
# ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

## «ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΤΩΝ ΒΕΛΤΙΣΤΩΝ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ KERNEL RIDGE REGRESSION ΓΙΑ ΤΑΞΙΝΟΜΗΣΗ ΓΟΝΙΔΙΑΚΩΝ ΔΕΔΟΜΕΝΩΝ»

**Μούκα Θεοδώρα**

**2000030040**

**Εξεταστική Επιτροπή**

Ζερβάκης Μιχάλης (επιβλέπων)

Μπάλας Κωνσταντίνος

Σταυρακάκης Γεώργιος

## *Ευχαριστίες*

*Αρχικά, θέλω να ευχαριστήσω τον καθηγητή μου, κύριο Ζερβάκη για την βοήθεια του στην εκπλήρωση αυτής της διπλωματικής εργασίας. Ευχαριστώ ακόμα πολύ τον κύριο Μιχάλη Μπλαζαντωνάκη για τη διαρκή του υποστήριξη-ακόμα και από μακριά- σε όλη την προσπάθεια υλοποίησης της εργασίας. Επίσης, ευχαριστώ τα μέλη της εξεταστικής επιτροπής Κύριο Μπάλα και κύριο Σταυρακάκη για την παρουσία και την ενασχόλησή τους με το κείμενο της διπλωματικής εργασίας.*

*Ακόμα, θα ήθελα να ευχαριστήσω την οικογένειά μου που αν και τόσο μακριά κατάφεραν να ναι πάντα δίπλα μου. Πολλά ευχαριστώ στους φίλους μου στα Χανιά που βοήθησαν να περάσουν όμορφα όλα αυτά τα χρόνια. Ιδιαίτερα ευχαριστώ στις αγαπημένες συγκατοίκους Μαρία και Juliana,στον επίτιμο συγκάτοικο Νίκο ,στον Δημήτρη και τον Θεοδόση που ήταν πάντα πρόθυμοι για βοήθεια αλλά και βόλτες, στην Έφη ,στον Αλέξανδρο , στον Μάριο και τον Μιχάλη ,στον Νίκο, στον Θωμά, στον Τάσο για όλες του τις συμβουλές, στον πάντα ενθουσιώδη Πέτρο, στον Βασίλη ,στον Θοδωρή ,τον Σπύρο, την Georgina, τον Θοδωρή ,τον Χρήστο, τον Κώστα και τον γεμάτο κατανόηση Θανάση ,τον ξενιτεμένο Κώστα και τον Διονύση.Ευχαριστώ πολύ τη Σουζάννα ,την Ευτυχία και την Αλίκη. Ευχαριστώ τον Γρηγόρη, τον Σταύρο και την Μαμασά για τις γεμάτες ενδιαφέρον και ιστορίες βόλτες και τις μικρές μου φίλες Μαρία και Αθηνά. Ακόμα ευχαριστώ τον Κωστάκη που γέμισε κάποια από αυτά τα χρόνια με πολλές όμορφες μικρές περιπέτειες.*

*Δε θα μπορούσα να μην ευχαριστήσω ιδιαίτερα και όλους εκείνους τους συναδέλφους που με πολλούς τρόπους με καναν να αισθανθώ τι σημαίνει συλλογικότητα στην καθημερινότητα της σχολής μέχρι και την ολοκλήρωση αυτής της εργασίας. Η εργασία αφιερώνεται στο φοιτητικό σύλλογο του οποίου οι διαδικασίες μου δίδαξαν πολλά.*

# Περίληψη

Οι όποιες προσπάθειες για έγκαιρη διάγνωση του καρκίνου στη βάση της μορφολογίας των καρκινικών όγκων έχουν σε πολύ μεγάλο βαθμό αποτύχει .Έτσι, υπάρχει αυξημένο επιστημονικό ενδιαφέρον στην προσπάθεια ταξινόμησης του όγκου με βάση μοριακά στοιχεία. Η γονιδιακή έκφραση των όγκων μπορεί να προσφέρει πολύ περισσότερη πληροφορία από τη μορφολογική και να προωθήσει την πιο αξιόπιστη ταξινόμηση των όγκων.

Εξαιτίας του πολύ μεγάλου όγκου δεδομένων, δημιουργούνται πολύ σημαντικά προβλήματα στην απόδοση και το κόστος της ταξινόμησης. Το πρόβλημα αυτό μπορεί να επιλυθεί με την επιλογή των πιο σημαντικών γονιδίων (informative genes), των οποίων ο συνδυασμός μπορεί να συγκεντρώσει τη συνολική πληροφορία του γονιδιώματος.

Υπάρχουν στη βιβλιογραφία πολλές μέθοδοι επιλογής γονιδίων , χωρίς καμία από αυτές να οδηγεί σε αξιόπιστα αποτελέσματα για την επιλογή ενός ελάχιστου συνόλου γονιδίων στα οποία θα επικεντρωθεί η μελέτη της βιολογίας, για την πρόβλεψη τυχόν καρκινικής προδιάθεσης. **Σκοπός της εργασίας είναι η υλοποίηση και βελτιστοποίηση του αλγορίθμου παλινδρόμησης κορυφής με χρήση της τεχνικής του πυρήνα (Kernel Ridge Regression-KRR) για την ταξινόμηση και επιλογή γονιδιακών δεδομένων.**

# Περιεχόμενα

# περιεχόμενα εικόνων- διαγραμμάτων- πινάκων

- **Εικόνες**



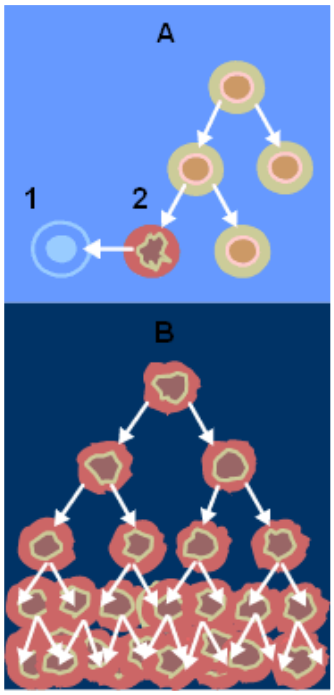
- **Πίνακες**

- **Διαγράμματα**

# Κεφάλαιο1: Εισαγωγή

## *1.1 Εισαγωγή – Γενετική έρευνα για τον καρκίνο*

Με τον όρο «καρκίνος» περιγράφεται μία ομάδα νοσημάτων, η αιτία των οποίων βρίσκεται σε κυτταρικό επίπεδο. **Ο όρος αναφέρεται στην υπερβολική, χωρίς προγραμματισμό, ανάπτυξη κυττάρων του οργανισμού, που ήταν φυσιολογικά, μέχρι τη στιγμή της έναρξης της διαδικασίας καρκινογένεσης.** Τα κύτταρα που έχουν προσβληθεί(καρκινικά κύτταρα)εμφανίζουν την δυνατότητα να εισβάλουν σε άλλους ιστούς παρακείμενους ή πιο απόμακρους μέσω μετάστασης. Αυτή η ανεξέλεγκτη αύξηση προκαλείται από βλάβη στο DNA, που οδηγεί σε μεταλλάξεις στα γονίδια που ελέγχουν την κυτταρική διαίρεση. Οι μεταλλάξεις που μετασχηματίζουν ένα υγιές κύτταρο σε κακοήθες είναι διαφόρων ειδών . Εάν δεν θεραπευθούν, οι καρκίνοι μπορούν τελικά να προκαλέσουν το θάνατο. Η επιβίωση των ασθενών με καρκίνο έχει βελτιωθεί σημαντικά τα τελευταία χρόνια. Ο καρκίνος ,ωστόσο, συνεχίζει να αποτελεί την δεύτερη αιτία θανάτου στις ανεπτυγμένες χώρες μετά τα καρδιαγγειακά νοσήματα με τάσεις αύξησης τόσο της συχνότητας εμφάνισής του όσο και της διάδοσής του. Κάθε χρόνο εμφανίζονται περίπου επτά εκατομμύρια νέα περιστατικά καρκίνου σε όλο τον κόσμο.




***Εικόνα 1**: σχηματική απεικόνιση δημιουργίας ενός καρκινικού κυττάρου:*
*A-κανονική διαίρεση κυττάρων,*
*B- καρκινική διαίρεση κυττάρων*
*2 - κατεστραμμένο κύτταρο.*

Σήμερα, ο καρκίνος μπορεί να αντιμετωπιστεί στις περισσότερες περιπτώσεις. Για να θεραπευτεί όμως, πρέπει η διάγνωση να γίνει σε πρώιμα στάδια της νόσου. Η διάγνωση, όμως, των διαφόρων ειδών του καρκίνου δεν είναι εύκολη, τουλάχιστο στα πρώτα στάδια.

Η ασυμπτωματική περίοδος μπορεί να διαρκέσει μήνες ή χρόνια. Υπάρχει λοιπόν ο κίνδυνος, η διάγνωση της νόσου να γίνει όταν ο όγκος έχει επεκταθεί τοπικά ή και να έχει χορηγήσει μεταστάσεις.

Οι όποιες προσπάθειες διάγνωσης αλλά και αξιολόγησης των όγκων στη βάση της μορφολογίας έχουν σε πολύ μεγάλο βαθμό αποτύχει, εξαιτίας των διαγνωστικών ασυμφωνιών που προκύπτουν από την μεταξύ και ενδιάμεσα αδυναμία αναπαραγωγής των παρατηρήσεων.

Επιπλέον ,έχει παρατηρηθεί ότι οι ασθενείς που υποφέρουν από καρκίνο του μαστού και βρίσκονται στην ίδια ακριβώς φάση της ασθένειας, μπορούν να έχουν αξιοσημείωτα διαφορετική απόκριση στην θεραπεία που υποβάλλονται. Οι πιο ισχυροί μέχρι τώρα παράγοντες που μπορούν να προβλέψουν μεταστάσεις (predictors), δεν μπορούν να ταξινομήσουν τους όγκους του μαστού σύμφωνα με την κλινική τους συμπεριφορά. Η χημειοθεραπεία ή η ορμονική θεραπεία μειώνουν τον κίνδυνο άμεσης μετάστασης κατά 30% περίπου, αλλά το 70%- 80% των ασθενών που υπόκεινται σε αυτή τη θεραπεία θα μπορούσαν να είχαν επιβιώσει και χωρίς αυτήν.

Για αυτόν, ακριβώς το λόγο, υπάρχει αυξημένο επιστημονικό ενδιαφέρον στην προσπάθεια ταξινόμησης (classification) του όγκου με βάση όχι μορφολογικά στοιχεία, αλλά μοριακά. Η γονιδιακή έκφραση των όγκων μπορεί να προσφέρει πολύ περισσότερη πληροφορία από τη μορφολογική και να προωθήσει εναλλακτικά σε αυτή, πιο αξιόπιστα συστήματα ταξινόμησης των όγκων. Η ανάπτυξη της microarray τεχνολογίας (κεφ. 1.2) και η συνακολούθα αναγκαία ανάπτυξη και εξέλιξη μεθόδων ταξινόμησης για μεγάλο όγκο δεδομένων(κεφ. 1.3) ανοίγουν προοπτικές στην παραπάνω κατεύθυνση.

## *1.2 Microarray Τεχνολογία*

Η microarray τεχνολογία είναι μια πολλά υποσχόμενη προσέγγιση για την ανάλυση μεγάλου μεγέθους δεδομένων και μας δίνει τη δυνατότητα να μελετήσουμε τα πρότυπα της γονιδιακής έκφρασης σε μια γενετική βαθμίδα. Με τη βοήθεια των microarrays μπορούμε να καθορίσουμε χιλιάδες τιμές

έκφρασης σε εκατοντάδες διαφορετικές περιπτώσεις, επιτρέποντας τη συγκέντρωση των γενετικών διεργασιών σε μια συνολική γενετική βαθμίδα. Με αυτόν τον τρόπο μπορούμε να καθορίσουμε τη συνεισφορά του γενετικού υλικού σε πολύπλοκές ανωμαλίες και να εξετάσουμε τις αλλαγές που μπορεί να υπάρχουν σε γονίδια (διαφορετικά επίπεδα έκφρασης) κατά την εκδήλωση ασθενειών.

Συνοπτικά, τα microarrays είναι:

- Ένα πείραμα της τάξης των 10000 στοιχείων
- Ένας τρόπος να εξερευνήσουμε τις λειτουργίες των γονιδίων
- Ένα στιγμιότυπο του επίπεδο έκφρασης ενός ολόκληρου φαινοτύπου κάτω από συγκεκριμένες συνθήκες.

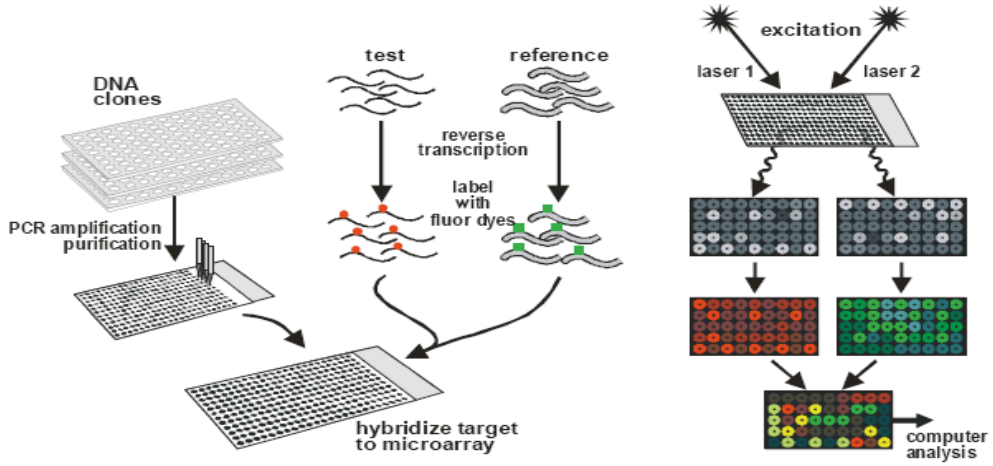
Η ουσία και οι λειτουργίες των κυττάρων χαρακτηρίζονται από την παραγωγή των πρωτεϊνών, οι οποίες αποτελούνται από αμινοξέα. Τα πρότυπα των αλληλουχιών των αμινοξέων κωδικοποιούνται σε γονίδια, τα οποία αποτελούν μέρος του γονιδιώματος. Για την παραγωγή της πρωτεΐνης, η οποία ορίζεται από το γονίδιο, ένα κύτταρο μεταγράφει τη γενετική του αλληλουχία σε αλληλουχία αγγελιοφόρου ριβονουκλεϊκού οξέος (mRNA). Έπειτα, η αλληλουχία mRNA μεταφράζεται σε τριπλέτες σε αλληλουχία αμινοξέων. Η **γενετική έκφραση (gene expression)** αναφέρεται στο επίπεδο παραγωγής των πρωτεϊνικών μορίων που ορίζονται από ένα γονίδιο. Η απεικόνιση της γενετικής έκφρασης είναι μια από τις πιο θεμελιακές προσεγγίσεις στη γενετική και μοριακή βιολογία. Η βασική τεχνική για τον υπολογισμό της γονιδιακής έκφρασης είναι να υπολογίσουμε το mRNA αντί των πρωτεϊνών, γιατί οι αλληλουχίες mRNA είναι υβριδικές με αυτές του συμπληρωματικού ριβονουκλεϊκού οξέος (cRNA) ή του δεοξυριβονουκλεϊκού οξέος (DNA), ιδιότητα η οποία απουσιάζει από τις πρωτεΐνες.

Οι πίνακες DNA (DNA arrays) είναι νέες τεχνολογίες οι οποίες έχουν σχεδιαστεί έτσι, ώστε να υπολογίζουν τη γενετική έκφραση δεκάδων χιλιάδων γονιδίων σε ένα μόνο πείραμα. Ένας πίνακας DNA αποτελείται από αλληλουχίες probe DNA, οι οποίες είναι εναποθετημένες σε μια επιφάνεια χρυσού ή γυαλιού. Για να αποτιμήσουμε τη γενετική έκφραση των ιστών,

εξάγουμε αρχικά από αυτούς τις αλληλουχίες του mRNA. Οι αλληλουχίες mRNA, έπειτα, μεταγράφονται αντίστροφα σε αλληλουχίες DNA (cDNA). Ένα microarray είναι ένα ειδικά επικαλυμμένο γυαλί μικροσκοπίου στο οποίο τα μόρια cDNA προσκολλώνται σε σταθερά σημεία, τα οποία ονομάζονται σημεία (spots). Σύμφωνα με την πιο σύγχρονη τεχνολογία 19.200 και περισσότερα σημεία μπορούν να τυπωθούν σε ένα γυαλί μικροσκοπίου, το καθένα από τα οποία αναπαριστά ένα μοναδικό γονίδιο. Οι αντίστροφα-μεταγραφόμενες αλληλουχίες DNA (cDNA), λαμβάνουν φθορίζουσα ετικέτα.

