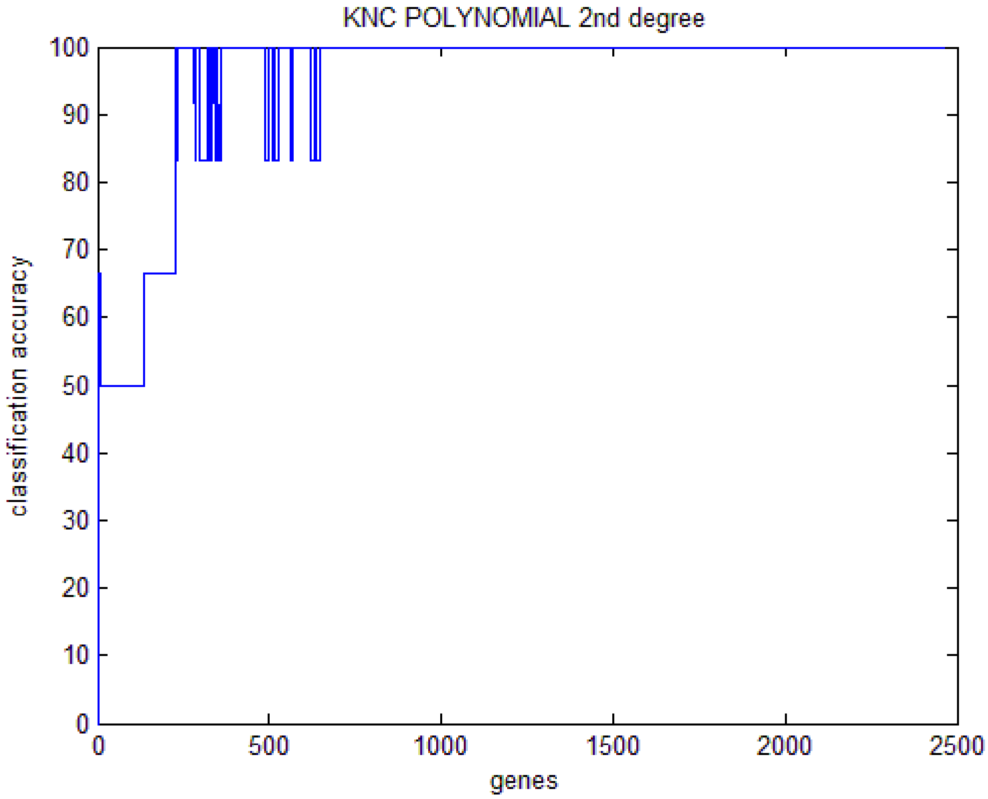

KS POLYNOMIAL 3rd degree
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
500
1000
1500
2000
2500

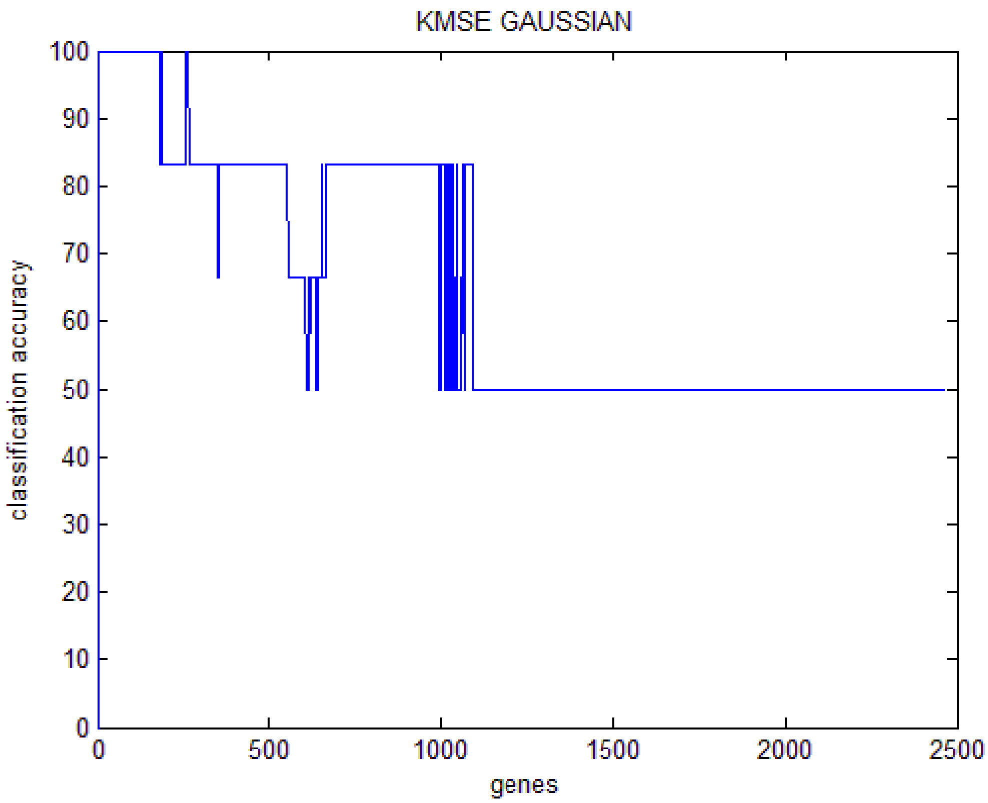

SVM QUADRATIC
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
0
500
1000
1500
2000
2500

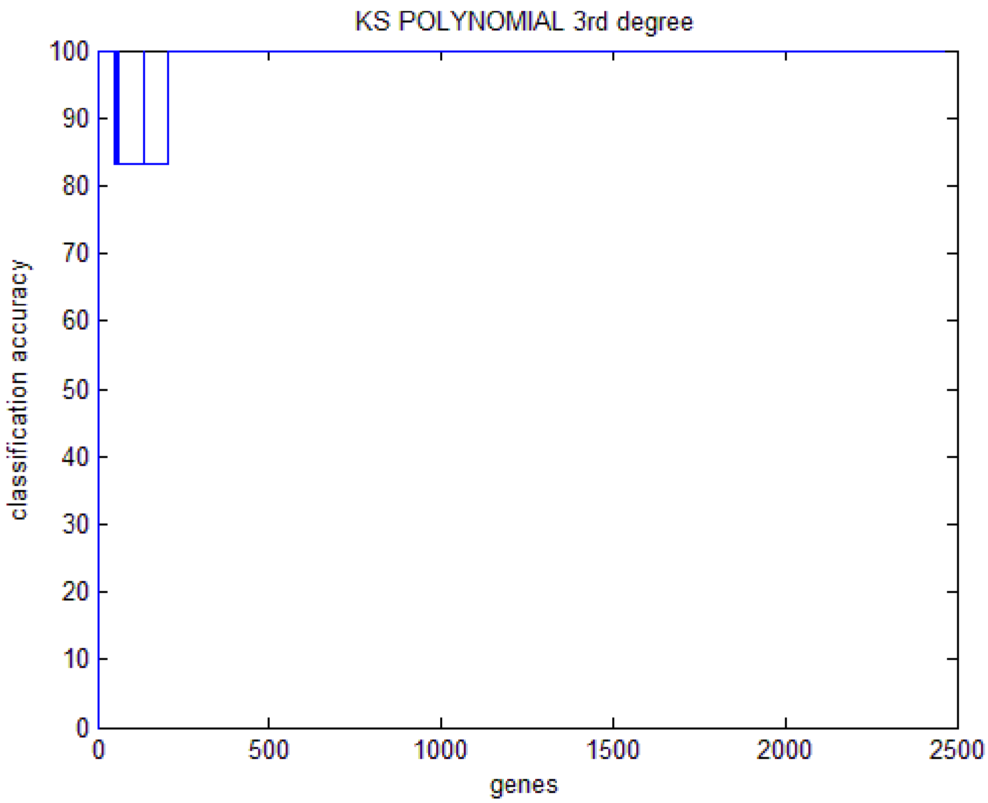

SVM LINEAR
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
0
500
1000
1500
2000
2500