Οι ταμπέλες που λαμβάνουν τα cDNA probes είναι είτε Cy3 dye είτε Cy5 dye. Τα Fluorescent cDNA probes, που έχουν ετικέτα Cy5 προέρχονται από κάθε δείγμα mRNA του πειράματος, ενώ αυτά με ετικέτα Cy3 προέρχονται από μια δεξαμενή των mRNA. Κάθε πειραματικό cDNA probe με ετικέτα Cy5 συνδυάζεται με cDNA probes που έχουν ετικέτα Cy3. Όλα τα probes ταυτόχρονα «επωάζονται» στο microarray, καθιστώντας δυνατή στις αλληλουχίες των γονιδίων την συνένωσή τους (hybridization), κάτω από αυστηρές συνθήκες, στους συμπληρωματικούς τους κλώνους και να εναποτίθενται στην επιφάνεια του πίνακα.

Έπειτα, με τη διέγερση του λέιζερ των ενσωματωμένων στόχων, μπορούμε να πετύχουμε μια εκπομπή με ένα χαρακτηριστικό φάσμα, το οποίο μετριέται με τη βοήθεια ενός σαρωτικού, ομοεστιακού μικροσκοπίου. Τότε εισάγονται στο λογισμικό μας μονόχρωμες εικόνες, οι οποίες ψευδο- χρωματίζονται και συγχωνεύονται.




Εικόνα 2*: cDNA microarray διάγραμμα. Αρχικά, οι κλώνοι DNA τοποθετούνται κατά διαστήματα πάνω σε ένα γυαλί μικροσκοπίου. Μετά τη συνένωση των δύο συμπληρωματικών αλυσίδων DNA ( hybridization) το γυαλί σαρώνεται με τη διέγερση λέιζερ και μας δίνει δυο εικόνες για περαιτέρω επεξεργασία.*

Ο λόγος φθορισμού (fluorescence ratio) ποσοτικοποιείται για κάθε γονίδιο και αντανακλά τη σχετική πληθώρα του γονιδίου σε κάθε πειραματικό δείγμα του mRNA σε σύγκριση με τη δεξαμενή αναφοράς mRNA. Η χρήση ενός ανιχνευτή κοινής αναφοράς μας επιτρέπει να θεωρήσουμε αυτούς τους λόγους φθορισμού ως μέτρα του σχετικού επιπέδου έκφρασης κάθε γονιδίου σε όλα τα πειραματικά μας δείγματα.

## *1.3. Σκοπός της εργασίας*

Κατά την ταξινόμηση των γονιδιακών δεδομένων που προκύπτουν από τη γονιδιακή έκφραση των καρκινικών κυττάρων συναντάμε σημαντικά προβλήματα .Ο μεγάλος όγκος των δεδομένων επιδρά αρνητικά στην απόδοση της ταξινόμησης, όπως επίσης και στο κόστος αυτής. Τα προβλήματα αυτά ώθησαν στην αναζήτηση ενός υποσυνόλου των εξεταζόμενων γονιδίων το οποίο να εμπεριέχει τα πιο σημαντικά γονίδια (informative genes). Ο συνδυασμός μπορεί ουσιαστικά να συγκεντρώσει τη συνολική πληροφορία του συνόλου του γονιδιώματος που μας ενδιαφέρει.

Κατά τη διαδικασία αυτή, υπάρχουν στη βιβλιογραφία πολλές μέθοδοι, χωρίς όμως κάποια από αυτές να οδηγεί σε αξιόπιστα αποτελέσματα για την επιλογή ενός ελάχιστου συνόλου γονιδίων στα οποία θα επικεντρωθεί η

μελέτη της βιολογίας, στην κατεύθυνση της δυνατότητας πρόβλεψης τυχόν καρκινικής προδιάθεσης.

Σκοπός της εργασίας αυτής είναι η υλοποίηση της μεθόδου παλινδρόμησης κορυφής με χρήση πυρήνα (Kernel ridge regression) για ταξινόμηση γονιδιακών δεδομένων που προκύπτουν από καρκινικά κύτταρα καθώς και για την αξιολόγηση της σημαντικότητας των γονιδίων. Σκοπός μας είναι ο αλγόριθμος προς υλοποίηση να έχει την καλύτερη δυνατή απόδοση .Για την επίτευξη του στόχου αυτού η εργασία θα ασχοληθεί με την εύρεση δύο βέλτιστων παραμέτρων (σταθερά κανονικοποίησης-πυρήνας) που επιδρούν στην απόδοση του αλγορίθμου.

Ως γενικές παρατηρήσεις για τα αποτελέσματά μας, μπορούμε να παρατηρήσουμε ότι, η μέθοδος της παλινδρόμησης κορυφής με χρήση πυρήνα μπορεί να αποτελέσει μια αξιόπιστη μέθοδο για την επιλογή χαρακτηριστικών γονιδίων στη μελέτη πρόβλεψης των καρκινικών παθήσεων. Η μέθοδος αυτή μας επιτρέπει ουσιαστικά να εκτελέσουμε μη γραμμική παλινδρόμηση κατασκευάζοντας μια γραμμική συνάρτηση παλινδρόμησης σε χώρο χαρακτηριστικών πολλών διαστάσεων. Ο μεγάλος χώρος χαρακτηριστικών που οφείλετε στον μεγάλο όγκο των γονιδιακών δεδομένων –χαρακτηριστικών μπορεί να έχει σαν συνέπεια μεγάλη αύξηση στον αριθμό των παραμέτρων του αλγορίθμου. Η μέθοδος μας αντιμετωπίζει το πρόβλημα αυτό της «κατάρας των πολλών διαστάσεων» με τη χρήση της τεχνικής του πυρήνα.

Σημαντικό ρόλο στη μελέτη της καταλληλότητας της μεθόδου έχει και το πεδίο στο οποίο εφαρμόζεται, δηλαδή το σύνολο δεδομένων, καθώς ανάλογα με την ποιότητά του (ύπαρξη θορύβου ή όχι), αλλά και την ασθένεια στην οποία αναφέρεται, μπορεί να καταστήσει τη μέθοδο περισσότερο ή λιγότερο κατάλληλη για κάποιο είδος καρκίνου.

# Κεφάλαιο2: Η συμβατική μέθοδος παλινδρόμησης κορυφής

Η εμφάνιση της τεχνολογίας micro-array παρέχει για τους αναλυτές δεδομένων μεγάλα δείγματα εκφρασμένων γονιδιακών δεομένων που αποθηκεύονται ταυτόχρονα σε ένα μόνο πείραμα. Διάφορα σύνολα δεδομένων είναι πλέον διαθέσιμα και στο διαδίκτυο. Αυτά τα σύνολα δεδομένων παρουσιάζουν πολλές προκλήσεις. Μία από αυτές οφείλεται στο πολύ μεγάλο αριθμό τιμών που προκύπτουν από την έκφραση γονιδίων ανά πείραμα( μερικές χιλιάδες έως και δεκάδες χιλιάδες),συγκριτικά με τον πολύ μικρό αριθμό πειραμάτων-δειγμάτων(μερικές δεκάδες). Το γεγονός αυτό μπορεί σε πολύ μεγάλο βαθμό να βλάψει την απόδοση της ταξινόμησης καρκινικών ιστών , όπως επίσης και να αυξήσει το κόστος. Έχει επίσης αποδειχθεί, ότι η επιλογή ενός μικρού μέγεθος γονιδίων (informative genes) μπορεί να οδηγήσει σε βελτιωμένα αποτελέσματα ακρίβειας. Το παραπάνω πρόβλημα μπορεί να οριστεί ως **πρόβλημα γονιδιακής επιλογής (gene selection problem).** Η γονιδιακή επιλογή περιλαμβάνει την αναζήτηση των υποσυνόλων ομάδων γονιδίων που μπορούν να διακρίνουν τον καρκινικό ιστό από τον κανονικό και μπορούν να έχουν κάποια ευθεία ή έμμεση συμμετοχή στο μοριακό μηχανισμό της ογκογένεσης. Πιο συγκεκριμένα, η γονιδιακή επιλογή περιλαμβάνει τη μελέτη των κριτηρίων και την αναπαράσταση των τεχνικών που έχουν ως σκοπό τη μείωση του αριθμού των γονιδίων και την επιλογή ενός βέλτιστου (ή κοντά στο βέλτιστο) υποσυνόλου γονιδίων από κάποιο αρχικό σύνολο γονιδίων για την ταξινόμηση όγκων (tumor classification).

Πολλές διαφορετικές μέθοδοι και τεχνικές έχουν αναλυθεί για τη γενετική επιλογή από σετ δεδομένων γενετικής έκφρασης τεχνολογίας microarray. Η ακρίβεια και η αξιοπιστία των αποτελεσμάτων εξαρτώνται από τις μεθόδους που χρησιμοποιούμε. Από τα μέχρι τώρα αποτελέσματα των μετρήσεων, δεν μπορούμε με την εφαρμογή μιας και μοναδικής μεθόδου να εγγυηθούμε κάποιο ικανοποιητικό αποτέλεσμα.

Στη συγκεκριμένη εργασία θα ασχοληθούμε ιδιαίτερα με τη μελέτη και τη βελτιστοποίηση της τεχνικής της παλινδρόμησης κορυφής(Ridge Regression-RR) .Η μέθοδος αυτή αποτελεί ουσιαστικά την κανονικοποιημένη τροποποίηση της μεθόδου των ελαχίστων τετραγώνων. Ο πολύ μεγάλος αριθμός εκφρασμένων γονιδίων συγκριτικά με το μικρό αριθμό πειραμάτων μπορεί να δημιουργήσει πρόβλημα υπερταυτοποιήσης στον αλγόριθμο ταξινόμησης. Για το σκοπό αυτό θα μελετήσουμε τη μέθοδο παλινδρόμησης κορυφής με χρήση πυρήνα για ταξινόμηση των γονιδιακών δεδομένων και επιλογή γονιδίων. Η εργασία εστιάζει στην βελτιστοποίηση των δύο παραπάνω αλγορίθμων μέσω της εύρεσης των ιδανικών τιμών για τη σταθερά κανονικοποίησης (RP) και τον πυρήνα (K) ,που είναι οι παράμετροι που επηρεάζουν την απόδοση του αλγορίθμου.

Αρχικά θα παρουσιάσουμε το πρόβλημα και την συμβατική μέθοδο παλινδρόμησης κορυφής.

## *2.1 Περιγραφή προβλήματος*

Σκοπός μας είναι η κατάταξη ενός αριθμού δειγμάτων σε δύο κλάσεις-εξόδους. Τα δείγματα είναι ασθενείς ,κάθε ένας από τους οποίους «χαρακτηρίζεται » από πολύ μεγάλο αριθμό γονιδίων. Οι τιμές των γονιδίων είναι ,δηλαδή, τα χαρακτηριστικά (features) για κάθε ασθενή-δείγμα(pattern). Για κάθε δείγμα ασθενή αντιστοιχεί μία κλάση –έξοδος που μπορεί να έχει τιμή -1/1.Εξαιτίας του μεγάλου αριθμού χαρακτηριστικών κάθε δείγματος δημιουργείται η ανάγκη για απεικόνιση των δεδομένων σε έναν χώρο πολλών διαστάσεων. Ο χώρος αυτός ονομάζεται χώρος χαρακτηριστικών (feature space), ενώ ο αρχικός χώρος δεδομένων –χώρος εισόδων. Η κατάταξη των δεδομένων στις δύο κλάσεις δεν μπορεί να γίνει –χωρίς προβλήματα- με γραμμικούς ταξινομητές , αφού πολύ σπάνια στην πραγματική ζωή μπορούμε να βρούμε σύνολα δεδομένων που είναι γραμμικώς διαχωρίσιμα. Ο πιο ακριβής διαχωρισμός των δειγμάτων απαιτεί τη δημιουργία διαχωριστικού υπερεπιπέδου (hyperplane) στον χώρο δεδομένων και όχι στο χώρο εισόδων.

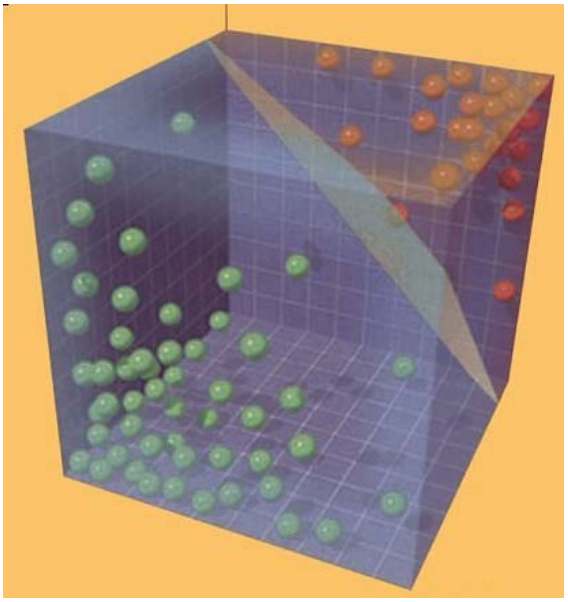
Ο προσδιορισμός της διαχωριστικής ευθείας μεταξύ των δεδομένων που ανήκουν σε δύο διαφορετικές κλάσεις μπορεί να επιτευχθεί με μια σειρά

μεθόδων. Η πιο απλή είναι η μέθοδος των ελαχίστων τετραγώνων. Είναι η μέθοδος που βασίζεται στην ελαχιστοποίηση του αθροίσματος των αποστάσεων των σημείων παρατήρησης από την ευθεία. Η επιλογή μεταβλητών με απλή γραμμική παλινδρόμηση(μέθοδος ελαχίστων τετραγώνων) μπορεί να εμφανίζει πολλά προβλήματα , ιδιαίτερα στην περίπτωση που ένας μεγάλος αριθμός δεδομένων εμφανίζουν υψηλή συσχέτιση. Σε αυτήν την κατηγορία ανήκουν και τα δεδομένα που προκύπτουν από την έκφραση γονίδιων. Σε αυτήν την εργασία θα περιγράψουμε ,θα υλοποιήσουμε ,θα βελτιστοποιήσουμε και θα εφαρμόσουμε πάνω σε γνωστά σύνολα δεδομένων την τεχνική Kernel Ridge Regression.

**Η μέθοδος Ridge Regression παρουσιάστηκε ως τροποποίηση της μεθόδου των ελαχίστων τετραγώνων από τους Hoerl και Kennard το 1970 .Πρόκειται ουσιαστικά για μία γενίκευση της μεθόδου των ελαχίστων τετραγώνων στην οποία όμως εφαρμόζουμε και έναν περιορισμό-«ποινή» στην τιμή του διανύσματος συντελεστών w.**

Η γραμμική μέθοδος Ridge Regression που περιγράψαμε παραπάνω δεν είναι κατάλληλη για το πρόβλημα που αντιμετωπίζουμε.Ο χώρος δεδομένων μας είναι αρχικά πολύ μεγάλων διαστάσεων μιας και οι διαστάσεις του καθορίζονται από τις μεταβλητές περιγραφής που είναι χιλιάδες γονίδια για κάθε ασθενή. Αρχικά , λοιπόν, ένας γραμμικός ταξινομητής όπως αυτός που προκύπτει από τη μέθοδο Ridge Regression θα αρκούσε για το διαχωρισμό των δεδομένων μας. Καθώς όμως σκοπός μας είναι να εντοπίσουμε μια μικρή ομάδα γονιδίων που καθορίζει την κλάση στην οποία ανήκουν τα δείγματα , ο χώρος μας θα μικρύνει σημαντικά. Όταν τα γονίδια είναι πλέον δεκάδες η πιθανότητα να μας αρκεί ένας γραμμικός ταξινομητής μειώνεται σημαντικά.Τα περισσότερα εξάλλου ρεαλιστικά προβλήματα περιλαμβάνουν μη διαχωρίσιμα δεδομένα και ο χώρος μας δεν αποτελεί εξαίρεση. Για το λόγο αυτό χρησιμοποιούμε μη γραμμικό αλγόριθμο ταξινόμησης εισάγοντας τη χρήση τεχνικών πυρήνα. Ο πυρήνας είναι μία απλή συνάρτηση αντιστοίχισης διανυσματικών γινομένων (dot products) στον χώρο χαρακτηριστικών (feature space) που είναι πολύ μεγάλων διαστάσεων.