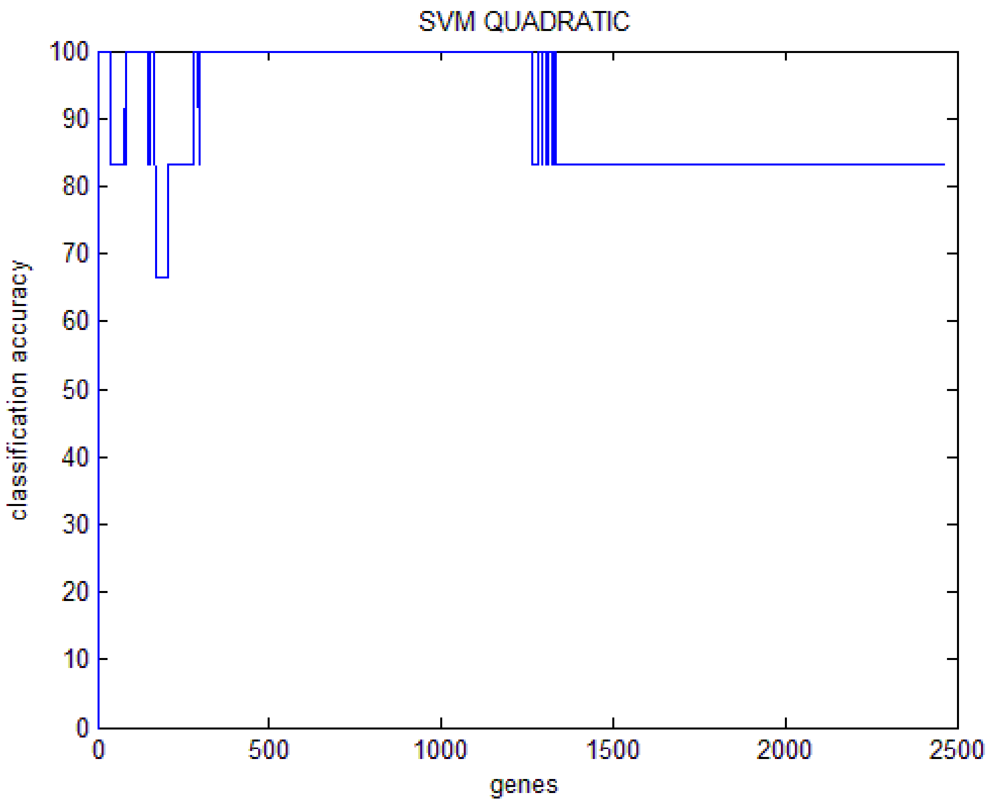
3rd FOLD

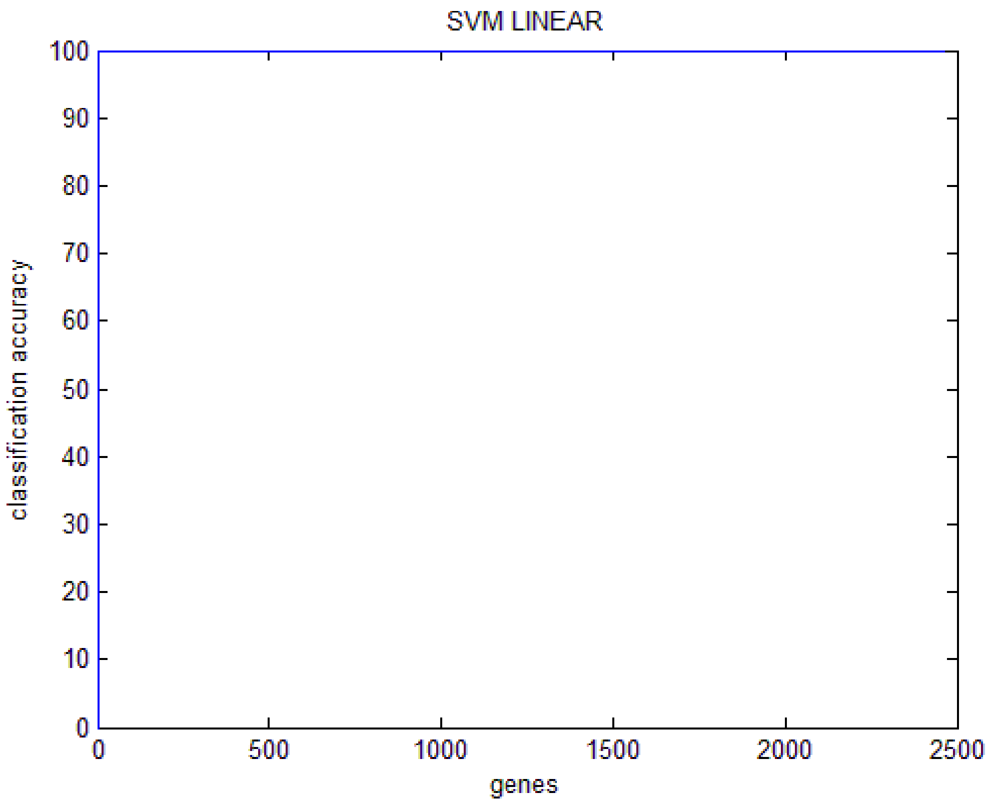

Mixture of Experts based on a fuzzy formulation
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
0
500
1000
1500
2000
2500


KNC POLYNOMIAL 2nd degree
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
0
500
1000
1500
2000
2500