**Η Kernel ridge regression (KRR) , είναι ίσως η πιο χαρακτηριστική μορφή της ταξινόμησης με χρήση πυρήνα.** Έχοντας τα αρχικά δεδομένα – ασθενείς, για τα οποία γνωρίζουμε την κλάση στην οποία ανήκουν, η μέθοδος kernel ridge regression κατασκευάζει ένα γραμμικό μοντέλο σε έναν δεδομένο χώρο χαρακτηριστικών ( feature space ) F που ορίζεται από έναν μετασχηματισμό του χώρου εισόδων. κατασκευάζει ένα γραμμικό μοντέλο-ταξινομητή: $\mu(x) = w \cdot \phi(x) + b$ για έναν καθορισμένο χώρο χαρακτηριστικών (Feature Space ) $\mathbb{F}$ που προσδιορίζεται από ένα καθορισμένο μετασχηματισμό στο χώρο εισόδων , $\phi(x) : \mathrm{X} \to \mathbb{F}$. Ωστόσο, αντί να προσδιορίσουμε τον χώρο χαρακτηριστικών $\mathbb{F}$ απευθείας, τον εξάγουμε με τη βοήθεια μιας θετικής και πεπερασμένης συνάρτησης πυρήνα [5] $\mathrm{K} : \mathrm{X} \times \mathrm{X} \to \mathbb{R}$ ,η οποία ορίζει το εσωτερικό γινόμενο μεταξύ διανυσμάτων που ορίζονται στον χώρο χαρακτηριστικών $\mathbb{F}$, π.χ $K(x, x') = \phi(x) \cdot \phi(x')$ .



*Εικόνα 3 : Υπερεπίδεδο διαχωρισμού μη γραμμικά διαχωρίσιμων δεδομένων .Στο σχήμα φαίνεται καθαρά ότι μια ευθεία δε θα αρκούσε για τον διαχωρισμό των δεδομένων μας.*

Η τεχνική του πυρήνα μας επιτρέπει να δημιουργήσουμε γραμμικά μοντέλα σε χώρους χαρακτηριστικών πολλών ή ακόμα και άπειρων διαστάσεων, χρησιμοποιώντας μόνο ποσότητες πεπερασμένων διαστάσεων , όπως είναι οι συναρτήσεις πυρήνα .Προσδιορίζουμε πλέον το ιδανικό υπερεπίπεδο και όχι

την ευθεία που διαχωρίζει τα δεδομένα μας. Το πρόβλημα μας έτσι από μη γραμμικό μετασχηματίζεται σε γραμμικό στον νέο χώρο μεγαλύτερων διαστάσεων. Έτσι προκύπτει η μέθοδος που χρησιμοποιούμε :Kernel Ridge Regression.

Η μέθοδος παλινδρόμησης κορυφής με χρήση πυρήνα θα μπορούσε να χαρακτηριστεί και ως κανονικοποιημένη μέθοδος των ελαχίστων τετραγώνων για ταξινόμηση και παλινδρόμηση. Η γραμμική εκδοχή της μεθόδου είναι παρόμοια με την **Fisher's discriminant** για ταξινόμηση. Η μη γραμμική εκδοχή είναι παρόμοια με τη μέθοδο ταξινόμησης με μηχανές διανυσμάτων υποστήριξης (support vector machines-SVMs) .Η διαφοροποίηση τω δύο τελευταίων μεθόδων έγκειται στο κριτήριο προς βελτιστοποίηση. Στη μέθοδο με χρήση διανυσμάτων υποστήριξης το κριτήριο βασίζεται μόνο στα δείγματα που βρίσκονται κοντά στο όριο απόφασης ενώ στην KRR σε όλα τα σημεία. Η λύση στην μέθοδο παλινδρόμησης κορυφής εξαρτάται από όλα τα δείγματα του συνόλου εκπαίδευσης. Συμπερασματικά, η μέθοδος KRR είναι κατάλληλη μόνο για σύνολα δεδομένων με λίγα δείγματα εκπαίδευσης. Για την γραμμική εκδοχή της μεθόδου , η εισαγωγή του πυρήνα είναι χρήσιμη αν ο αριθμός των χαρακτηριστικών είναι μεγάλος και ο αριθμός των δειγμάτων μικρός.
Η μέθοδος ridge regression που περιγράφηκε παραπάνω έχει ανακαλυφθεί πολλές φορές και είναι γνωστή με πολλά ονόματα π.χ. "ridge regression" (Hoerl-Kennard), "regularization networks" (Poggio-Girosi), neural network "weight-decay", see historical in http://www.anc.ed.ac.uk/~mjo/intro/node19.html ).

Για να κατανοήσουμε καλύτερα τη μέθοδο Kernel Ridge Regression θα περιγράψουμε αρχικά σύντομα τη συμβατική μέθοδο των ελαχίστων τετραγώνων και θα παρουσιάσουμε τους λόγους που οδήγησαν στην τροποποίηση της και στην δημιουργία της τεχνικής Ridge Regression.Στη συνέχεια θα παρουσιάσουμε τις μεταβολές της τεχνικής με την εισαγωγή και του πυρήνα.

## *2.2 Προσδιορισμός της διαχωριστικής ευθείας ταξινόμησης*

Έστω ότι έχουμε ένα σύνολο δειγμάτων ($X_1$,…..,$X_T$).Κάθε δείγμα είναι ένα διάνυσμα που αντιπροσωπεύει έναν ασθενή και κάθε στοιχείο του διανύσματος είναι η τιμή για κάποιο γονίδιό του.

Έστω $y_t$ η κλάση στην οποία ανήκει ο κάθε ασθενής και ότι κάποιος παρατηρητής είναι σε θέση να μας δώσει τις πραγματικές τιμές της κλάσης (θανατηφόρα ή μη θανατηφόρα περίπτωση) για κάθε ένα από τα παραπάνω δείγματα.

Το πρόβλημα μας είναι να κατασκευάσουμε μια μηχανή εκμάθησης με βάση τα δοσμένα δεδομένα. Ο ταξινομητής μας θα πρέπει να κατατάσσει οποιοδήποτε νέα δεδομένα σε μία από της δύο κλάσεις κάνοντας έτσι μία πρόβλεψη .Αν συμβολίσουμε με $y_t$ την πρόβλεψη που κάνει ο ταξινομητής μας , τότε απαιτούμε η πρόβλεψη να ναι τέτοια ώστε να ελαχιστοποιείται κάποιο μέτρο σύγκρισης ανάμεσα στην πρόβλεψη και την πραγματική τιμή. Το μέτρο σύγκρισης είναι διαφορετικό ανάλογα με τον αλγόριθμο κατάταξης που χρησιμοποιούμε.

Για να κατανοήσουμε καλύτερα την αξία του μέτρου σύγκρισης, υπενθυμίζουμε ότι ο αλγόριθμος κατάταξης απεικονίζεται γραφικά ως μία ευθεία(αν είναι γραμμικός) ή ως ένα υπερεπίπεδο τόσων διαστάσεων όσα και τα χαρακτηριστικά-γονίδια των δειγμάτων .Η ευθεία αυτή ή το επίπεδο διαχωρίζουν το χώρο δεδομένων στις δύο κλάσεις. Για να έχουμε , λοιπόν, βελτιστοποίηση του αλγορίθμου ή αλλιώς τον πλησιέστερο στην πραγματικό δυνατό διαχωρισμό πρέπει να εντοπίσουμε ποια είναι η καλύτερη ευθεία/ή το καλύτερο επίπεδου διαχωρισμού. Ψάχνουμε δηλαδή την εξίσωση της διαχωριστικής ευθείας/ επιπέδου που διαχωρίζει πιο σωστά τα δεδομένα μας. Ο καθορισμός της ευθείας γίνεται με βάση κάποια δείγματα-ασθενείς για τα οποία γνωρίζουμε τις εξόδους-κλάσεις(training set). Έχοντας λοιπόν κάποια αρχικά σημεία στο χώρο ψάχνω να βρω την ευθεία που τα διαχωρίζει καλύτερα.

Ένας απλός τρόπος για να εντοπίσουμε αυτήν την ευθεία είναι η ελαχιστοποίηση της απόστασης από την ευθεία διαχωρισμού των σημείων προς ταξινόμηση. Αυτή η ιδέα αποτελεί τη βάση της μεθόδου των ελαχίστων τετραγώνων.

## *2.3 Απλή γραμμική παλινδρόμηση*

Έστω $Y = wX + \varepsilon$ είναι η διαχωριστική ευθεία προς εκτίμηση που διαχωρίζει τα δεδομένα παρατήρησης $(\mathrm{X}_1, \mathrm{Y}_1), ..., (\mathrm{X}_n, \mathrm{Y}_n)$ σε δύο κλάσεις.

Στην παραπάνω σχέση :

*Y είναι ένα nx*1 διάνυσμα των μεταβλητών εξόδου-εξαρτημένων μεταβλητών,

*X είναι ένας nxp πίνακας που περιέχει τα διανύσματα των χαρακτηριστικών – των ερμηνευτικών(ανεξάρτητων)μεταβλητών και*

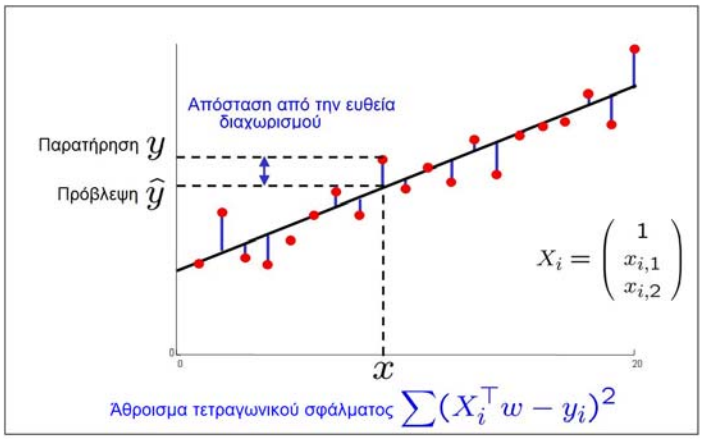
*ε είναι ένα nx1 διάνυσμα των τυχαίων σφαλμάτων.*

*Το διάνυσμα w είναι ένα px* 1 διάνυσμα που οι παράμετροι του εκφράζουν το βαθμό συμμετοχής των χαρακτηριστικών στη διαμόρφωση της εξόδου .Οι παράμετροι αυτοί είναι αρχικά άγνωστες και εκτιμούνται βάσει των δεδομένων παρατήρησης (X,Y)

Η απόσταση των σημείων παρατήρησης από την ευθεία παλινδρόμησης είναι:

$\sum_{i=1}^{n} (Y_i - X_i w)^2$ ,i=1,2,3..n (1.1)

Στόχος της μεθόδου είναι η ελαχιστοποίηση του αθροίσματος των τετραγώνων των αποκλίσεων των πραγματικών τιμών y από την ευθεία γραμμή που υπολογίζουμε




*Εικόνα 4:Η συμβατική μέθοδος των ελαχίστων τετραγώνων*

Αρκεί να ελαχιστοποιήσουμε την εξίσωση των ελαχίστων τετραγώνων(εξίσωση 1.1) ως προς την μεταβλητή w που μας δείχνει την κλίση της ευθείας/επιπέδου προκειμένου να βρούμε την ιδανική ευθεία παλινδρόμησης. Ουσιαστικά υπολογίζουμε την ευθεία εκείνη για την οποία το θροισμα των αποστάσεων όλων των σημείων από την ευθεία είναι το μικρότερο δυνατό

$$L_T(w) = \min_w \sum_{t=1}^{T} (y_t - wx_t)^2 \quad (1.2)$$

Από την (1.2) προκύπτει ότι το ιδανικό w σύμφωνα με τη μέθοδο των ελαχίστων τετραγώνων υπολογίζεται από τη σχέση:

$w_{LS} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$ (1.3) και έχει διακύμανση: $\mathrm{var}(w_{LS}) = (X^T X)^{-1}\sigma^2$ (1.4)

Υπενθυμίζουμε ακόμα ότι για το μέσο τετραγωνικό σφάλμα MSE ισχύει η σχέση:

$$MSE = bias^2 + \mathrm{var}\,iance \quad (1.5),$$

όπου bias:σταθερά της ευθείας διαχωρισμού ( Y=wX+bias ), variance:διακύμανση, MSE: το μέσο τετραγωνικό σφάλμα των πραγματικών από τις εκτιμούμενες τιμές.

Παρατηρούμε στην εξίσωση (1.3) ότι η βασική προϋπόθεση για την εκτίμηση με τη μέθοδο των ελαχίστων τετραγώνων της διαχωριστικής ευθείας

παλινδρόμησης είναι να υπάρχει ο $(X^{\mathrm{T}}X)^{-1}$. Ο πίνακας $(X^{\mathrm{T}}X)^{-1}$ δεν ορίζεται σε δύο περιπτώσεις: (1) Όταν p >> n και (2)όταν εμφανίζεται το πρόβλημα της συγγραμμικότητας των ανεξαρτήτων μεταβλητών X . Στην πρώτη περίπτωση αντιμετωπίζουμε το πρόβλημα overfitting. Πιο συγκεκριμένα Έχει παρατηρηθεί ο κίνδυνος χειροτέρευσης της απόδοσης των νέων δεδομένων, την ίδια στιγμή που η απόδοση των εκπαιδευόμενων παραδειγμάτων θα αύξανε. Το πρόβλημα αυτό σχετίζεται και με τον μικρό αριθμό δειγμάτων συγκριτικά με των πολύ μεγάλο αριθμό χαρακτηριστικών-διαστάσεων.

Στη δεύτερη περίπτωση , ο πίνακας $(X^{\mathrm{T}}X)$ είναι μη αντιστρέψιμος λόγω της υψηλής συσχέτισης που εμφανίζουν οι ανεξάρτητες μεταβλητές (φαινόμενο συγγραμμικότητας) Λόγω του μεγάλου αριθμού των ανεξάρτητων μεταβλητών – χαρακτηριστικών είναι πιθανό να αντιμετωπίσουμε το πρόβλημα της συγγραμμικότητας. Είναι δυνατό κάποιες από τις ανεξάρτητες μεταβλητές – γονίδια $X_1$, $X_2$,…,$X_{n-1}$ να είναι γραμμικά εξαρτημένες μεταξύ τους με συνέπεια ο πίνακας πληροφορίας $X^T X$ να μην αντιστρέφεται (η ορίζουσα του είναι 0). Αυτό το πρόβλημα είναι γνωστό ως πρόβλημα πολυσυγγραμικότητας. Ένας απλός τρόπος αντιμετώπισης του είναι η αφαίρεση κάποιων ανεξάρτητων μεταβλητών (χάνοντας όμως «πληροφορία»).Ακόμη και όταν η ορίζουσα του $X^T X$ δεν είναι ακριβώς 0 αλλά «κοντά» στο 0 (ασθενής πολυσυγγραμικότητα) παρουσιάζεται πρόβλημα. Σε αυτή την περίπτωση μπορεί να εμφανιστούν σφάλματα στρογγυλοποίησης κατά την αντιστροφή του $X^T X$ με συνέπεια οι εκτιμήσεις που παίρνουμε να μην είναι αξιόπιστες. Η ορίζουσα του $X^T X$ είναι κοντά στο 0 όταν υπάρχει ισχυρή συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών. Ακόμα παρατηρώντας την σχέση (1.5) μπορούμε να συμπεράνουμε ότι όταν ο $X^T X$ είναι κοντά στο 0, ο $(X^T X)^{-1}$ μεγαλώνει πολύ λόγω αντιστροφής και κατά συνέπεια αυξάνεται η διακύμανση(σχέση (1.4)) και το μέσο τετραγωνικό σφάλμα κατά πολύ. Συνέπεια αυτού είναι να έχουμε κακές εκτιμήσεις και προβλέψεις. Ακόμα ένα πρόβλημα που παρατηρείται όταν τα δεδομένα είναι γραμμικά εξαρτημένα μεταξύ τους είναι οι μεγάλες αλλαγές στον υπολογισμό των παραμέτρων κατά το σβήσιμο ή την προσθήκη μεταβλητών καθώς και οι μεγάλες διακυμάνσεις των

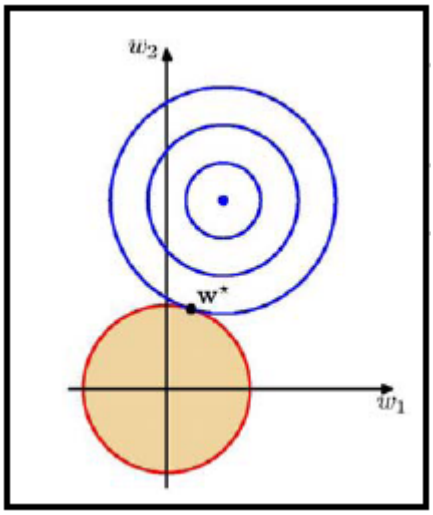
παραμέτρων από δείγμα σε δείγμα. Επιπλέον μπορεί ένα δεδομένο να το θεωρούμε σημαντικό και μετά από την προσθήκη ή αφαίρεση κάποιου να μην προκύπτει ως σημαντικό, λόγω της υψηλής συσχέτισης με κάποιο άλλο δεδομένο.

Τέτοιου είδους προβλήματα , τα λεγόμενα “badly conditioned linear regression problems” (Hoerl and Kennard, 1970) είναι εδώ και πολύ καιρό ένα σημαντικό πρόβλημα στη στατιστική και την επιστήμη των υπολογιστών. Το πρόβλημα είναι ακόμα πιο διακριτό σε δεδομένα πάρα πολλών διαστάσεων. Τέτοια είναι τα περισσότερα δεδομένα που προκύπτουν από έκφραση γονιδίων.

Μία από τις πιο αποτελεσματικές μεθόδους αντιμετώπισης τέτοιων προβλημάτων είναι και η και ο αλγόριθμος Ridge Regression.Είναι μάλιστα αξιοσημείωτο ότι ενώ η δημιουργία του βασίζεται σε μια πολύ απλή ιδέα και οι υπολογισμοί του πλησιάζουν πολύ αυτούς της μεθόδου των ελαχίστων τετραγώνων , είναι μια αρκετά αποδοτική μέθοδος.

## *2.4 Κανονικοποιημένη γραμμική παλινδρόμηση*

Εξαιτίας των κινδύνων που εντοπίσαμε παραπάνω στην συμβατική μέθοδο των ελαχίστων τετραγώνων, χρειάζεται να κανονικοποιήσουμε την αρχική μας σχέση και να θέσουμε ένα όριο στη μεταβλητή w.Για το σκοπό αυτό θα χρειαστούμε μία σταθερά κανονικοποιήσης η οποία θέτει απλά ένα όριο στην τιμή της νόρμας του συντελεστή W .Τη σταθερά αυτή την ονομάζουμε L και τη συμβολίζουμε με RP στους κώδικες υλοποίησης του αλγορίθμου και για το σκοπό που μόλις αναφέρουμε στη σχέση (1.1) προσθέτουμε τον όρο $l \| w \|^2$.Τροποποιούμε με τον τρόπο αυτό «τεχνητά» την ορίζουσα του $X^T X$ σε $X^T X + RP\mathrm{I}$ ώστε να μην είναι ο $X^T X$ μοναδιαίος και να μπορεί να γίνει αντιστροφή.




*Εικόνα 5 : Η επίδραση της κανονικοποίησης .Η συνολική συνάρτηση κόστους της μεθόδου ταξινόμησης είναι το άθροισμα δύο σφαιρών. Το άθροισμα είναι και αυτό μία σφαίρα. Το ελάχιστο του συνδυασμού βρίσκεται στη γραμμή του ελάχιστου του τετραγωνικού λάθους και της αρχικής τιμής .Ο «κανονικοποιητής» απλά συρρικνώνει το w.*

Εγγυούμαστε έτσι με έναν απλό τρόπο για να εγγυηθούμε την αντιστρεψιμότητα του πίνακα $X^T X$ .Ουσιαστικά προσθέτουμε έναν διαγώνιο πίνακα διαστάσεων (p+1)×(p+1) στο γινόμενο $X^{\mathrm{T}} X$ πριν το αντιστρέψουμε. Προσθέτουμε ,δηλαδή ,μία θετική σταθερά σε όλα τα στοιχεία της κύριας διαγωνίου του $X^{\mathrm{T}} X$ . Από τον καθορισμό της τιμής της σταθεράς αυτής (l) εξαρτάται και ο βαθμός βελτίωσης του αλγορίθμου μας συγκριτικά με την απλή μέθοδο των ελαχίστων τετραγώνων

Έτσι από την εξίσωση (1.1) μετά και την προσθήκη της «ποινής» η απόσταση των σημείων από την ευθεία υπολογίζεται από: $\sum_{i=1}^{n}(Y_i - X_i w)^2 + l \,||\, w||^2$ (1.6).Αυτήν την απόσταση πλέον προσπαθούμε να ελαχιστοποιήσουμε ως προς w: $\min_{w} \sum_{i=1}^{n}(Y_i - X_i w)^2 + l \,||\, w||^2$ (1.7) και ο τύπος (1.7) αποτελεί την έκφραση της τεχνικής Ridge Regression.

Προσθέτοντας τον όρο $l \,||\, w||^2$ βάζουμε ουσιαστικά ένα όριο στο w .

Δεν είναι δύσκολο να καταλήξουμε από τη σχέση (1.7) στον παρακάτω τύπο:

$\hat{W}_R = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + l\mathrm{I})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$ (1.8) που υπολογίζει ουσιαστικά την κλίση της διαχωριστικής ευθείας.

Η τιμή της παραμέτρου RP αντικατοπτρίζει το βαθμό διαφοροποίησης της μεθόδου Ridge Regression από τη μέθοδο ελαχίστων τετραγώνων. για l=0 από τις σχέσεις (1.6),(1.7),(1.8)ανακτούμε τους τύπους της μεθόδου των ελαχίστων τετραγώνων(1.1),(1.2),(1.3).Αξίζει εδώ να σημειώσουμε ότι με μεγαλύτερο l το «όριο-ποινή» για το W τείνει να γίνεται πιο αυστηρό και η λύση για το W είναι μικρότερη. Ακόμα Η επίλυση του προβλήματος της συγγραμμικότητας δεν επιτυγχάνεται χωρίς κανένα κόστος. Από τη σχέση 1.1 άλλωστε εύκολα παρατηρούμε ότι όσο μεγαλύτερο είναι το w τόσο μεγαλύτερο θα είναι και το άθροισμα των ελαχίστων τετραγώνων .Όσο μεγαλύτερη είναι η τιμή του w , τόσο πιο εύκολο είναι να αντιμετωπιστεί το πρόβλημα της συγγραμμικότητας. Υπάρχει μία ιδανική τιμή για την παράμετρο RP , της οποίας ο καθορισμός εξαρτάται από εμάς.

Διακύμανση του εκτιμητή Ridge regression ( $\mathrm{var}(\hat{w}_R)$ ) :

Υπενθυμίζουμε ότι: $\mathrm{E}\hat{w}_{LS} = w$ (1.8) και $\mathrm{var}(\mathrm{E}) = \sigma^2\mathrm{I}_n$ (1.9) και $\mathrm{var}(\hat{w}_{LS}) = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\sigma^2$ (1.10)

Υπολογίζουμε με βάση τα παραπάνω την διακύμανση του εκτιμητή ridge regression $w_R$

$\mathrm{var}(\hat{w}_R) = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + l\mathrm{I})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X} + l\mathrm{I})^{-1}\sigma^2$ (1.11)

Μέσο Τετραγωνικό σφάλμα του εκτιμητή Ridge regression $MSE_R$ =( $E \parallel \hat{w}_R - w \parallel^2$ ):

$MSE_R = E \parallel \hat{w}_R - w \parallel^2 = E(\hat{w}_R - w)^T(\hat{w}_R - w)$ =

= $E(\hat{\beta}_R - E\hat{w}_R + E\hat{w}_R - w)^T(\hat{w}_R - E\hat{w}_R + E\hat{w}_R - w)$

= $E \parallel \hat{\beta}_R - E\hat{\beta}_R \parallel^2 + 2E\{(E\hat{w}_R - w)^{\mathrm{T}}(\hat{w}_R - \mathrm{E}\hat{w}_R)\} + E \parallel \hat{w}_R - w \parallel^2$

= $E \parallel \hat{w}_R - E\hat{w}_R \parallel^2 + \parallel E\hat{w}_R - w \parallel^2$

= $true(Variance) + \parallel bias \parallel^2$

Τελικά $MSE_R = true(Variance) + ||bias||^2$ (1.12)

Από τη σχέση (1.12) θα μπορούσε να συμπεράνει κανείς ότι $MSE_{LS} < MSE_R$ μιας και στο μέσο τετραγωνικό σφάλμα του εκτιμητή της συμβατικής μεθόδου ελαχίστων τετραγώνων $MSE_{LS}$ δεν υπάρχει ο όρος $||bias||^2$ που προστέθηκε εδώ λόγω της σταθεράς που προσθέσαμε. Με κατάλληλη, όμως ,επιλογή της παραμέτρου l η διακύμανση στην περίπτωση του Ridge Regression μπορεί να γίνει σημαντικά μικρότερη της διακύμανσης με τη συμβατική μέθοδο ελαχίστων τετραγώνων. Έχει μάλιστα αποδειχτεί (Hoerl and Kennard 1970) ότι πάντα μπορούμε να βρούμε μια τουλάχιστον τιμή του L για την οποία να ισχύει: $MSE(\hat{w}_R) < MSE(\hat{w}_{LS})$

Με άλλα λόγια η μέθοδος ridge regression μπορεί να βελτιώσει την εκτίμηση του w και επομένως και της ευθείας διαχωρισμού των δεδομένων μας.
Συμπερασματικά: Η μέθοδος Ridge regression προτάθηκε για πρώτη φορά από τους Hoerl and Kennard το 1971.Είναι μία από τις πιο δημοφιλείς μεθόδους για την αντιμετώπιση του προβλήματος της συγγραμμικότητας. Στην περίπτωση που τα δεδομένα μας εμφανίζουν συγγραμμικότητα, οι συνηθισμένοι εκτιμητές της μεθόδου των ελαχίστων τετραγώνων –χωρίς τη χρήση ορίου μπορούν να γίνουν πολύ ασταθείς , λόγω των μεγάλων διακυμάνσεων που παρουσιάζουν. Οι μεγάλες αυτές διακυμάνσεις οδηγούν και σε μεγάλα σφάλματα και ανακριβείς προβλέψεις. Στη μέθοδο Ridge Regression επιτρέπονται «μεροληπτικοί» εκτιμητές για τους συντελεστές παλινδρόμησης με τροποποίηση της μεθόδου των ελαχίστων τετραγώνων. Επιτυγχάνεται έτσι μια ουσιαστική μείωση στη διακύμανση που ακολουθείται και από αύξηση στη σταθερότητα των συντελεστών.

Οι εκτιμητές Ridge Regression είναι γνωστό ότι έχουν πολλές ευνοϊκές ιδιότητες. Για παράδειγμα , οι Hoerl και Kennard το 1970 απέδειξαν ότι εμφανίζουν πολύ μικρότερο τετραγωνικό σφάλμα συγκριτικά με αυτό τη συμβατικής μεθόδου ελαχίστων τετραγώνων αν γίνει κατάλληλη επιλογή της μεροληπτικής παραμέτρου Regression.Ακόμα οι εκτιμητές RR διαδραματίζουν σημαντικό ρόλο στα Bayesian μοντέλα. Πολλές μέθοδοι που

χρησιμοποιούνται για την καταπολέμηση του φαινομένου της συγγραμμικότητας σχετίζονται πολύ με τη μέθοδο ridge regression.Ανάμεσα σε αυτές τις μεθόδους είναι η «διαρκής παλινδρόμηση» (continuum regression) (Stone & Brooks 1990,Sundberg ,1993), partial Least squares(PLS), παλινδρόμηση κύριου συστατικό(principal component regression(PCR)) και άλλες (Bjoerkstroem and Sundberg,1999(Least squares ridge regression)).Αυτές οι παρατηρήσεις καθιστούν ακόμα πιο ενδιαφέρουσα τη μελέτη των ιδιοτήτων της μεθόδου Ridge Regression.

## *2.5 Εκπαίδευση και ταξινόμηση με παλινδρόμηση κορυφής(Ridge Regression)*

Έχοντας καταλήξει στην επιλογή της μεθόδου Ridge Regression για όλους τους λόγους που περιγράψαμε στην προηγούμενη παράγραφο, επανερχόμαστε στο πρόβλημα του ακριβή προσδιορισμού της διαχωριστικής ευθείας(εκπαίδευση) και στο διαχωρισμό των νέων δεδομένων με βάση αυτήν (ταξινόμηση).

Για να βρούμε τη διαχωριστική ευθεία/επίπεδο που διαχωρίζει πιο σωστά τα δεδομένα μας με βάση τον αλγόριθμο Ridge Regression ελαχιστοποιούμε τη σχέση (1.7) ως προς w.

$$\min_{w} \sum_{i=1}^{n} (Y_i - X_i w)^2 + l \,||\, w \,||^2 \quad (1.7)$$

Για $x_t$ , $y_t$ χρησιμοποιούμε ένα σύνολο δεδομένων για τα οποία γνωρίζουμε από πριν τόσο σύνολο τιμών των χαρακτηριστικών – γονιδίων όσο και την κλάση στην οποία ανήκει κάθε δείγμα-ασθενής. Το σύνολο αυτό ονομάζεται training set.Έτσι προκύπτει η τιμή του w, δηλαδή η κλίση της διαχωριστικής ευθείας/επιπέδου.

$$\hat{W}_R = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + l\mathrm{I})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y} \quad (1.8)$$

Με τον καθορισμό του w με τη βοήθεια του training set ολοκληρώνεται η διαδικασία της εκπαίδευσης του αλγορίθμου μας. Από εδώ και πέρα

μπορούμε να ταξινομήσουμε κάθε νέο δείγμα γνωρίζοντας μόνο το διάνυσμα των χαρακτηριστικών του. Μπορούμε δηλαδή να κατατάξουμε κάθε νέο ασθενή σε μία από τις κλάσεις βάσει των τιμών των γονιδίων του. Η διαδικασία αυτή ονομάζεται ταξινόμηση νέων δειγμάτων και περιγράφεται από τη σχέση (1.13):

$$y_{new} = \hat{w}_R x_{new} \quad (1.13)$$

## *2.6 Εισαγωγή των πολλαπλασιαστών Lagrange στην ridge regression*

Συνδυάζοντας τις σχέσεις 1.8,1.13 μπορούμε εύκολα να συμπεράνουμε ότι η πρόβλεψη της νέας τιμής για την παλινδρόμηση κορυφής εξαρτάται από το εσωτερικό γινόμενο των εισόδων X. Από εδώ προκύπτει και η ιδέα για την εισαγωγή της τεχνικής του πυρήνα , μιας συνάρτησης δηλαδή που θα λειτουργεί ως μετασχηματισμός και θα αντιστοιχεί τον χώρο εισόδων στον μεγαλύτερων διαστάσεων χώρο χαρακτηριστικών. Με κατάλληλη επιλογή του πυρήνα μπορούν να απλουστευτούν κατά πολύ οι υπολογισμοί όπως θα εξηγήσουμε αναλυτικότερα και στην παράγραφο 2.7.Προκειμένου να εκμεταλλευτούμε περισσότερο την παρατήρηση αυτή θα χρειαστεί να εισάγουμε τους πολλαπλασιαστές Lagrange.

Μία πολύ ενδιαφέρουσα μαθηματική μέθοδος βελτιστοποίησης είναι η μέθοδος των πολλαπλασιαστών Lagrange.Με τη βοήθεια αυτής θα εκφράσουμε τις παραπάνω σχέσεις (1.8), (1.13) που αντιστοιχούν στην διαδικασία της εκπαίδευσης και ταξινόμησης/πρόβλεψης . Η γνωστή μέθοδος των πολλαπλασιαστών Lagrange μας βοηθά να λύσουμε το πρόβλημα της ελαχιστοποίησης της σχέσης (1.7) διευκολύνοντας τους υπολογισμούς.

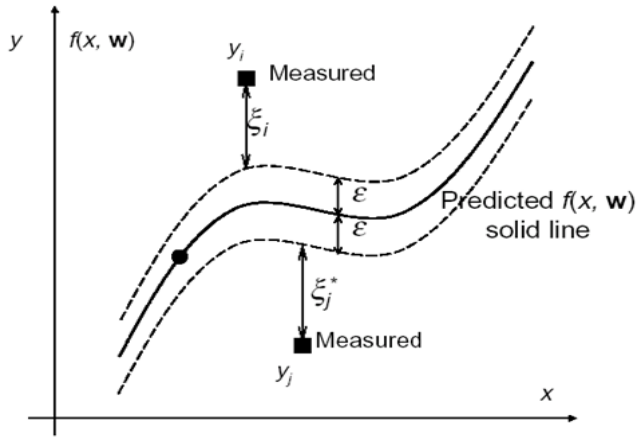
Ξαναεκφράζουμε την σχέση (1.5) ως εξής

$$L_T(w) = l\,||W||^2 + \sum_{t=1}^{T} \xi_t^2 \quad (1.5)$$

Υπό τους περιορισμούς

$$\xi_t = y_t - w x_t \quad (1.14)$$

$\xi_t$ :είναι η διαφορά ανάμεσα στην πραγματική και την προβλεπόμενη έξοδο που απεικονίζεται ενδεικτικά στο παρακάτω παράδειγμα




*Εικόνα 6: $\xi_t$ - **η** διαφορά ανάμεσα στην πραγματική και την προβλεπόμενη έξοδο*

Προκειμένου να μπορέσουμε εφαρμόσουμε το θεώρημα Kuhn –Tucker και να εισάγουμε στη συνάρτηση ελαχιστοποίησης του w τους πολλαπλασιαστές Lagrange .