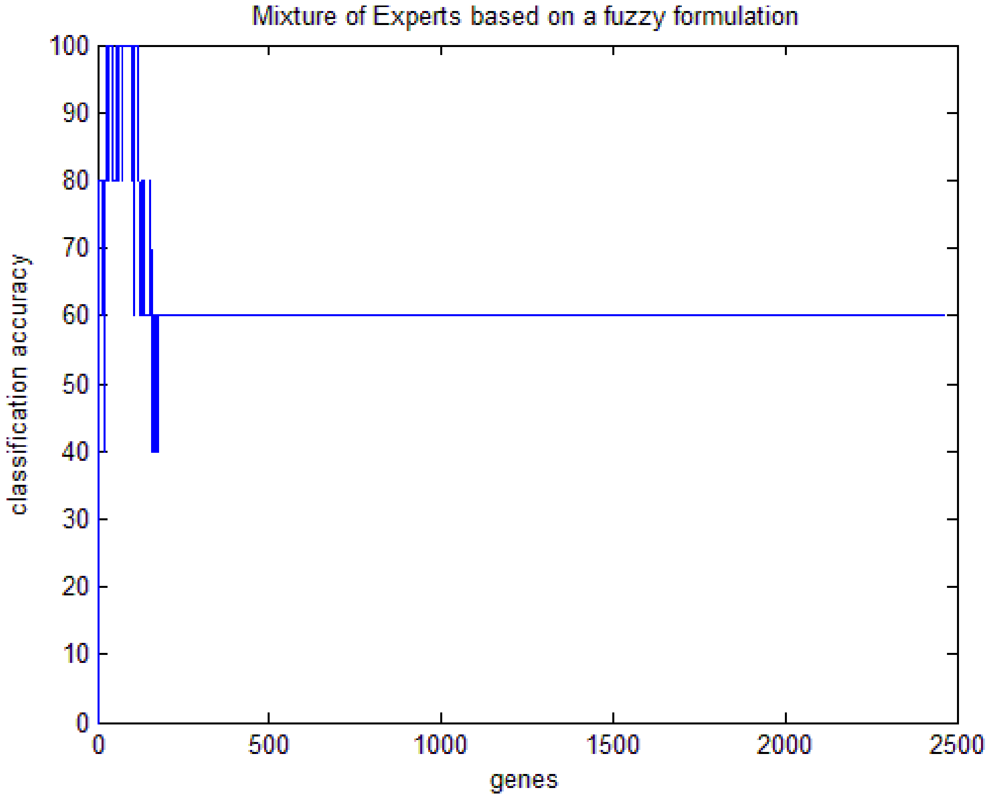

KMSE GAUSSIAN
classification accuracy
genes


KS POLYNOMIAL 3rd degree
classification accuracy
genes

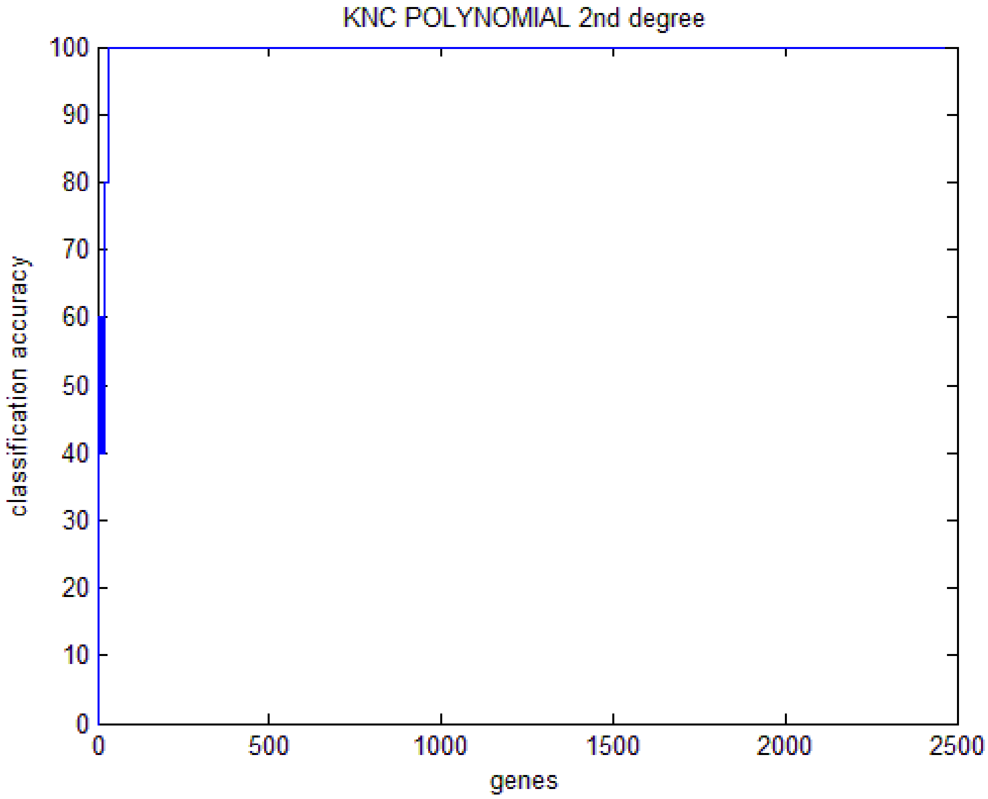

SVM QUADRATIC
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
0
500
1000
1500
2000
2500


SVM LINEAR
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
0
500
1000
1500
2000
2500