Σύμφωνα με το θεώρημα Kuhn –Tucker και εφόσον ισχύουν οι (1.5) και (1.14) υπάρχει τιμή για τους πολλαπλασιαστές Lagrange τέτοια ώστε η ελαχιστοποίηση της (1.5) να ισοδυναμεί με την ελαχιστοποίηση της

$$L_T(w) = l \,||W||^2 + \sum_{t=1}^{T} \xi_t^2 + \sum_{t=1}^{T} (y_t - wx_t - \xi_t) \quad (1.15)$$

Για να ελαχιστοποιήσουμε ως προς w την (1.15) την διαφορίζουμε ως προς w , οπότε προκύπτει ότι:

$$w = \frac{1}{2l} \sum_{t=1}^{T} a_t x_t \quad (1.16)$$

*Εξίσωση 1.16:Παρατηρώντας την παραπάνω σχέση αξίζει να σημειώσουμε ότι οι πολλαπλασιαστές Lagrange μας δείχνουν τον βαθμό στον οποίο συμβάλλουν τα δείγματα στα οποία αντιστοιχούν στην τελική λύση-κατάταξη. Ακόμα η μορφή αυτή του w θα μας διευκολύνει παρακάτω στην εισαγωγή μιας τεχνική που απλοποιεί σημαντικά τους αναγκαίους υπολογισμούς του αλγορίθμου.*

Είναι φανερό ότι το διάνυσμα βαρών ανά πάσα στιγμή είναι γραμμικός συνδυασμός των προτύπων. Προκύπτει δηλαδή από το άθροισμα των προτύπων πολλαπλασιασμένων με θετικούς συντελεστές (έστω $a_t$ ) θα μπορούσαμε να θεωρήσουμε τους $a_s$ ως τις παραμέτρους που ψάχνουμε να βρούμε αντί των συναπτικών βαρών . Τα $a_t$ ονομάζονται δυϊκές μεταβλητές των $w_i$ και αντίστροφα. Στην περίπτωσή μας οι μεταβλητές αυτές είναι οι συντελεστές Lagrange.


*Εικόνα 7:Δυϊκές μεταβλητές*

Για να εκμεταλλευτούμε την παρατήρηση αυτή αντικαθιστούμε την (1.16) στην (1.15) και στη συνέχεια διαφορίζουμε ως προς ξt και αντικαθιστούμε ξανά στην (1.16).Οπότε καταλήγουμε στην

$$\frac{1}{4l}\sum_{s,t=1}^{T} a_s a_t (x_s x_t) - \frac{1}{4}\sum_{t=1}^{T} a_t^2 + \sum_{t=1}^{T} y_t a_t \quad (1.17)$$

Αξίζει να παρατηρήσουμε ότι στη σχέση 1.17 τα δεδομένα της εισόδου εμφανίζονται υπό τη μορφή **εσωτερικού γινομένου**. Ακόμα παρατηρούμε το διάνυσμα των βαρών δεν υπεισέρχεται καν στους αναγκαίους υπολογισμούς. Μας αρκεί η γνώση των δυικών συντελεστών $a_t$

Ας δούμε όμως σε τι μας χρησιμεύουν οι παραπάνω παρατηρήσεις, για να κατανοήσουμε γιατί εισάγουμε τις δυϊκές μεταβλητές- μεταβλητές Lagrange στον αλγόριθμό Ridge Regression.

Στην πλειοψηφία των περιπτώσεων δεν είναι δυνατό να βρεθεί αναλυτικά η λύση που προσδιορίζεται μονοσήμαντα από τις συνθήκες Kuhn-Tucker,οπότε

είμαστε υποχρεωμένοι να λύσουμε αριθμητικά-επαναληπτικά το πρόβλημα της βελτιστοποίησης με κατάλληλους αλγόριθμους .Στην ουσία εκμεταλλευόμαστε τις συνθήκες αυτές για να επαναδιατυπώσουμε το πρόβλημα σε συνάρτηση με τις δυϊκές μεταβλητές μόνο. Το πρόβλημα έχει έτσι απλούστερη μορφή. Το δυϊκό πρόβλημα αποτελεί την αφετηρία για την εφαρμογή επαναληπτικών αριθμητικών μεθόδων για την εύρεση διαχωριστικού υπερεπιπέδου και των βαρών. Τέλος τα πρότυπα εισέρχονται τόσο στη Lagrangian, όσο και στον κανόνα ταξινόμησης, υπό μορφή εσωτερικού γινομένου. Αυτό είναι πολύ σημαντικό για τη ***γενίκευση της Ridge Regression για ταξινόμηση μη γραμμικά διαχωρίσιμων δεδομένων.***

**Τελικά:**

| | |
|---|---|
| **Συνάρτηση κόστους:** | $L_T(w) = l \|W\|^2 + \sum_{t=1}^{T} \xi_t^2 + \sum_{t=1}^{T} (y_t - wx_t - \xi_t)$ |
| **Εκπαίδευση-εύρεση w:** | $w = \frac{1}{2l} \sum_{t=1}^{T} a_t x_t$ |
| **Πρόβλεψη νέων τιμών:** | $y_{new} = wx_{new}$ |

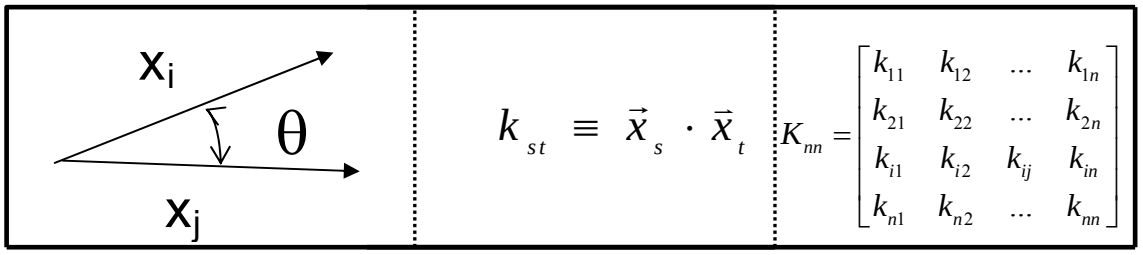
## *2.7 Η εισαγωγή της τεχνικής του πυρήνα (Kernel Ridge Regression – KRR )*

Παρατηρούμε ότι λόγω των πολλών διαστάσεων η πολυπλοκότητα των υπολογισμών στις σχέσεις που έχουμε μέχρι τώρα περιγράψει είναι μεγάλη. Ακόμα παρατηρούμε ότι η συνάρτηση του ridge regression όσο και η συνάρτηση κόστους δεν εξαρτώνται από τα ίδια τα δείγματα αλλά από τα εσωτερικά γινόμενα των δειγμάτων μας. Έτσι χρησιμοποιώντας αυτήν την παρατήρηση φτιάχνω συναρτήσεις που αντιστοιχίζουν των χώρο εισόδων στον χώρο καταστάσεων , βοηθώντας να κατανικήσουμε το πρόβλημα της πολυπλοκότητας των υπολογισμών. Οι συναρτήσεις αυτές μπορεί να είναι διαφόρων τύπων (πολυωνυμικές , εκθετικές..) και ονομάζονται συναρτήσεις πυρήνα.

Πιο συγκεκριμένα:

Ορίζουμε τον πίνακα - πυρήνα K που περιέχει τα εσωτερικά γινόμενα των διανυσμάτων των χαρακτηριστικών. Ο πυρήνας για τα διανύσματα $x_s, x_t$ είναι $K_{s,t} = x_s \cdot x_t$ (1.18), s=1,2,..,n και t=1,2,..,..n

ή καλύτερα( επειδή πρόκειται για διανύσματα)




*Εικόνα 8: Σχηματική απεικόνιση πυρήνα. Ο K είναι ένας τετραγωνικός πίνακας n x n διαστάσεων που περιέχει όλα τα εσωτερικά γινόμενα του διανύσματος χαρακτηριστικών των δειγμάτων προς εκπαίδευση μεταξύ τους .Θα περιγράψουμε στην επόμενη παράγραφο λίγο πιο αναλυτικά την αξία της χρησιμοποίησης του πυρήνα*

Τελικά ο πυρήνας θα ναι ένας πίνακας θα ναι διαστάσεων nxn και θα έχει τη μορφή του $K_{nn}$ (Εικόνα 8) .

Αντικαθιστώντας την (1.18) στην (1.17) και διαφορίζοντας ως προς $a_t$ προκύπτει η (1.19)

$a = 2l(K + lI)^{-1} y$ (1.19) (Εύρεση συντελεστών Lagrange)

W= $\frac{1}{2l}\sum_{t=1}^{T} a_t x_t = (K + lI)^{-1} yx$ (εκπαίδευση) (1.20)

Στην παραπάνω σχέση y,x είναι το διάνυσμα των κλάσεων και ο πίνακας διανυσμάτων τιμών των χαρακτηριστικών κάθε ασθενή αντίστοιχα για το training set.Χρησιμοποιούμε δηλαδή πια τις συναρτήσεις πυρήνα των οποίων ο υπολογισμός είναι πιο εύκολος, για να υπολογίσουμε τους συντελεστές Lagrange , το W(διαδικασία εκπαίδευσης) και στη συνέχεια να κάνουμε τις προβλέψεις .

**Τελικά :**

**Συνάρτηση κόστους**: $L_T$ (w)= $\frac{1}{4l}\sum_{s,t=1}^{T} a_s a_t K - \frac{1}{4}\sum_{t=1}^{T} a_t^2 + \sum_{t=1}^{T} y_t a_t$

**Εκπαίδευση-εύρεση** w. W= $\frac{1}{2l}\sum_{t=1}^{T} a_t x_t = (K + lI)^{-1} yx$

y: Διάνυσμα κλάσεων που ανήκουν όλοι οι ασθενείς που περιλαμβάνονται στο training set

x: πίνακας που περιέχει τα διανύσματα με το σύνολο τιμών των χαρακτηριστικών-γονιδίων για κάθε ασθενή που ανήκει στο training set

**Πρόβλεψη νέων τιμών**: $y_{new} = wx_{new} = (K + lI)^{-1} yxx_{new}$

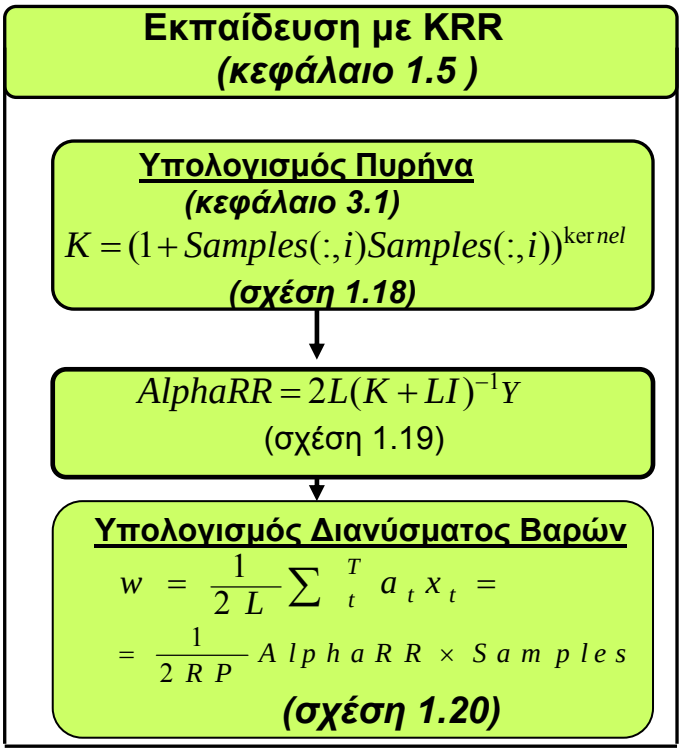
$= (K + lI)^{-1} yK = y'(K + lI)^{-1} k$

$x_{new}$ : οι τιμές των χαρακτηριστικών των νέων δειγμάτων-ασθενών που θέλω να ταξινομήσω

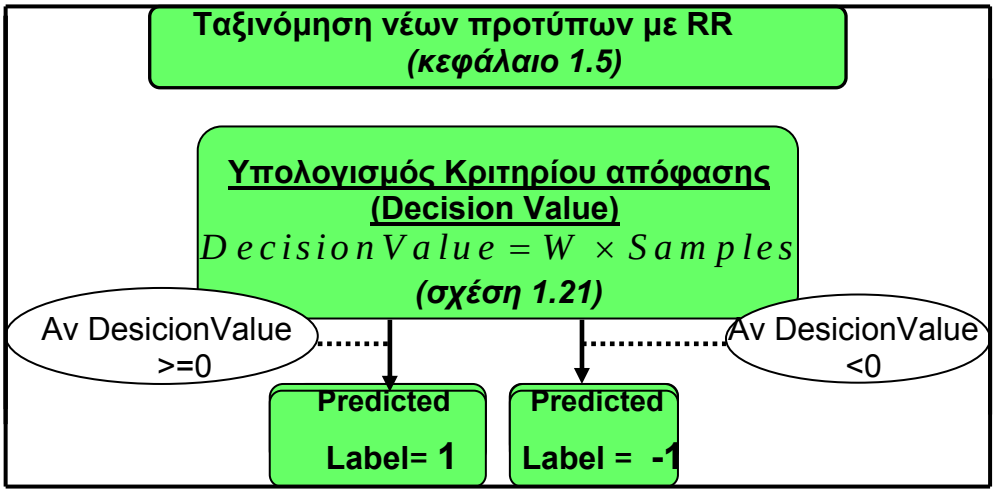
$y_{new}$ : η κλάση στην οποία προβλέπουμε ότι ανήκει το νέο δείγμα

K : $(k_1, k_2, ..k_n)$είναι ο πίνακας των εσωτερικών γινομένων(dot products) των διανυσμάτων του training set $K_{s,t} = K(x_s x_t)$,s=1,..n,t=1,..n

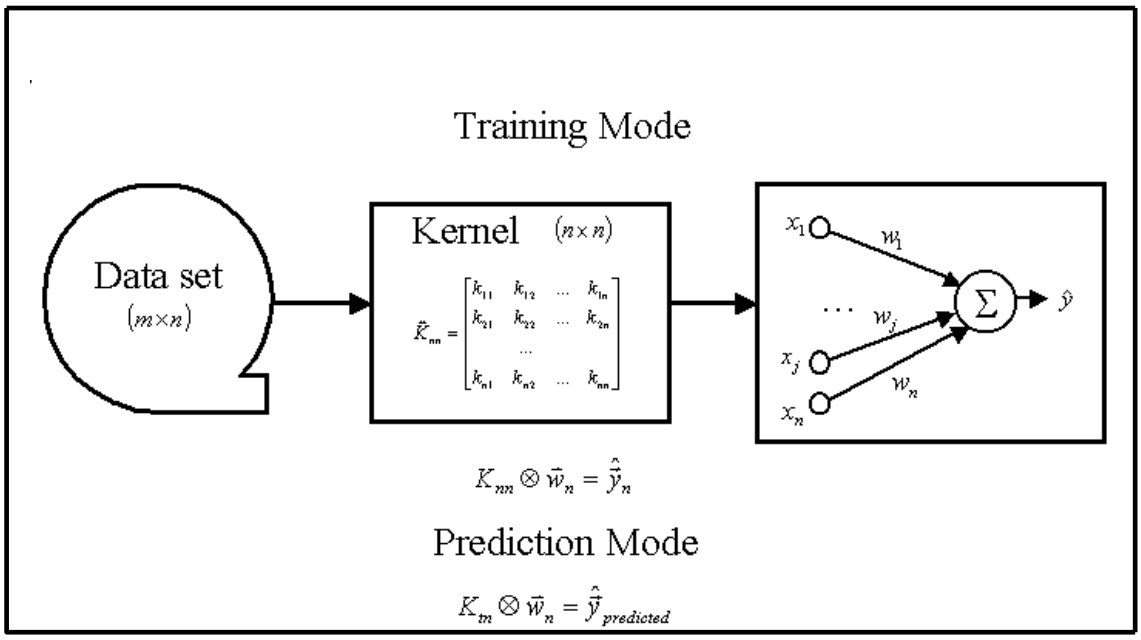
Στις εικόνες 9,10 βλέπουμε τη διαδικασία εκπαίδευσης και ταξινόμησης




*Εικόνα 9 : Σχηματική απεικόνιση της διαδικασίας εκπαίδευσης*




Εικόνα 10 : *Σχηματική απεικόνιση της διαδικασίας ταξινόμησης*




*Εικόνα 11 : Ο πυρήνας στη διαδικασία της εκπαίδευσης και ταξινόμησης*.

Στην Παραπάνω εικόνα φαίνεται πιο συγκεκριμένα και ο ρόλος του πυρήνα κατά τη φάση της εκπαίδευσης (training Mode ) και ταξινόμησης-πρόβλεψης (Prediction Mode) .

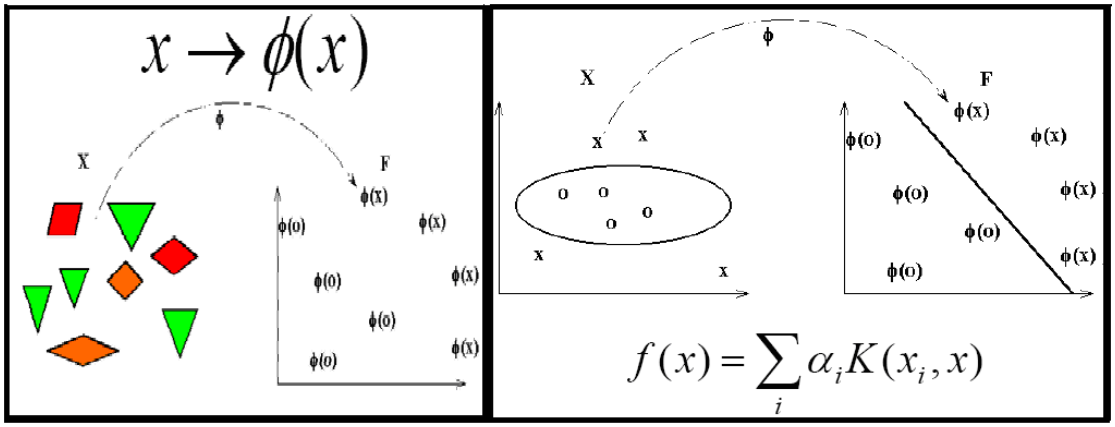
## *2.8 Ο ρόλος της συνάρτησης πυρήνα*

Ορίσαμε στην προηγούμενη παράγραφο έναν πίνακα K , τον οποίο ονομάσαμε πυρήνα και ο οποίος ορίζεται ως το εσωτερικό γινόμενο(dot product) των πινάκων $\mathrm{X}_s, \mathrm{X}_t$ .

Θα ορίσουμε εδώ την συνάρτηση K( $x_i, x_j$ )που είναι απλά μία συνάρτηση που επιστρέφει το εσωτερικό γινόμενο (dot products) των διανυσμάτων $x_i, x_j$ που μας δίνονται ως δεδομένα.. Η *K*(.) ονομάζεται ΣΥΝΑΡΤΗΣΗ ΠΥΡΗΝΑ. Ουσιαστικά ,η συνάρτηση υπολογίζει τον πίνακα με καθένα από τα εσωτερικά γινόμενα που απαιτούνται για την εκπαίδευση απευθείας χωρίς να χρειαστεί να υπολογιστούν τα αθροίσματα, που μπορεί στην πραγματικότητα να είναι και άπειρες σειρές ή ολοκληρώματα σε ορισμένες περιπτώσεις. Η πολυπλοκότητα του υπολογισμού του πίνακα των εσωτερικών γινομένων που απαιτούνται για τη διαδικασία της εκπαίδευσης είναι O( $P^2$ ) ανεξάρτητα από τη διάσταση του χώρου των προτύπων, δηλαδή ανεξάρτητα από το πλήθος των χαρακτηριστικών. Αντίστοιχο όφελος υπάρχει και στη φάση της ταξινόμησης.

Δε χρειάζεται να υπολογίζονται τα υψηλής (ίσως και άπειρης) πολυπλοκότητας εσωτερικά γινόμενα. Πάλι, η εύρεση της κλάσης στην οποία ανήκει ένα νέο πρότυπο, είναι ανεξάρτητη από τη διάσταση του χώρου των προτύπων., δηλαδή από το πλήθος των χαρακτηριστικών. Ο πυρήνας είναι αρκετός τόσο για την εκπαίδευση όσο και για την ταξινόμηση. Δεν είναι καν απαραίτητο να γνωρίζουμε τις συναρτήσεις Φ. Μας αρκεί ο πυρήνας.

Στην παρακάτω εικόνα μπορούμε να παρατηρήσουμε τη βασική ιδέα της τεχνικής του πυρήνα.




Εικόνα 12:*Βασική ιδέα μεθόδων ταξινόμησης με πυρήνα. Αρχικά ενσωματώνουμε τα δεδομένα στον διανυσματικό χώρο των εισόδων και στη συνέχεια ψάχνουμε για γραμμικές συσχετίσεις σε αυτόν τον χώρο.*

Η σχέση (1.21) αντιστοιχεί στην υλοποίηση γραμμικής παλινδρόμησης μέσα στο χώρο εισόδων $\mathbb{R}^n$ που ορίζεται από τα δεδομένα. Για να υλοποιήσουμε όμως τη γραμμική παλινδρόμηση σε κάποιον χώρο δεδομένων $\mathbb{R}^n$, πρέπει πρώτα να επιλέξουμε κάποιον τρόπο αντιστοίχισης από τον αρχικό χώρο των εισόδων X σε έναν μεγαλύτερων διαστάσεων χώρο δεδομένων F $(\phi : X \to F)$.Για να μπορέσουμε να χρησιμοποιήσουμε τον τύπο (1.21) για να υλοποιήσουμε την παλινδρόμηση στον χώρο δεδομένων, η συνάρτηση K που ορίσαμε παραπάνω πρέπει να ανταποκρίνεται στο εσωτερικό γινόμενο $\phi(x_i) \cdot \phi(x_j)$.Δεν είναι απαραίτητο να γνωρίζουμε ακριβώς ποια είναι η $\phi(x)$.Μας αρκεί να γνωρίζουμε την K( $x_i, x_j$ )= $\phi(x_i) \cdot \phi(x_j)$.Για να θεωρείται μια συνάρτηση ικανή να αντιστοιχίσει σε κάποιον χώρο δεδομένων μέσω ενός

εσωτερικό γινομένου ,αρκεί να πληρεί τις προϋποθέσεις του θεωρήματος Mercer's.
Για να γίνει πιο κατανοητή η τεχνική του πυρήνα, παρουσιάζουμε παρακάτω το παράδειγμα μιας απλής συνάρτησης πυρήνα.

Ας υποθέσουμε ότι υπάρχει μια συνάρτησης αντιστοίχισης $\phi$ που αντιστοιχεί ένα διάνυσμα δύο διαστάσεων σε χώρο 6 διαστάσεων όπως φαίνεται παρακάτω: $\phi:(x^1,x^2)\mapsto((x^1)^2,(x^2)^2,\sqrt{2}x^1,\sqrt{2}x^2,\sqrt{2}x^1x^2,1)$,

Στην περίπτωση αυτή αν υπολογίσουμε τα εσωτερικά γινόμενα στην F θα δούμε:

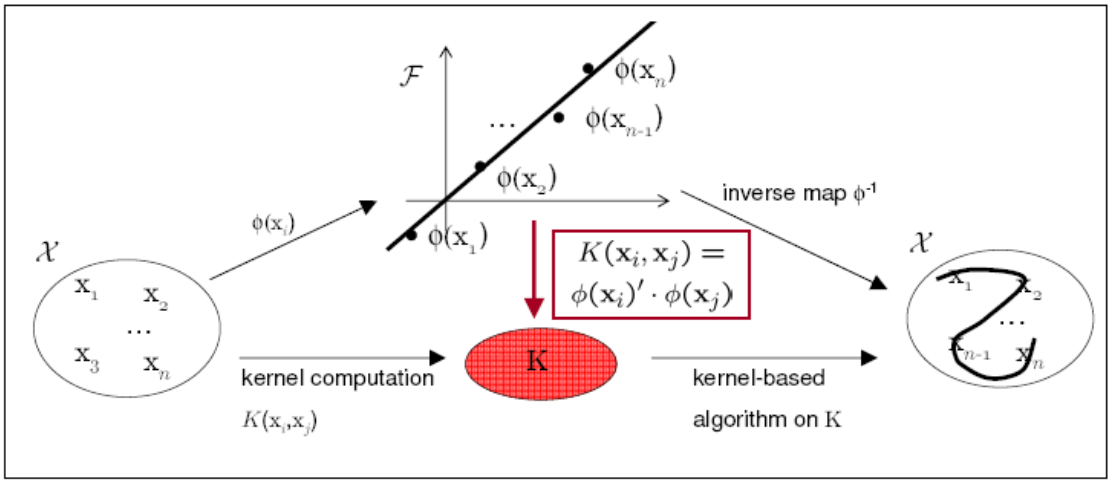
$$(\phi(x)\cdot\phi(y)) = (x^1)^2(y^1)^2+(x^2)^2(y^2)^2+2x^1y^1+2x^2y^2+2x^1y^1x^2y^2+1$$
$$= ((x\cdot y)+1)^2 \quad (1.22)$$

Από τη σχέση (1.22) εύκολα συμπεραίνουμε ότι μια πιθανή συνάρτηση πυρήνα για την περίπτωση που περιγράψαμε είναι η $K(x,y)=((x\cdot y)+1)^2$.Η τελευταία μπορεί ακόμα να γίνει πιο γενική : $K(x,y)=((x\cdot y)+1)^d$ και να αναφέρεται σε περισσότερες διαστάσεις.

Η χρήση συναρτήσεων πυρήνα μας επιτρέπει να κατασκευάσουμε μία γραμμική συνάρτηση πυρήνα σε ένα χώρο χαρακτηριστικών μεγάλων διαστάσεων (που αντιστοιχεί σε μη γραμμική παλινδρόμηση στον χώρο εισόδων).Με τον τρόπο αυτό αποφεύγουμε να κάνουμε υπολογισμούς σε χώρο μεγάλων διαστάσεων που θα ήταν πολύ πολύπλοκοι.
Η τεχνική του πυρήνα είναι τελικά ένα πού χρήσιμο εργαλείο που μας βοηθά με απλούς υπολογισμούς να βρίσκουμε πρότυπα για ένα μεγάλο φάσμα δεδομένων. Αυτό συμβαίνει ,γιατί:

• Αντιστοιχούν τα δεδομένα από το χώρο εισόδων x σε έναν διανυσματικό χώρο πολύ μεγαλύτερων διαστάσεων τον χώρο χαρακτηριστικών (Feature Space) με την εφαρμογή της συνάρτησης χάρτη των χαρακτηριστικών Φ στα δείγματα.

• Με την εφαρμογή της αντίστροφής συνάρτησης χάρτη μια μη γραμμική συνάρτηση μπορεί να μετατραπεί σε γραμμική στον χώρο χαρακτηριστικών.
• Όλα αυτά επιτυγχάνονται μόνο από το γεγονός ότι πλέον λόγω του πυρήνα η πρόβλεψη εξαρτάται από το εσωτερικό γινόμενο και όχι από τα ίδια τα δεδομένα. Οι υπολογισμοί κατά συνέπεια είναι πολύ πιο απλοί και γρήγοροι




*Εικόνα 13 : η τεχνική του πυρήνα*

# Κεφάλαιο3: Μείωση γονιδίων με τη μέθοδο Kernel Ridge Regression

Εκτός από τη διαδικασία της εκπαίδευσης και της ταξινόμησης ο αλγόριθμός μας περιλαμβάνει και τη διαδικασία της επιλογής και αφαίρεσης των γονιδίων που θεωρούνται λιγότερο σημαντικά προκειμένου να καταλήξουμε σε αυτά που επηρεάζουν περισσότερο τον καθορισμό της κλάσης –εξόδου στην οποία ανήκει το κάθε δείγμα. Αρχικά είναι απαραίτητο να επιλέξουμε ένα μέτρο σύγκρισης της σημαντικότητας του κάθε γονιδίου. Είναι βασικό εδώ να ορίσουμε ότι σημαντικότερο από κάποιο άλλο θεωρούμε εκείνο το γονίδιο που συνεισφέρει περισσότερο στον καθορισμό της κλάσης εξόδου.

Πρόσφατες μελέτες έχουν δείξει ότι η microarray γονιδιακή έκφραση των δεδομένων (microarray gene expression data) μπορεί να φανεί χρήσιμη στην ταξινόμηση (classification) φαινοτύπων σε πολλές περιπτώσεις ασθενειών. Σε αυτό το πρόβλημα ο αριθμός των χαρακτηριστικών (στη συγκεκριμένη περίπτωση των γονιδίων) ξεπερνά σε πολύ μεγάλο βαθμό το μέγεθος των στιγμιότυπων (δείγματα ιστών). Το γεγονός αυτό μπορεί σε πολύ μεγάλο βαθμό να βλάψει την απόδοση της ταξινόμησης καρκινικών ιστών , όπως επίσης και να αυξήσει το κόστος. Αυτό , γιατί στην περίπτωση μας, θα μπορούσαμε να προσδιορίσουμε εύκολα έναν ταξινομητή (ακόμα και γραμμικό) που να διαχωρίζει τα δεδομένα εκπαίδευσης , ο οποίος όμως θα εμφάνιζε πολύ κακή απόδοση σε νέα δεδομένα. Το πρόβλημα αυτό είναι γνωστό ως πρόβλημα overfitting . Η χρήση κανονικοποιημένων μεθόδων (Ridge Regression, SVMs) βοηθά σημαντικά στην αντιμετώπιση του προβλήματος αυτού. Έχει επίσης αποδειχθεί, ότι επιλέγοντας ένα μικρό μέγεθος γονιδίων (informative genes) μπορεί να οδηγήσει σε βελτιωμένα αποτελέσματα ακρίβειας. Το παραπάνω πρόβλημα μπορεί να οριστεί ως **πρόβλημα γονιδιακής επιλογής (gene selection problem).** Η γονιδιακή επιλογή περιλαμβάνει την αναζήτηση των υποσυνόλων ομάδων γονιδίων που μπορούν να διακρίνουν τον καρκινικό ιστό από τον κανονικό και μπορούν να έχουν κάποια ευθεία ή έμμεση συμμετοχή στο μοριακό μηχανισμό της ογκογένεσης .

Τα πρακτικά πλεονεκτήματα της γονιδιακής επιλογής, τα οποία ισχύουν ακόμα και σε σχέση με κάποιες άλλες μεθόδους μείωσης των διαστάσεων ενός αρχικού συνόλου δεδομένων /dataset (principal component analysis κ.ά.) είναι:

- η μείωση της πολυπλοκότητας. Παρόλο που δύο χαρακτηριστικά (γονίδια) μπορούν να φέρουν σημαντική πληροφορία για την ταξινόμηση όταν αυτά τα μεταχειριζόμαστε ξεχωριστά, το κέρδος είναι πολύ μικρό όταν συνδυάζουμε τα δύο χαρακτηριστικά μαζί σε ένα διάνυσμα εξαιτίας της υψηλής αμοιβαίας συσχέτισης. Έτσι με τη γονιδιακή επιλογή μπορούμε να μειώσουμε την πολυπλοκότητα με πολύ μικρή μείωση του κέρδους.

- η μείωση του κόστους υπολογισμού, αφού στη θέση ενός τεράστιου συνόλου δεδομένων (dataset) υπάρχει ένα σαφώς πολύ μικρότερο σύνολο για επεξεργασία, μειώνοντας έτσι, τις απαιτήσεις για τις μετρήσεις και την αποθήκευση των αποτελεσμάτων.

- η διατήρηση της πολύ βασικής ιδιότητας για την ταξινόμηση (classification) της γενίκευσης (generalization), αφού όσο μεγαλύτερη είναι η αναλογία των εκπαιδευόμενων δειγμάτων προς τις ελεύθερες παραμέτρους του ταξινομητή, τόσο καλύτερο είναι το αποτέλεσμα της ταξινόμησης που προκύπτει. Ένας μεγάλος αριθμός χαρακτηριστικών (γονιδίων) μεταφράζεται ευθέως σε μεγάλο αριθμό παραμέτρων ταξινόμησης (π.χ. τα βάρη σε έναν γραμμικό ταξινομητή ή τα συναπτικά βάρη σε ένα νευρωνικό δίκτυο). Έτσι, για έναν περιορισμένο αριθμό δειγμάτων ιστών, μειώνοντας τον αριθμό των γονιδίων όσο περισσότερο γίνεται, μπορούμε να επιτύχουμε καλύτερη γενίκευση στην σχεδιαζόμενη ταξινόμηση.