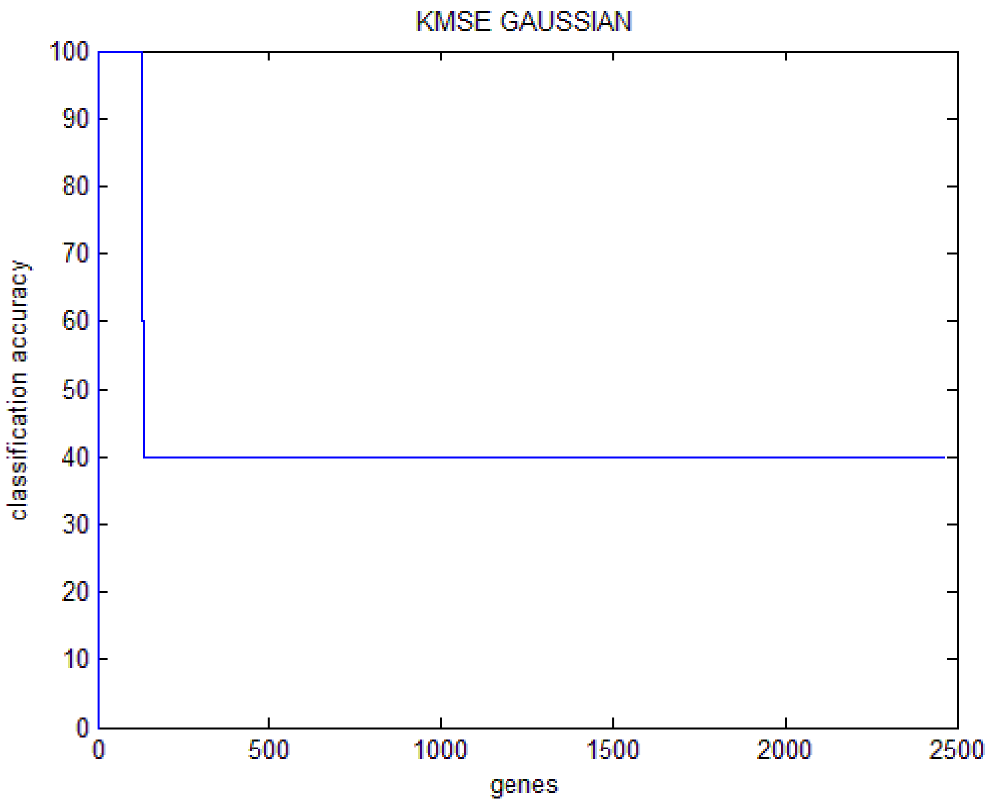
4th FOLD

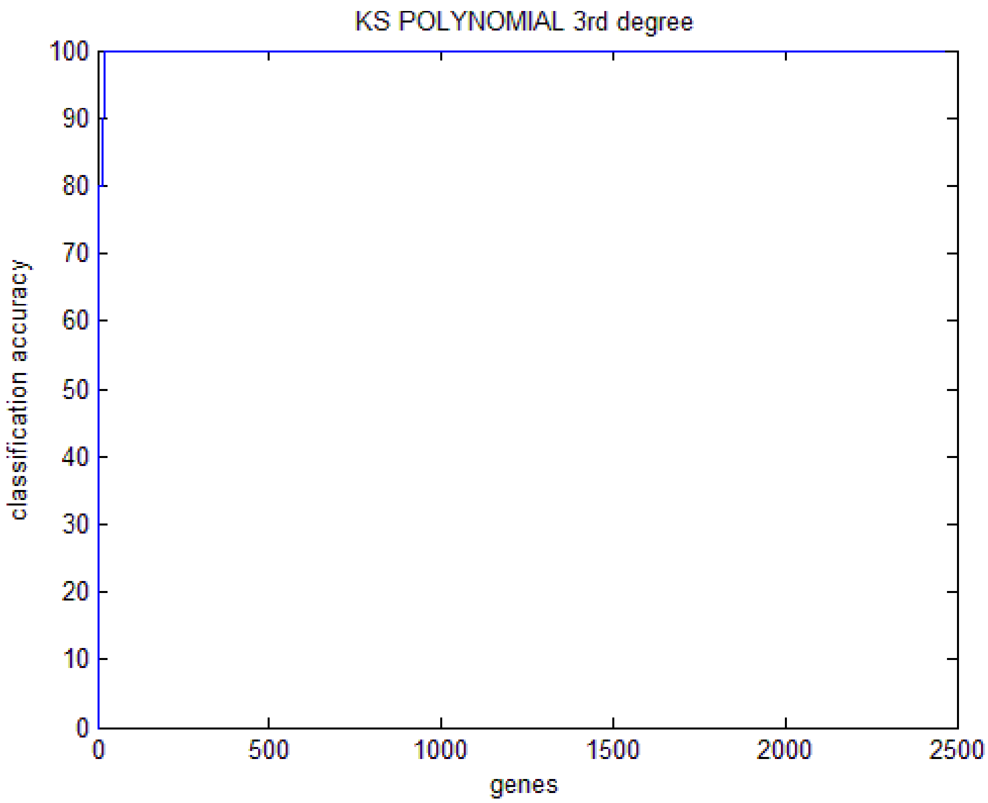


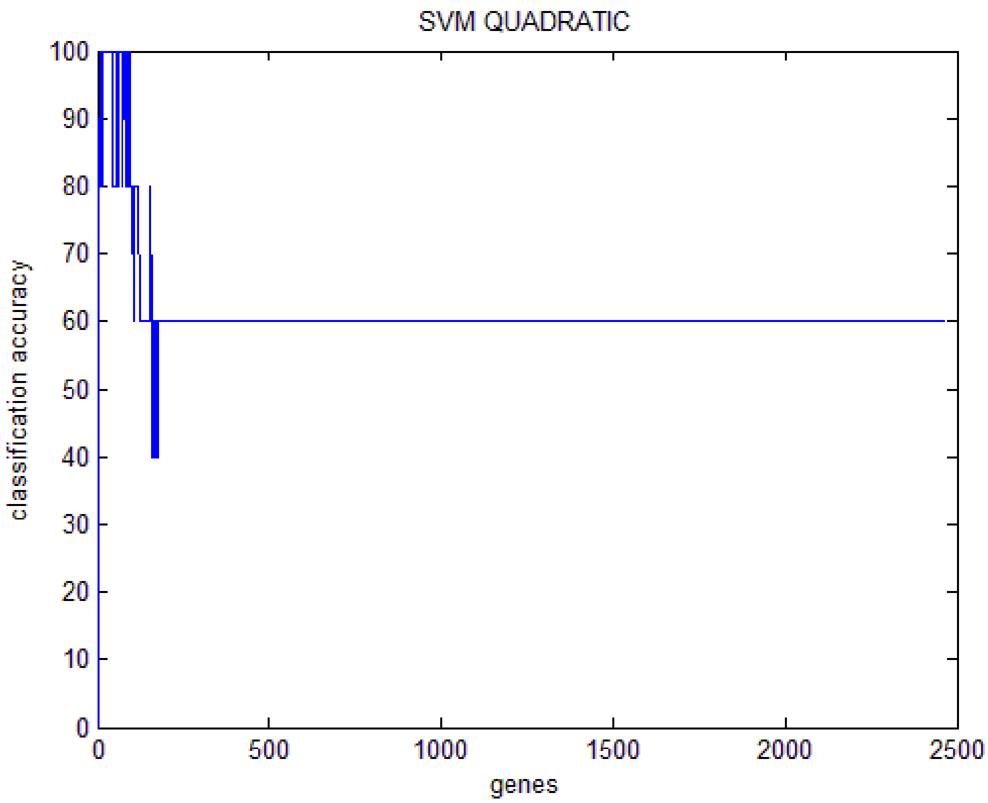
Mixture of Experts based on a fuzzy formulation
classification accuracy
genes
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500
2000
2500



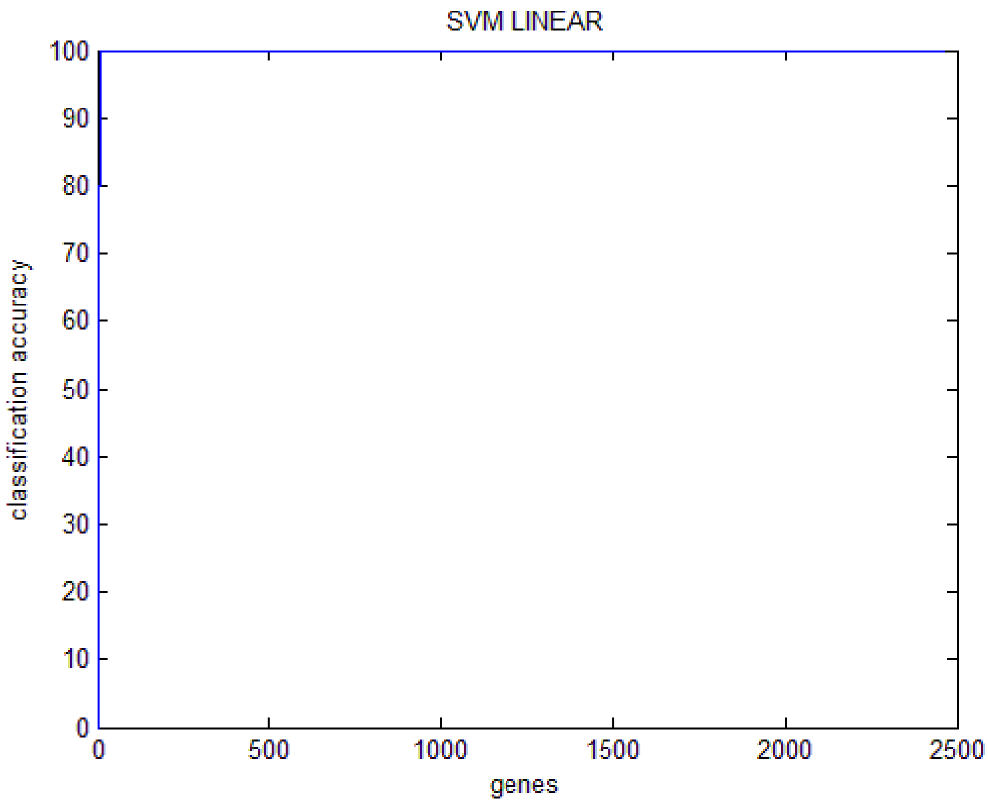
KNC POLYNOMIAL 2nd degree
classification accuracy
genes
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500
2000
2500


KMSE GAUSSIAN
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500
2000
2500
classification accuracy
genes


KS POLYNOMIAL 3rd degree
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500
2000
2500
classification accuracy
genes

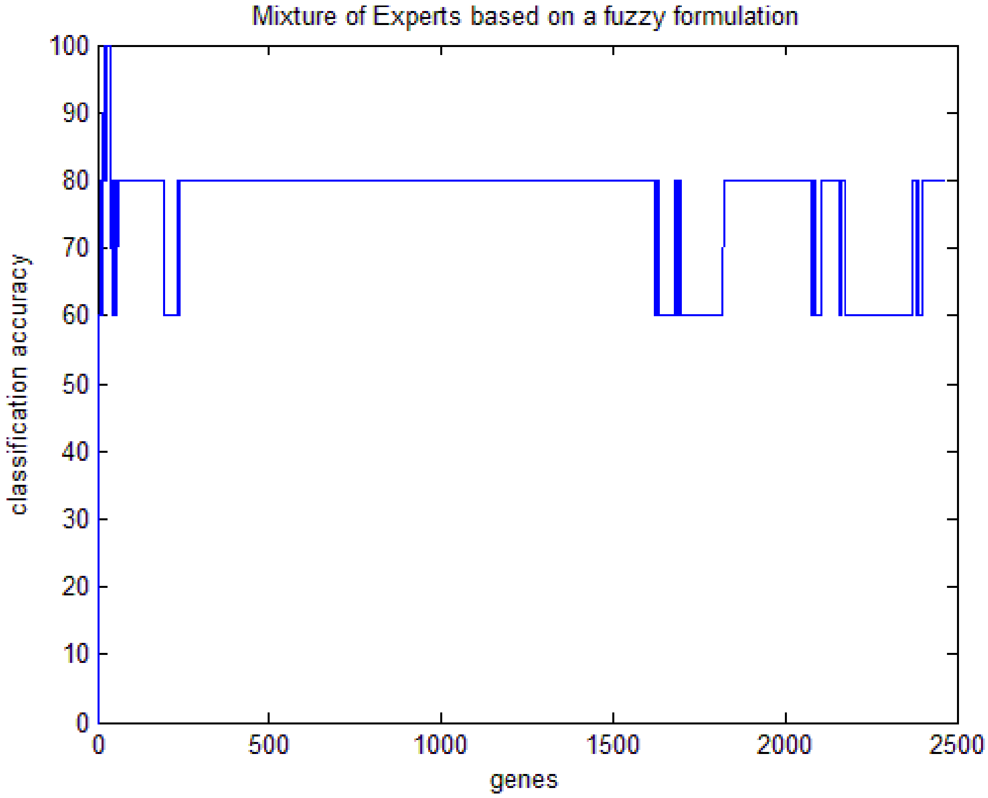


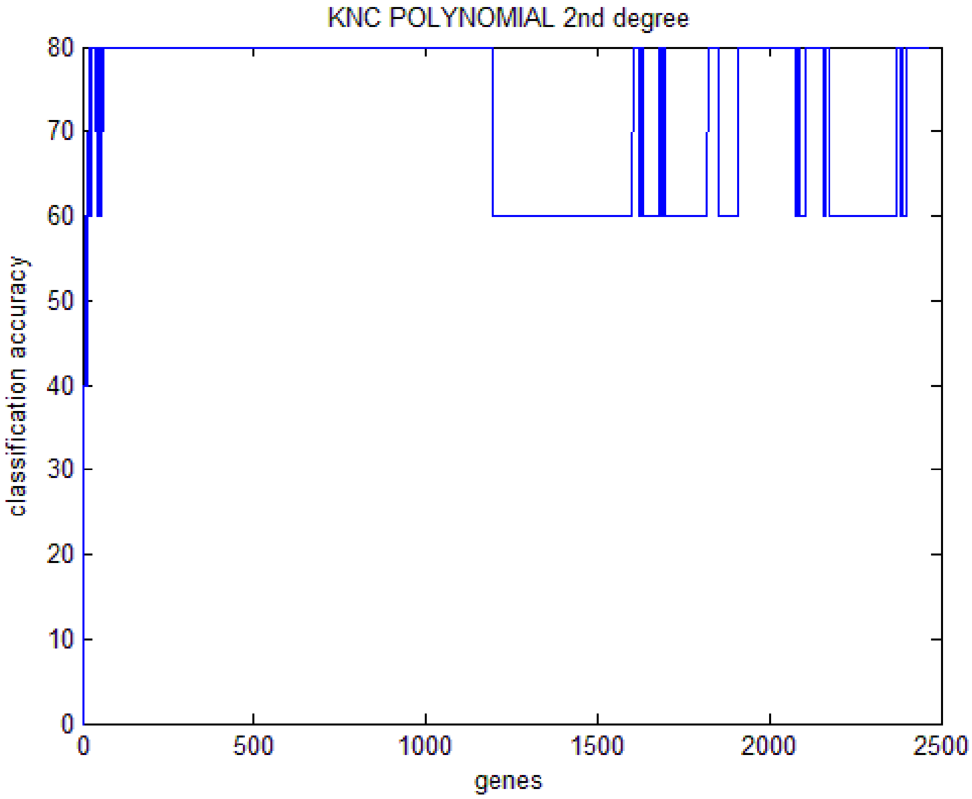
SVM QUADRATIC
classification accuracy
genes

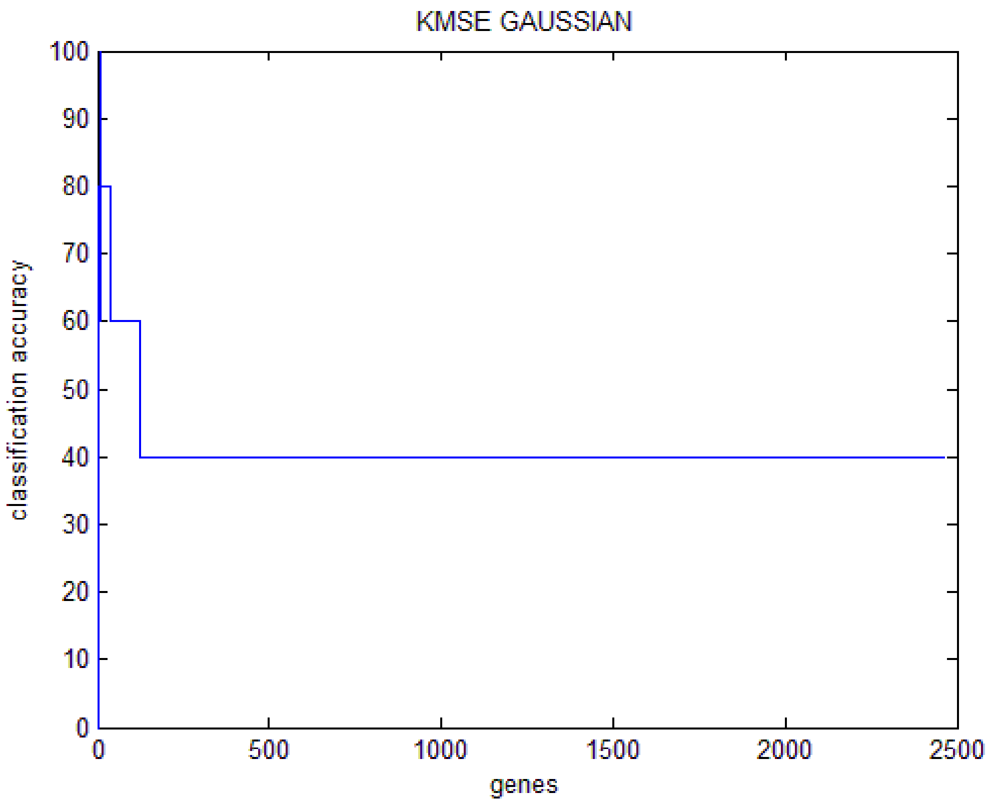


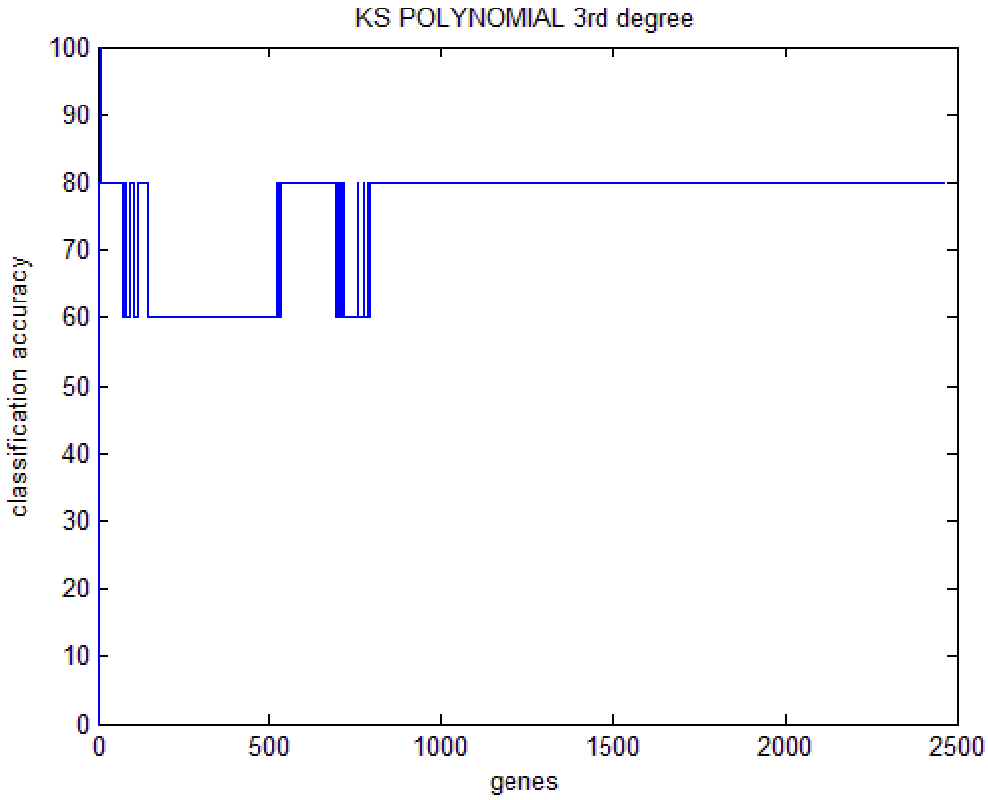
SVM LINEAR
classification accuracy
genes