- η μεγαλύτερη πιθανότητα υιοθέτησης του μοντέλου σε κλινικό περιβάλλον. Τα επιλεχθέντα υποσύνολα γονιδίων με την μεγάλη ακρίβεια στην ταξινόμησης, είναι πιθανόν να εμπλέκονται σε έναν βαθμό στη διαδικασία ανάπτυξης του όγκου. Έτσι, τα υποσύνολα γονιδίων που

επιλέγουμε είναι δυνατόν να έχουν σημαντική βιολογική αξία και μπορούν να αποτελέσουν πεδίο έρευνας για την επιστήμη της Βιολογίας στην προσπάθεια της θεραπείας, αλλά και της πρόληψης του καρκίνου.

- Η βελτίωση του λάθους ταξινόμησης (classification error), αφού όπως φαίνεται και παραπάνω, ένα πολύ σημαντικό βήμα στο σχεδιασμό της ταξινόμησης είναι το στάδιο της εκτίμησης της απόδοσης, όπου υπολογίζεται η πιθανότητα σφάλματος του ταξινομητή.

Το πρόβλημα της επιλογής χαρακτηριστικών αντιμετωπίζεται με διάφορες μεθόδους .Για μία δεδομένη μέθοδο ταξινόμησης , μπορεί να γίνει επιλογή ενός υποσυνόλου χαρακτηριστικών που ικανοποιούν ένα δεδομένο κριτήριο επιλογής μετά από εξαντλητική διερεύνηση όλων των δυνατών υποσυνόλων χαρακτηριστικών. Η εξαντλητική απαρίθμηση όλων των υποσυνόλων για πολύ μεγάλο αριθμό χαρακτηριστικών (στην περίπτωση μας μερικές χιλιάδες γονιδίων) δεν είναι καθόλου πρακτική , μιας και ο συνδυασμός του ήδη μεγάλου αριθμού χαρακτηριστικών θα εκτοξεύσει τον αριθμό των υποσυνόλων. Η διεξαγωγή της επιλογής χαρακτηριστικών για χώρο πολύ μεγάλων διαστάσεων προϋποθέτει άλλες μεθόδους. Ανάμεσα σε διάφορες πιθανές μεθόδους οι τεχνικές που βασίζονται στην κατάταξη των χαρακτηριστικών είναι ιδιαίτερα δημοφιλείς. Πιο συγκεκριμένα ένας προκαθορισμένος κάθε φορά αριθμός χαρακτηριστικών επιλέγονται από τη λίστα κατάταξης των συνόλου των χαρακτηριστικών. Το υποσύνολο αυτό των χαρακτηριστικών χρησιμοποιείται για την κατασκευή του ταξινομητή ή για παραπέρα ανάλυση. Εναλλακτικά, μπορούμε να θέσουμε ένα συγκεκριμένο όριο στην τιμή του κριτηρίου κατάταξης των χαρακτηριστικών. Στην περίπτωση αυτή δεν είμαστε εμείς που προκαθορίζουμε το πλήθος των χαρακτηριστικών που θα επιλεχτούν .Σε κάθε περίπτωση βασικό ρόλο διαδραματίζει το κριτήριο με βάση το οποίο γίνεται η κατάταξη των χαρακτηριστικών αλλά και η μέθοδος με την οποία το αξιοποιούμε. Στην προσέγγιση που ακολουθεί επιλέξαμε να προκαθορίσουμε εμείς τον αριθμό των επιλεγόμενων γονιδίων, προκριμένου να μπορούμε να ελέγξουμε την απόδοση του αλγορίθμου για διάφορα μεγέθη υποσυνόλων .

## *3.1 Κριτήριο κατάταξης γονιδίων και ταξινόμηση*

Στα προβλήματα ταξινόμησης που εξετάζουμε , είναι αδύνατο να πετύχουμε διαχωρισμό χωρίς λάθη με ένα και μόνο γονίδιο . Καλύτερα αποτελέσματα επιτυγχάνονται καθώς αυξάνουμε τον αριθμό των γονιδίων. Οι πιο κλασσικές μέθοδοι επιλογής γονιδίων επιλέγουν τα γονίδια που κατά μονάδα ταξινομούν καλύτερα τα δεδομένα εκπαίδευσης. Αυτές οι μέθοδοι περιλαμβάνουν μεθόδους συσχέτισης και μεθόδους έκφρασης ποσοστού. Σβήνουν τα γονίδια που είναι άχρηστα για τον διαχωρισμό (θόρυβος). Ενδεικτικά αναφέρουμε τον συντελεστής Golub(1999) που ορίζεται ως:

$w_i = (\mu_i(+) - \mu_i(-))/(\sigma_i(+) - \sigma_i(-))$ (1.23)

Όπου $\mu_i$ και $\sigma_i$ είναι η μέση και τυπική απόκλιση των τιμών από τη γονιδιακή έκφραση του γονιδίου i.

Αυτό που γενικά χαρακτηρίζει τις μεθόδους συσχέτισης είναι το ότι κάθε συντελεστής εκφράζει τη συσχέτιση του κάθε μεμονωμένου γονιδίου στην ταξινόμηση των δειγμάτων , χωρίς αναλαμβάνει υποψιών αμοιβαίες πληροφορίες γονιδίων.

Η εκτίμηση του μεγέθους της συνεισφοράς κάθε γονιδίου στο διαχωρισμό μπορεί να δημιουργήσει μία απλή μέθοδο κατάταξης χαρακτηριστικών. Διάφοροι συντελεστές συσχέτισης χρησιμοποιούνται ως κριτήρια κατάταξης.

Μία πιθανή αξιοποίηση της διαδικασίας κατάταξης των χαρακτηριστικών-γονιδίων είναι η κατασκευή ενός ταξινομητή βασισμένου σε ένα προ-επιλεγμένο υποσύνολο του συνολικού αριθμού των χαρακτηριστικών (γονιδίων). Κάθε χαρακτηριστικό που είναι συσχετισμένο με το διαχωρισμό που μας ενδιαφέρει μπορεί να αποτελέσει από μόνο του έναν –έστω και ατελή- προβλεπτή εξόδου-κλάσης. Με τον τρόπο αυτό προκύπτει μια απλή μέθοδος ταξινόμησης που βασίζεται στη σταθμισμένη συνεισφορά κάθε χαρακτηριστικού: κάθε χαρακτηριστικό συνεισφέρει ανάλογα με το συντελεστή συσχέτισής του .Τέτοια είναι η μέθοδος Golub(1999).Έτσι προκύπτει ένας συγκεκριμένος γραμμικός ταξινομητής:

$D(x) = w \cdot (x - \mu)$ (1.24)

Όπου $w = (\mu(+) - \mu(-))/(\sigma(+) - \sigma(-))$ (1.25) και $\mu = (\mu(+) - \mu(-))/2$ . (1.26)

Αποδείξαμε παραπάνω ότι οι συντελεστές συσχέτισης και κατάταξης των χαρακτηριστικών μπορούν να χρησιμοποιηθούν σαν τα βάρη του ταξινομητή. Επιπλέον ισχύει και το αντίστροφο. **Τα βάρη w που πολλαπλασιάζουν τις εισόδους-χαρακτηριστικά για έναν δεδομένο ταξινομητή μπορούν να χρησιμοποιηθούν ως κριτήρια ταξινόμησης.** Οι είσοδοι που πολλαπλασιάζονται μεγαλύτερα w επιδρούν περισσότερο στην απόφαση του ταξινομητή. Για το λόγο αυτό , σε έναν καλό ταξινομητή, οι είσοδοι που πολλαπλασιάζονται με τα μεγαλύτερα βάρη αντιστοιχούν στα πιο σημαντικά γονίδια(most informative genes).

## *3.2 Κατάταξη γονιδίων με ανάλυση ευαισθησίας*

Σε αυτήν την ενότητα θα αποδείξουμε πιο συγκεκριμένα ότι η κατάταξη των χαρακτηριστικών βάσει του μέτρου των βαρών ενός ταξινομητή είναι βάσιμη. Από πολλούς έχει προταθεί η χρήση της μεταβολής που παρατηρείται στην αντικειμενική συνάρτηση του αλγορίθμου ταξινόμησης όταν ένα χαρακτηριστικό απομακρύνεται, σαν κριτήριο κατάταξης . Για τα προβλήματα ταξινόμησης , η ιδανική αντικειμενική συνάρτηση εκφράζει το αναμενόμενο ποσοστό λάθους. Αυτό προκύπτει από το ποσοστό λάθους που υπολογίζεται από έναν μη πεπερασμένο αριθμό δειγμάτων. Για τους σκοπούς της διαδικασίας της εκπαίδευσης αυτή η εξιδανικευμένη συνάρτηση(αντικειμενική συνάρτηση-objective function) αντικαθίσταται από τη συνάρτηση κόστους (cost function) J που υπολογίζεται από έναν πεπερασμένο αριθμό δειγμάτων-το σύνολο εκπαίδευσης. Έτσι θα μπορούσε να χρησιμοποιηθεί ως κριτήριο κατάταξης η μεταβολή στην συνάρτηση κόστους J που προκύπτει από την αφαίρεση ενός συγκεκριμένου χαρακτηριστικού. Αυτή η μεταβολή μπορεί εύκολα να υπολογιστεί με τη βοήθεια παραγώγων και σειρών Taylor (OBD αλγόριθμος- LeCun ,1990):

$$DJ(i) = (1/2)\frac{\partial^2 J}{\partial w_i^{\,2}}(Dw_i)^2 \quad (1.27)$$

Όπου $Dw_i$ είναι η μεταβολή στο βάρος που αντιστοιχεί στην αφαίρεση του χαρακτηριστικού i. Η χρήση του παραπάνω αλγορίθμου , δηλαδή η χρήση του DJ(i) ως κριτηρίου κατάταξης είναι πολλές φορές προτιμότερη της χρήσης του μέτρου των βαρών .Δεν ισχύει όμως το ίδιο για ταξινομητές των οποίων η συνάρτηση κόστους είναι τετραγωνική συνάρτηση του w.Στην περίπτωση αυτή οι δύο μέθοδοι είναι ισοδύναμες. Τέτοιοι ταξινομητές είναι η ο ταξινομητής ελαχίστων τετραγώνων (Duda 1973) με συνάρτηση κόστους $J = \sum_{x \in X} || w \cdot x - y ||^2$ (1.28) ,ο γραμμικός ταξινομητής με χρήση διανυσμάτων υποστήριξης (Boser 1992;Vapnik,1998;Cristianini,1999) ,καθώς και η μέθοδος παλινδρόμησης κορυφής (Ridge Regression).

Γι αυτό στην εργασία καταλήξαμε στην αξιοποίηση των βαρών του αλγορίθμου Kernel Ridge Regression ως κριτήριο κατάταξης των γονιδίων. Υπενθυμίζουμε εδώ τον τύπο $w = \frac{1}{2l} \sum_{t=1}^{T} a_t x_t$ (1.16) που μας αποκαλύπτει πως για τον καθορισμό του w σημαντικό ρόλο παίζουν οι συντελεστές Lagrange.Αποτελούν ουσιαστικά το βαθμό συμμετοχής κάθε γονιδίου, μιας και η μεταβλητή l είναι ίδια για όλα τα γονίδια(διανύσματα x).Ακόμα η κλάση στην οποία ανήκε ο κάθε ασθενής είναι ανάλογη της τιμής του w: $y_{new} = w x_{new}$(1.21) . Κατά συνέπεια η του W ή καλύτερα η απόλυτη τιμή του w αποτελεί ένα πολύ καλό μέτρο της συνεισφοράς του κάθε γονιδίου στη διαμόρφωση της εξόδου.

## *3.3 Αναδρομική αφαίρεση γονιδίων με Kernel Ridge Regression*

Η επιλογή του κατάλληλου κριτηρίου κατάταξης χαρακτηριστικών διαδραματίζει σημαντικό ρόλο, αλλά δεν αρκεί. Ένα καλό κριτήριο κατάταξης μεμονωμένων χαρακτηριστικών δεν είναι απαραίτητα καλό κριτήριο για την κατάταξη ενός υποσυνόλου. Το κριτήριο κατάταξης DJ(i) ή το $(w_i)^2$ εκτιμούν την επίδραση στην συνάρτηση κόστους από την αφαίρεση ενός μεμονωμένου γονιδίου. Είναι όμως αρκετά αναποτελεσματικά όταν αφαιρούμε περισσότερα γονίδια ταυτόχρονα , πράγμα το οποίο είναι απαραίτητο προκειμένου από

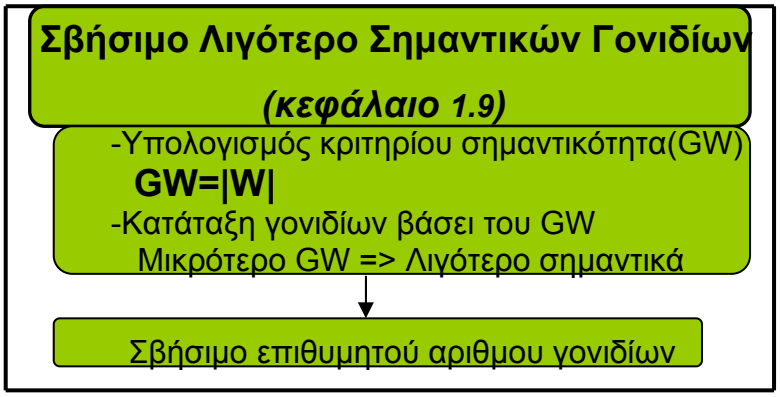
δεκάδες χιλιάδες γονιδίων να καταλήξουμε σε ένα μικρό υποσύνολο. Αυτό το πρόβλημα μπορεί να ξεπεραστεί αν χρησιμοποιήσουμε την ακόλουθη διαδικασία , την όποια και θα ονομάσουμε αναδρομική αφαίρεση χαρακτηριστικών(Recursive Feature Elimination- RFE)(Kohavi ,2000):

1.Εκπαίδευση του ταξινομητή με το σύνολο εκπαίδευσης(προσδιορισμός βέλτιστου w από τη συνάρτηση κόστους).

2.Υπολογισμός του κριτηρίου κατάταξης $(w_i)^2$ για κάθε χαρακτηριστικό

3.Αφαίρεσε το υποσύνολο προεπιλεγμένου μεγέθους γονιδίων που έχουν μικρότερο $(w_i)^2$

Το παρακάτω σχήμα απεικονίζει τη διαδικασία επιλογής και αφαίρεσης γονιδίων:

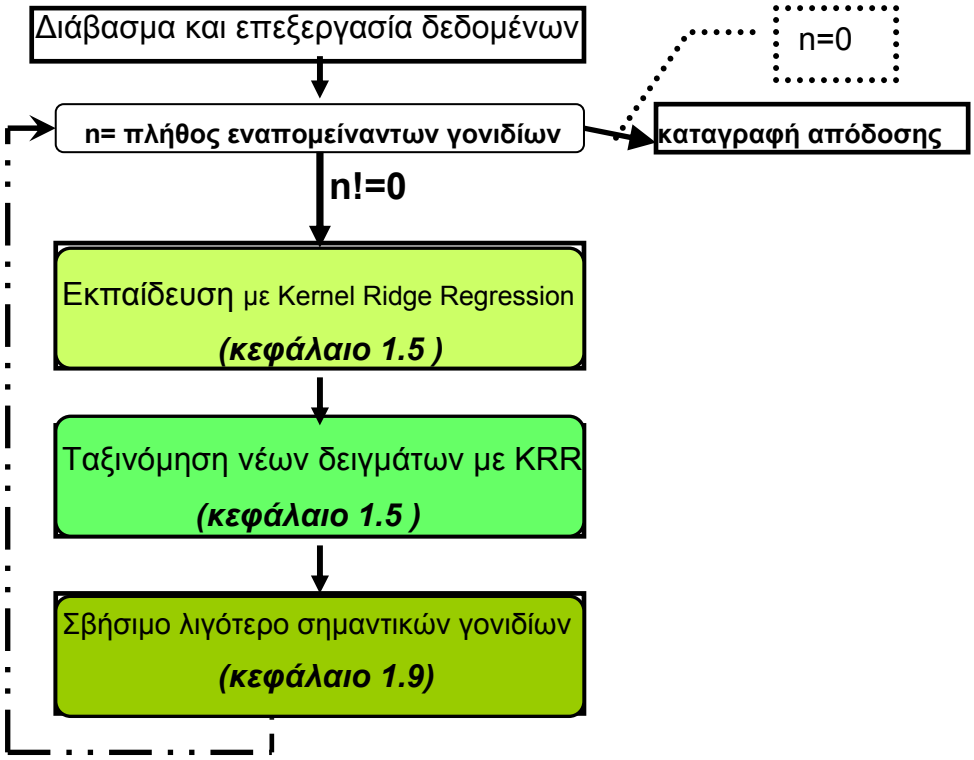



*Εικόνα 14 : Σβήσιμο Λιγότερο σημαντικών γονιδίων*

Ας περιγράψουμε τώρα την διαδικασία επιλογής των γονιδίων (Εικόνα 14) στον αλγόριθμό μας. Μετά τη διαδικασία της εκπαίδευσης και της ταξινόμησης , υπολογίζουμε για κάθε γονίδιο την τιμή του μέτρου σημαντικότητα (|w|) .Βάσει αυτού κατατάσσουμε τα γονίδια και στη συνέχεια απλά διαγράφουμε έναν επιθυμητό αριθμό γονιδίων. Ο αλγόριθμος μας στη συνεχεία επαναλαμβάνει την παραπάνω διαδικασία σβήνοντας διαφορετικό αριθμό γονιδίων κάθε φορά μέχρι να καταλήξουμε στα 0 γονίδια. Σε κάθε επανάληψη καταγράφουμε φυσικά τις επιδόσεις του ,καθώς μεταβάλλουμε την τιμή των παραμέτρων K και I στη σχέση (1.21) του αλγορίθμου. Με τον τρόπο αυτό θα καθορίσουμε τελικά τις βέλτιστες τιμές για τις παραπάνω παραμέτρους, διεκπεραιώνοντας έτσι και έναν από τους βασικούς στόχους της εργασίας αυτής. Οι παράμετροι προς βελτιστοποίηση K και I εκφράζουν τη συνάρτηση

πυρήνα και την παράμετρο κανονικοποίησης (Ridge parameter).Οι παράμετροι αυτοί περιγράφονται αναλυτικότερα στο κεφάλαιο 5.2 .