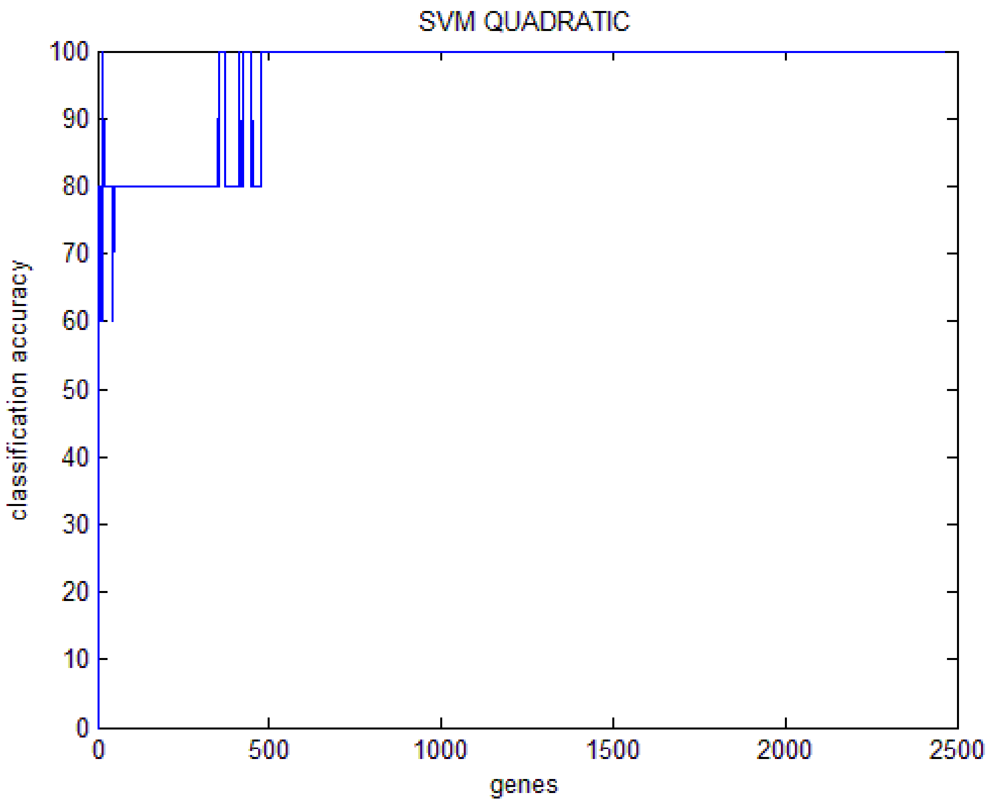
5th FOLD

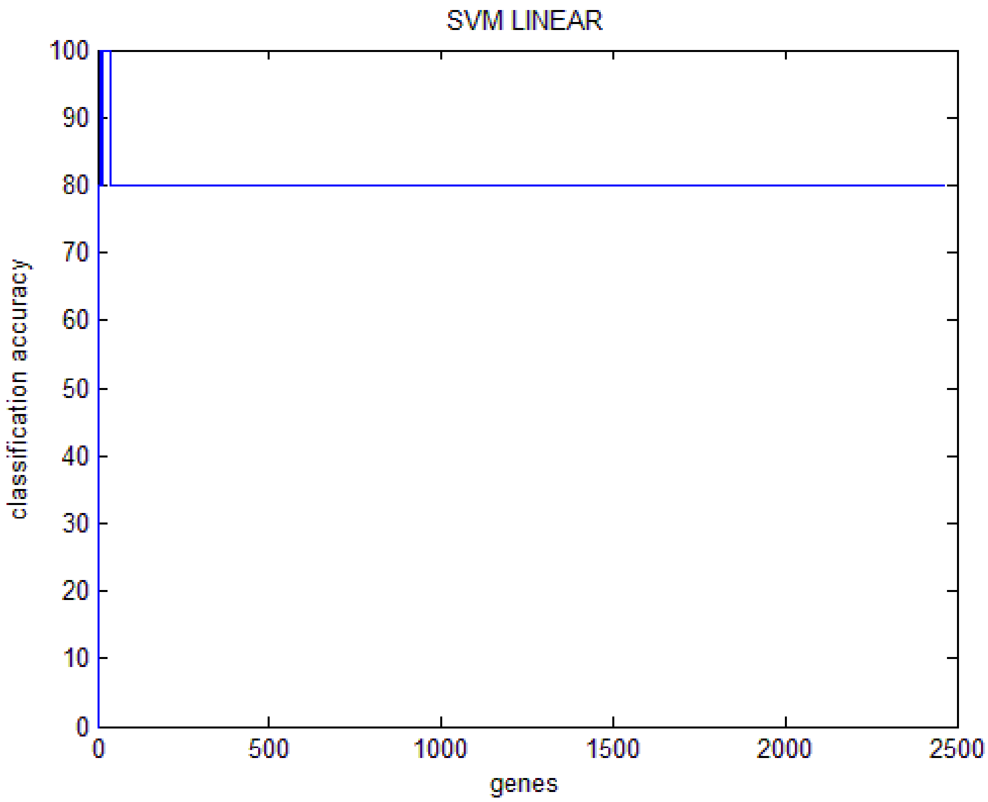

Mixture of Experts based on a fuzzy formulation
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
0
500
1000
1500
2000
2500


KNC POLYNOMIAL 2nd degree
classification accuracy
genes
0
10
20
30
40
50
60
70
80
0
500
1000
1500
2000
2500


KMSE GAUSSIAN
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
0
500
1000
1500
2000
2500


KS POLYNOMIAL 3rd degree
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
0
500
1000
1500
2000
2500


SVM QUADRATIC
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
0
500
1000
1500
2000
2500


SVM LINEAR
classification accuracy
genes
0
10
20
30
40
50
60
70
80
90
100
0
500
1000
1500
2000
2500

Παράρτημα Β

Ακολουθούν τα ονόματα και τα σύμβολα των γονιδίων που αποτελούν τις γονιδιακές υπογραφές για όλα τα μοντέλα ταξινόμησης της παρούσας εργασίας, και που περιλαμβάνονται στους πίνακες 23 και 24 της ενότητας 5.1. Επίσης, στον τελευταίο πίνακα σημειώνονται με ίδιο χρώμα τα κοινά γονίδια όλων των υπογραφών.

Ατομικός Ταξινομητής	Κωδικός γονιδίου	ORF	Σύμβολο γονιδίου	Περιγραφή
SVM LINEAR	104	YML098W	TAF19	TRANSCRIPTION TFIID 19 KD SUBUNIT
	475	YFL029C	CAK1	CELL CYCLE PROTEIN KINASE
	1066	YOL148C	SPT20	CHROMATIN STRUCTURE HISTONE ACETYLTRANSFERASE COMPLEX SUBUNIT
	1213	YJL008C	CCT8	PROTEIN FOLDING CYTOPLASMIC CHAPERONIN COMPLEX
	1262	YNR052C	POP2	GLUCOSE DEREPRESSION COMPONENT OF CCR4 COMPLEX
	1283	YDR228C	PCF11	MRNA 3'-END PROCESSING CLEAVAGE/POLYADENYLATION FACTOR CF I COMPONENT
	1307	YGL058W	RAD6	PROTEIN DEGRADATION, UBI E2 UB.-CONJUGATING ENZYME
	1723	YJL168C	SET2	GALACTOSE REGULATION TRANSCRIPTIONAL REPRESSOR OF GAL4
	2116	YPL149W	APG5	AUTOPHAGY UNKNOWN
	2364	YLR163C	MAS1	PROTEIN PROCESSING MITOCHONDRIAL PROCESSING PROTEASE SUBUNIT
	2466	YPR198W	SGE1	CRYSTAL VIOLET RESISTANC TRANSPORTER, MAJOR FACILITATOR SUPERFAMILY
SVM QDC	104	YML098W	TAF19	TRANSCRIPTION TFIID 19 KD SUBUNIT
	475	YFL029C	CAK1	CELL CYCLE PROTEIN KINASE
	934	YBR176W	ECM31	CELL WALL BIOGENESIS UNKNOWN
	1066	YOL148C	SPT20	CHROMATIN STRUCTURE HISTONE ACETYLTRANSFERASE COMPLEX SUBUNIT
	1213	YJL008C	CCT8	PROTEIN FOLDING CYTOPLASMIC CHAPERONIN COMPLEX
	1262	YNR052C	POP2	GLUCOSE DEREPRESSION COMPONENT OF CCR4 COMPLEX
	1283	YDR228C	PCF11	MRNA 3'-END PROCESSING CLEAVAGE/POLYADENYLATION FACTOR CF I COMPONENT
	1307	YGL058W	RAD6	PROTEIN DEGRADATION, UBI E2 UB.-CONJUGATING ENZYME
	1723	YJL168C	SET2	GALACTOSE REGULATION TRANSCRIPTIONAL REPRESSOR OF GAL4
	1819	YBR289W	SNF5	TRANSCRIPTION COMPONENT OF SWI/SNF GLOBAL ACTIVATOR COMPLEX
	2116	YPL149W	APG5	AUTOPHAGY UNKNOWN
	2364	YLR163C	MAS1	PROTEIN PROCESSING MITOCHONDRIAL PROCESSING PROTEASE SUBUNIT
	2466	YPR198W	SGE1	CRYSTAL VIOLET RESISTANC TRANSPORTER, MAJOR FACILITATOR SUPERFAMILY
KS POLYNOMIAL 3	7	YHL025W	SNF6	TRANSCRIPTION COMPONENT OF SWI/SNF GLOBAL ACTIVATOR COMPLEX
	30	YML051W	GAL80	GALACTOSE REGULATION TRANSCRIPTIONAL REPRESSOR
	104	YML098W	TAF19	TRANSCRIPTION TFIID 19 KD SUBUNIT
	475	YFL029C	CAK1	CELL CYCLE PROTEIN KINASE
	1066	YOL148C	SPT20	CHROMATIN STRUCTURE HISTONE ACETYLTRANSFERASE COMPLEX SUBUNIT
	1213	YJL008C	CCT8	PROTEIN FOLDING CYTOPLASMIC CHAPERONIN COMPLEX
	1262	YNR052C	POP2	GLUCOSE DEREPRESSION COMPONENT OF CCR4 COMPLEX
	1283	YDR228C	PCF11	MRNA 3'-END PROCESSING CLEAVAGE/POLYADENYLATION FACTOR CF I COMPONENT
	1307	YGL058W	RAD6	PROTEIN DEGRADATION, UBI E2 UB.-CONJUGATING ENZYME
	1723	YJL168C	SET2	GALACTOSE REGULATION TRANSCRIPTIONAL REPRESSOR OF GAL4
	2116	YPL149W	APG5	AUTOPHAGY UNKNOWN
	2364	YLR163C	MAS1	PROTEIN PROCESSING MITOCHONDRIAL PROCESSING PROTEASE SUBUNIT
	2466	YPR198W	SGE1	CRYSTAL VIOLET RESISTANC TRANSPORTER, MAJOR FACILITATOR SUPERFAMILY
KMSE GAUSSIAN	7	YHL025W	SNF6	TRANSCRIPTION COMPONENT OF SWI/SNF GLOBAL ACTIVATOR COMPLEX
	30	YML051W	GAL80	GALACTOSE REGULATION TRANSCRIPTIONAL REPRESSOR
	104	YML098W	TAF19	TRANSCRIPTION TFIID 19 KD SUBUNIT
	475	YFL029C	CAK1	CELL CYCLE PROTEIN KINASE
	1066	YOL148C	SPT20	CHROMATIN STRUCTURE HISTONE ACETYLTRANSFERASE COMPLEX SUBUNIT
	1213	YJL008C	CCT8	PROTEIN FOLDING CYTOPLASMIC CHAPERONIN COMPLEX
	1262	YNR052C	POP2	GLUCOSE DEREPRESSION COMPONENT OF CCR4 COMPLEX
	1283	YDR228C	PCF11	MRNA 3'-END PROCESSING CLEAVAGE/POLYADENYLATION FACTOR CF I COMPONENT
	1307	YGL058W	RAD6	PROTEIN DEGRADATION, UBI E2 UB.-CONJUGATING ENZYME
	1723	YJL168C	SET2	GALACTOSE REGULATION TRANSCRIPTIONAL REPRESSOR OF GAL4
	2116	YPL149W	APG5	AUTOPHAGY UNKNOWN
	2364	YLR163C	MAS1	PROTEIN PROCESSING MITOCHONDRIAL PROCESSING PROTEASE SUBUNIT
	2466	YPR198W	SGE1	CRYSTAL VIOLET RESISTANC TRANSPORTER, MAJOR FACILITATOR SUPERFAMILY

Ατομικός Ταξινομητής	Κωδικός γονιδίου	ORF	Σύμβολο γονιδίου	Περιγραφή
KNC POLYNOMIAL 2	7	YHL025W	SNF6	TRANSCRIPTION COMPONENT OF SWI/SNF GLOBAL ACTIVATOR COMPLEX
	104	YML098W	TAF19	TRANSCRIPTION TFIID 19 KD SUBUNIT
	136	YIL139C	REV7	DNA REPAIR DNA POLYMERASE ZETA
	144	YCR002C	CDC10	CYTOKINESIS GTP BINDING PROTEIN
	231	YJL085W	EXO70	SECRETION EXOCYST COMPLEX SUBUNIT
	235	YDR323C	PEP7	VACUOLAR PROTEIN TARGETI VACUOLAR SEGREGATION PROTEIN
	307	YJL219W	HXT9	TRANSPORT HEXOSE PERMEASE
	327	YNL012W	SPO1	MEIOSIS (SPOR.) TRANSCRIPTIONAL REGULATOR
	396	YOR117W	RPT5	PROTEIN DEGRADATION 26S PROTEASOME REGULATORY SUBUNIT
	475	YFL029C	CAK1	CELL CYCLE PROTEIN KINASE
	556	YNR030W	ECM39	CELL WALL BIOGENESIS UNKNOWN
	576	YGL038C	OCH1	PROTEIN GLYCOSYLATION MEMBRANE-BOUND MANNOSYLTRANSFERASE
	584	YMR236W	TAF17	TRANSCRIPTION TFIID 17 KD SUBUNIT
	616	YHR044C	DOG1	2-DEOXYGLUCOSE RESISTANC 2-DEOXYGLUCOSE-6-PHOSPHATE PHOSPHATASE
	624	YOR002W	ALG6	PROTEIN GLYCOSYLATION GLUCOSYLTRANSFERASE
	644	YNL052W	COX5A	OXIDATIVE PHOSPHORYLATIO CYTOCHROME-C OXIDASE SUBUNIT VA
	833	YOR069W	VPS5	VACUOLAR PROTEIN TARGETI SORTING NEXIN FAMILY
	934	YBR176W	ECM31	CELL WALL BIOGENESIS UNKNOWN
	938	YIL022W	TIM44	MITOCHONDRIAL PROTEIN TA INNER MEMBRANE TRANSLOCASE COMPONENT
	952	YGR171C	MSM1	PROTEIN SYNTHESIS MITOCHONDRIAL METHIONYL-TRNA SYNTHETASE
	1065	YGL134W	PCL10	CELL CYCLE CYCLIN (PHO85P)
	1074	YEL066W	HPA3	CHROMATIN STRUCTURE HISTONE ACETYLTRANSFERASE COMPLEX SUBUNIT
	1092	YDL045C	FAD1	FLAVIN BIOSYNTHESIS FLAVIN ADENINE DINUCLEOTIDE (FAD) SYNTHETASE
	1126	YML055W	SPC2	SECRETION SIGNAL PEPTIDASE SUBUNIT
	1138	YHR012W	VPS29	VACUOLAR PROTEIN TARGETI TARGETS VACUOLAR RECEPTOR, VPS10
	1213	YJL008C	CCT8	PROTEIN FOLDING CYTOPLASMIC CHAPERONIN COMPLEX
	1238	YLR417W	VPS36	VACUOLAR PROTEIN TARGETI UNKNOWN?
	1250	YNL048W	ALG11	PROTEIN GLYCOSYLATION UNKNOWN
	1260	YPL161C	BEM4	BUD EMERGENCE INTERACTS WITH RHO-TYPE GTPASES
	1266	YBL040C	ERD2	ER PROTEIN RETENTION HDEL RECEPTOR
	1414	YEL009C	GCN4	AMINO ACID, PURINE BIOSY TRANSCRIPTION FACTOR
	1819	YBR289W	SNF5	TRANSCRIPTION COMPONENT OF SWI/SNF GLOBAL ACTIVATOR COMPLEX
	1916	YFR002W	NIC96	NUCLEAR PROTEIN TARGETIN NUCLEAR PORE PROTEIN
	1992	YOR005C	DNL4	DNA REPAIR DNA LIGASE IV HOMOLOGUE
	2062	YBR131W	CCZ1	CALCIUM, CAFFEINE, AND Z UNKNOWN
	2101	YMR274C	RCE1	PROTEIN PROCESSING PROTEASE, ACTS ON RAS AND A-FACTOR C-TERMINI
	2116	YPL149W	APG5	AUTOPHAGY UNKNOWN
	2167	YGR186W	TFG1	TRANSCRIPTION TFIIF 105 KD SUBUNIT
	2231	YPR056W	TFB4	TRANSCRIPTION TFIIH 37 KD SUBUNIT
	2338	YBR003W	COQ1	UBIQUINONE BIOSYNTHESIS EXAPRENYL PYROPHOSPHATE SYNTHETASE
	2359	YGR183C	QCR9	OXIDATIVE PHOSPHORYLATIO UBIQUINOL CYTOCHROME-C REDUCTASE SUBUNIT 9
	2364	YLR163C	MAS1	PROTEIN PROCESSING MITOCHONDRIAL PROCESSING PROTEASE SUBUNIT
	2427	YFR033C	QCR6	OXIDATIVE PHOSPHORYLATIO UBIQUINOL CYTOCHROME-C REDUCTASE SUBUNIT
	2466	YPR198W	SGE1	CRYSTAL VIOLET RESISTANC TRANSPORTER, MAJOR FACILITATOR SUPERFAMILY

Συνδυασμός Ατομικών Ταξινομητών	Κωδικός γονιδίου	ORF	Σύμβολο γονιδίου	Περιγραφή
MoE/fuzzy	30	YML051W	GAL80	GALACTOSE REGULATION TRANSCRIPTIONAL REPRESSOR
	120	YKR002W	PAP1	MRNA POLYADENYLATION POLY(A) POLYMERASE
	475	YFL029C	CAK1	CELL CYCLE PROTEIN KINASE
	1066	YOL148C	SPT20	CHROMATIN STRUCTURE HISTONE ACETYLTRANSFERASE COMPLEX SUBUNIT
	1213	YJL008C	CCT8	PROTEIN FOLDING CYTOPLASMIC CHAPERONIN COMPLEX
	1262	YNR052C	POP2	GLUCOSE DEREPRESSION COMPONENT OF CCR4 COMPLEX
	1307	YGL058W	RAD6	PROTEIN DEGRADATION, UBI E2 UB.-CONJUGATING ENZYME
	2116	YPL149W	APG5	AUTOPHAGY UNKNOWN
	2466	YPR198W	SGE1	CRYSTAL VIOLET RESISTANC TRANSPORTER, MAJOR FACILITATOR SUPERFAMILY
MoE/som	30	YML051W	GAL80	GALACTOSE REGULATION TRANSCRIPTIONAL REPRESSOR
	104	YML098W	TAF19	TRANSCRIPTION TFIID 19 KD SUBUNIT
	136	YIL139C	REV7	DNA REPAIR DNA POLYMERASE ZETA
	475	YFL029C	CAK1	CELL CYCLE PROTEIN KINASE
	1066	YOL148C	SPT20	CHROMATIN STRUCTURE HISTONE ACETYLTRANSFERASE COMPLEX SUBUNIT
	1213	YJL008C	CCT8	PROTEIN FOLDING CYTOPLASMIC CHAPERONIN COMPLEX
	1262	YNR052C	POP2	GLUCOSE DEREPRESSION COMPONENT OF CCR4 COMPLEX
	1283	YDR228C	PCF11	MRNA 3'-END PROCESSING CLEAVAGE/POLYADENYLATION FACTOR CF I COMPONENT
	1307	YGL058W	RAD6	PROTEIN DEGRADATION, UBI E2 UB.-CONJUGATING ENZYME
	1723	YJL168C	SET2	GALACTOSE REGULATION TRANSCRIPTIONAL REPRESSOR OF GAL4
	2116	YPL149W	APG5	AUTOPHAGY UNKNOWN
	2364	YLR163C	MAS1	PROTEIN PROCESSING MITOCHONDRIAL PROCESSING PROTEASE SUBUNIT
	2466	YPR198W	SGE1	CRYSTAL VIOLET RESISTANC TRANSPORTER, MAJOR FACILITATOR SUPERFAMILY
MoE/Gradient	7	YHL025W	SNF6	TRANSCRIPTION COMPONENT OF SWI/SNF GLOBAL ACTIVATOR COMPLEX
	30	YML051W	GAL80	GALACTOSE REGULATION TRANSCRIPTIONAL REPRESSOR
	104	YML098W	TAF19	TRANSCRIPTION TFIID 19 KD SUBUNIT
	475	YFL029C	CAK1	CELL CYCLE PROTEIN KINASE
	1066	YOL148C	SPT20	CHROMATIN STRUCTURE HISTONE ACETYLTRANSFERASE COMPLEX SUBUNIT
	1213	YJL008C	CCT8	PROTEIN FOLDING CYTOPLASMIC CHAPERONIN COMPLEX
	1262	YNR052C	POP2	GLUCOSE DEREPRESSION COMPONENT OF CCR4 COMPLEX
	1283	YDR228C	PCF11	MRNA 3'-END PROCESSING CLEAVAGE/POLYADENYLATION FACTOR CF I COMPONENT
	1307	YGL058W	RAD6	PROTEIN DEGRADATION, UBI E2 UB.-CONJUGATING ENZYME
	1723	YJL168C	SET2	GALACTOSE REGULATION TRANSCRIPTIONAL REPRESSOR OF GAL4
	2116	YPL149W	APG5	AUTOPHAGY UNKNOWN
	2364	YLR163C	MAS1	PROTEIN PROCESSING MITOCHONDRIAL PROCESSING PROTEASE SUBUNIT
	2466	YPR198W	SGE1	CRYSTAL VIOLET RESISTANC TRANSPORTER, MAJOR FACILITATOR SUPERFAMILY
MoE/2 LOGSIGS	104	YML098W	TAF19	TRANSCRIPTION TFIID 19 KD SUBUNIT
	475	YFL029C	CAK1	CELL CYCLE PROTEIN KINASE
	625	YDL066W	IDP1	TCA CYCLE ISOCITRATE DEHYDROGENASE (NADP+)
	934	YBR176W	ECM31	CELL WALL BIOGENESIS UNKNOWN
	1066	YOL148C	SPT20	CHROMATIN STRUCTURE HISTONE ACETYLTRANSFERASE COMPLEX SUBUNIT
	1213	YJL008C	CCT8	PROTEIN FOLDING CYTOPLASMIC CHAPERONIN COMPLEX
	1262	YNR052C	POP2	GLUCOSE DEREPRESSION COMPONENT OF CCR4 COMPLEX
	1283	YDR228C	PCF11	MRNA 3'-END PROCESSING CLEAVAGE/POLYADENYLATION FACTOR CF I COMPONENT
	1307	YGL058W	RAD6	PROTEIN DEGRADATION, UBI E2 UB.-CONJUGATING ENZYME
	1723	YJL168C	SET2	GALACTOSE REGULATION TRANSCRIPTIONAL REPRESSOR OF GAL4
	2116	YPL149W	APG5	AUTOPHAGY UNKNOWN
	2364	YLR163C	MAS1	PROTEIN PROCESSING MITOCHONDRIAL PROCESSING PROTEASE SUBUNIT
	2466	YPR198W	SGE1	CRYSTAL VIOLET RESISTANC TRANSPORTER, MAJOR FACILITATOR SUPERFAMILY

ΟΜΟΙΟΤΗΤΕΣ ΓΟΝΙΔΙΩΝ ΜΕΤΑΞΥ ΟΛΩΝ ΤΩΝ ΑΤΟΜΙΚΩΝ ΤΑΞΙΝΟΜΗΤΩΝ ΚΑΙ ΟΛΩΝ ΤΩΝ ΣΥΝΔΥΑΣΜΩΝ ΤΩΝ ΑΤΟΜΙΚΩΝ ΤΑΞΙΝΟΜΗΤΩΝ								
SVM LINEAR	SVM QUADRATIC	KS POLYNOMIAL 3	KMSE GAUSSIAN	KNC POLYNOMIAL 2	MoE (FUZZY SET FORMULATION)	MoE (SOM FORMULATION)	MoE (GRADIENT)	MoE (2 LOGSIGS)
Κωδικός γονιδίου	Κωδικός γονιδίου	Κωδικός γονιδίου	Κωδικός γονιδίου	Κωδικός γονιδίου	Κωδικός γονιδίου	Κωδικός γονιδίου	Κωδικός γονιδίου	Κωδικός γονιδίου
104	104	7	7	7	30	30	7	104
475	475	30	30	104	120	104	30	475
1066	934	104	104	136	475	136	104	625
1213	1066	475	475	144	1066	475	475	934
1262	1213	1066	1066	231	1213	1066	1066	1066
1283	1262	1213	1213	235	1262	1213	1213	1213
1307	1283	1262	1262	307	1307	1262	1262	1262
1723	1307	1283	1283	327	2116	1283	1283	1283
2116	1723	1307	1307	396	2466	1307	1307	1307
2364	1819	1723	1723	475		1723	1723	1723
2466	2116	2116	2116	556		2116	2116	2116
	2364	2364	2364	576		2364	2364	2364
	2466	2466	2466	584		2466	2466	2466
				616				
				624				
				644				
				833				
				934				
				938				
				952				
				1065				
				1074				
				1092				
				1126				
				1138				
				1213				
				1238				
				1250				
				1260				
				1266				
				1414				
				1819				
				1916				
				1992				
				2062				
				2101				
				2116				
				2167				
				2231				
				2338				
				2359				
				2364				
				2427				
				2466				