Ο αλγόριθμος που περιγράφηκε παραπάνω απεικονίζεται συνοπτικά από το παρακάτω block diagram (***Εικόνα 15***Εικόνα 15)




*Εικόνα 15 : Σχηματική απεικόνιση του αλγορίθμου*

# Κεφάλαιο4: Γονιδιακά δεδομένα και πειραματική διαδικασία

Η υλοποίηση του αλγόριθμου παλινδρόμησης κορυφής με χρήση πυρήνα( Kernel Ridge Regression) έγινε με βάση τους τελευταίους τύπους (1.17),(1.20).(1.21) που αφορούν την εκπαίδευση και την ταξινόμηση και τη διαδικασία που περιγράψαμε στην προηγούμενη παράγραφο για την επιλογή και αφαίρεση γονιδίων.

## *4.1 Περιγραφή των συνόλων δεδομένων(datasets)*

Ελέγξαμε τις αποδόσεις του αλγορίθμου για διάφορες τιμές των παραμέτρων RP, πυρήνας εφαρμόζοντας τον αλγόριθμο σε δύο διαφορετικά dataset.

- Πρώτο σύνολο δεδομένων :Ως δεδομένα εισόδου χρησιμοποιήσαμε :δεδομένα εκπαίδευσης που περιείχαν 24188 γονίδια για 79 ασθενείς και δεδομένα εκπαίδευσης που περιείχαν 24188 τιμές γονιδίων για 20 ασθενείς. Πιο συγκεκριμένα τα δεδομένα ήταν σε δύο αρχεία excel κάθε στήλη του οποίου αντιπροσώπευε έναν ασθενή και κάθε γραμμή του ένα συγκεκριμένο γονίδιο. Εξαίρεση αποτελεί η τελευταία γραμμή που περιέχει την κλάση στην οποία ανήκει ο κάθε ασθενής. Εφόσον διαπιστώσαμε κάποιες συγκεκριμένες επιδόσεις παίζοντας με τις παραμέτρους RP και το είδος του πυρήνα, δοκιμάσαμε την απόδοση του αλγορίθμου και σε δεύτερο training και test set
- Δεύτερο σύνολο δεδομένων: Ως δεδομένα εισόδου χρησιμοποιήσαμε :δεδομένα εκπαίδευσης που περιείχαν 7000 γονίδια για 79 ασθενείς και δεδομένα εκπαίδευσης που περιείχαν 7000 τιμές γονιδίων για 20 ασθενείς. Πιο συγκεκριμένα τα δεδομένα ήταν σε δύο αρχεία excel κάθε στήλη του οποίου αντιπροσώπευε έναν ασθενή και κάθε γραμμή του ένα συγκεκριμένο γονίδιο. Εξαίρεση αποτελεί η τελευταία γραμμή που περιέχει την κλάση στην οποία ανήκει ο κάθε ασθενής. Εφόσον διαπιστώσαμε κάποιες συγκεκριμένες επιδόσεις παίζοντας με τις παραμέτρους RP και το είδος του πυρήνα, δοκιμάσαμε την απόδοση του αλγορίθμου και σε δεύτερο training και test set.

Επιπλέον ως είσοδο του αλγορίθμου εισάγουμε και την τιμή της παραμέτρου κανονικοποίησης RP, τη σημασία της οποίας περιγράφουμε αναλυτικά στην παράγραφο 4.2.

Τέλος ο αριθμός γονιδίων που σβήνουμε σε κάθε επανάληψη του αλγορίθμου καθορίζεται επίσης από εμάς. Για τις πρώτες χιλιάδες επαναλήψεων (σβήσιμο χιλιάδων γονιδίων) παρατηρήσαμε γενικά ελάχιστες μεταβολές στις επιδόσεις και γι αυτό επιλέξαμε να σβήνουμε μεγαλύτερο αριθμό γονιδίων .Αντίθετα στις τελευταίες εκατοντάδες παρατηρήσαμε σημαντικές μεταβολές στις επιδόσεις και για να μπορούμε να τις καταγράψουμε , μειώσαμε τον αριθμό γονιδίων που αφαιρούσαμε σε κάθε επανάληψη.)

## *4.2 Βελτιστοποίηση παραμέτρων της KRR*

Ο καθορισμός των βέλτιστων τιμών των παραμέτρων RP καθώς και του πυρήνα που χρησιμοποιούμε είναι ένας από τους βασικούς σκοπούς αυτής της εργασίας. Οι δύο παραπάνω παράμετροι είναι αυτοί που ουσιαστικά επηρεάζουν την απόδοση του αλγορίθμου μας. Παρακάτω θα περιγράψουμε αναλυτικά τη σημασία των δύο αυτών παραμέτρων καθώς επίσης.

**Α)Σταθερά κανονικοποίησης** l:

Υπενθυμίζουμε σε αυτό το σημείο ότι η μεταβλητή RP ή L είναι η σταθερά κανονικοποίησης και επιδρά στον αλγόριθμο

στη διαδικασία της εκπαίδευσης: $W = (K + lI)^{-1} yx$

και στη διαδικασία της πρόβλεψης νέων τιμών : $y_{new} = y'(K + lI)^{-1} K$

Η σταθερά L μπορεί να ονομαστεί και παράμετρος συρρίκνωσης μιας και είναι η παράμετρος που είναι υπεύθυνη για τον περιορισμό της τιμής του w .Υπενθυμίζουμε ότι η χρήση της σταθεράς αυτής προτάθηκε για να αντιμετωπιστεί το πρόβλημα της συγγραμμικότητας καθώς και του overfitting. Ακόμα παραπάνω υπολογίσαμε ότι : $MSE_R = true(Variance) + || bias ||^2$ .Από τον τύπο αυτό συμπεραίνουμε ότι το μέσο τετραγωνικό σφάλμα στη μέθοδο Ridge Regression εξαρτάται άμεσα από την τιμή της διακύμανσης. Η διακύμανση μειώνεται ως ένα σημείο λόγω του ορίου που έχουμε θέσει στο w , αλλά από

ένα όριο και πάνω αυξάνει .Υπενθυμίζουμε ακόμα ότι βάσει των Hoerl και Kennard υπάρχει πάντα τιμή του L τέτοια ώστε $MSE_R < MSE_{LS}$ .Η δυσκολία της μεθόδου ridge regression βρίσκεται και στη δυσκολία προσδιορισμού της παραμέτρου αυτής. Εμείς επιλέξαμε να τρέξουμε τον αλγόριθμο για διάφορες τιμές , προκείμενου να διαπιστώσουμε ποιο είναι το πεδίο τιμών που μας εξασφαλίζει την παραπάνω προϋπόθεση. Οι τιμές RP που δοκιμάσαμε είναι 0.05, 0.5,100:100:1000,1500

**Β)Πυρήνας:** Στους παραπάνω τύπους $K = (1 + X_s \cdot X_t)^{\mathrm{ker}nel}$

Η μεταβλητή kernel αντιπροσωπεύει την τάξη του πυρήνα Η επιλογή πυρήνα πουωνυμικής τάξης διάφορη του 1 ήταν απαραίτητη μιας και τα δεδομένα μας ,όπως τα περισσότερα ρεαλιστικά, δεν είναι γραμμικά διαχωρίσιμα. Μόνο στην περίπτωση των πρώτων επαναλήψεων, όποτε οι διαστάσεις είναι τόσο πολλές, οι γραμμικοί πυρήνες μας αρκούν. . Οι τιμές που δοκιμάσαμε είναι οι :1,2,3,4,6,8,10,12,14,16,18, 20,40,45,45.

## *4.3 Αξιολόγηση της απόδοσης του αλγορίθμου KRR*

Για να μπορέσουμε να συγκρίνουμε την απόδοση του αλγορίθμου καθώς μεταβάλλουμε τις τιμές των παραμέτρων RP και kernel που περιγράψαμε παραπάνω είναι απαραίτητο να καθορίσουμε και μέτρα με βάσει τα οποία θα συγκρίνουμε την επίδοση του. Οι μεταβλητές που επιλέξαμε ως κριτήρια σύγκρισης είναι οι: Success Rate,Line Accuracy, Sensitivity , Specificity .Ας δούμε , όμως, πιο συγκεκριμένα τι εκφράζει κάθε μια από αυτές.

**Α. SUCCESS RATE :** είναι το ποσοστό επιτυχίας του αλγορίθμου όταν ο έλεγχος επίδοσης γίνεται με βάση νέο test set .

**Β. LINE ACCURACY:** είναι το ποσοστό επιτυχίας του αλγορίθμου όταν ο έλεγχος επίδοσης γίνεται με βάση το ίδιο set με το οποίο έγινε η εκπαίδευση

**Γ.SENSITIVITY:**(SENSITIVITY=TruePositives/(TruePositives+FalseNegatives )) ορίζεται ως η πιθανότητα σωστής θετικής πρόβλεψης ανάμεσα σε ασθενείς που είναι θετικοί . Για παράδειγμα αν ο classifier κάνει πρόβλεψη θετική σε 85από τους 100 (πάσχοντες-θετικούς ασθενείς)και στους υπόλοιπους 15

αρνητική(δηλαδή λάθος), τότε:TP=85,FN=15 και SE=Sensitivity=85/(85+15)=0.85

**Δ. SPECIFICITY:** Το SPECIFICITY ορίζεται ως η πιθανότητα σωστής αρνητικής πρόβλεψης ανάμεσα σε ασθενείς που είναι αρνητικοί.

# Κεφάλαιο5: Παρουσίαση αποτελεσμάτων

## Πειραματικός προσδιορισμός βέλτιστων RP,kernel

Βασικός σκοπός αυτής της εργασίας είναι η υλοποίηση και βελτιστοποίηση του αλγορίθμου Kernel Ridge Regression που έχει περιγραφεί αναλυτικά σε όλα τα προηγούμενα κεφάλαια. Οι μεταβλητές παράμετροι από τις οποίες εξαρτάται η απόδοση της μεθόδου που χρησιμοποιούμε είναι οι RP και kernel. Υπενθυμίζουμε ότι RP είναι η σταθερά κανονικοποιήσης .Μάλιστα όσο πιο μικρή είναι αυτή η σταθερά τόσο πιο πολλή η μέθοδος μας πλησιάζει αυτή της συμβατικής μεθόδου των ελαχίστων τετραγώνων. Η σταθερά kernel είναι ουσιαστικά η τάξη του πολυωνυμικού πυρήνα που χρησιμοποιήθηκε. Όσο μεγαλύτερη είναι η τάξη ,τόσο μεγαλώνουν και οι διαστάσεις του επιπέδου που διαχωρίζει τα δεδομένα μας. Έτσι για kernel=1 , προκύπτει ο ταξινομητής έχει τη μορφή ευθείας, για kernel=2 υπερβολής .

Η ανάγκη προσδιορισμού της σταθεράς RP από εμάς είναι ίσως το μόνο μειονέκτημα της μεθόδου Ridge Regression όπως αναφέραμε και στην παράγραφο 2.4 Κανονικοποιημένη γραμμική παλινδρόμηση.Έχουν μάλιστα αναπτυχθεί διάφορες μέθοδοι για τον προσδιορισμό της. Κατανοώντας την μεγάλη σημασία της σταθεράς RP , επιλέξαμε να δοκιμάσουμε τις επιδόσεις του αλγορίθμου για ένα ευρύ φάσμα τιμών .Δοκιμάσαμε πολύ μικρές τιμές 0,05 και 0, 5 , αλλά και τιμές από 100-1500 προκειμένου να προσδιορίσουμε πότε καταγράφονται οι καλύτερες τιμές για τις παραμέτρους αξιολόγησης απόδοσης(Success Rate, Line Accuracy, Sensitivity , Specificity).Αρχικά η δοκιμή για όλο αυτό το φάσμα των τιμών έγινε για γραμμικό πυρήνα (δηλαδή για σταθερή τιμή kernel ίση με 1).Δοκιμάσαμε όμως στη συνέχεια τις επιδόσεις του αλγορίθμου για όλο αυτό το εύρος τιμών του RP και για τους πυρήνες που παρατηρήσαμε βέλτιστη απόδοση του αλγορίθμου.

Ακόμα κρατώντας σταθερό το βέλτιστο RP προσπαθήσαμε να προσδιορίσουμε τον πυρήνα για τον οποίο παρατηρήθηκαν οι καλύτερες

επιδόσεις του αλγορίθμου. Δοκιμάσαμε και εδώ ένα μεγάλο φάσμα τιμών( πυρήνες τάξης:1/2/3/4/6/8/10/12/14/16/18/20/40/45/50).

Μεγάλη σημασία έχει να σημειώσουμε εδώ πως η επίδοση του αλγορίθμου κρίθηκε από τις τιμές που καταγράφαμε για τις παραμέτρους αξιολόγησης απόδοσης (Success Rate, Line Accuracy, Sensitivity , Specificity) για τις τιμές RP και kernel που προαναφέραμε όχι για ένα μόνο συγκεκριμένο αριθμό γονιδίων. Για κάθε συγκεκριμένη δοκιμαζόμενη τιμή RP,kernel καταγράφαμε ένα σύνολο τιμών για τις παραμέτρους αξιολόγησης απόδοσης. Κάθε τιμή αντιστοιχούσε σε ένα συγκεκριμένο αριθμό γονιδίων, μιας και αλγόριθμος μας σε κάθε επανάληψη αφαιρεί σταδιακά γονίδια και τερματίζει όταν μηδενιστούν. Για κάθε συγκεκριμένο ζεύγος τιμών RP,kernel η επίδοση του αλγορίθμου μπορεί να απεικονιστεί με 4 γραφικές παραστάσεις κάθε μία από τις οποίες απεικονίζει τη διακύμανση των τιμών μίας από της παραμέτρους αξιολόγησης απόδοσης (Success Rate, Line Accuracy, Sensitivity , Specificity) συναρτήσει του αριθμού γονιδίων που μειώνουμε σταδιακά. Έτσι συγκρίνοντας τις γραφικές παραστάσεις για διαφορετικές τιμές RP και σταθερό kernel μπορούμε να συμπεράνουμε ποια τιμή του RP βελτιστοποιεί τον αλγόριθμο. Αντίστοιχα συγκρίνοντας τις γραφικές παραστάσεις για διαφορετικές τιμές kernel και σταθερό RP μπορούμε να συμπεράνουμε ποια τιμή του kernel βελτιστοποιεί τον αλγόριθμο. Εκτός από την παρατήρηση των γραφικών για όλα τα γονίδια , ιδιαίτερο βάρος δώσαμε στην σύγκριση των διακυμάνσεων των τιμών των παραμέτρων αξιολόγησης για τα διάφορα ζεύγη τιμών (RP, kernel) για τα τελευταία 200 περίπου γονίδια. Αυτό γιατί , όπως έχουμε ήδη αναλύσει μας ενδιαφέρει ιδιαίτερα ο προσδιορισμός μια συγκεκριμένης μικρής ομάδας γονιδίων (informative genes) που δίνουν περισσότερες πληροφορίες για την κλάση των δειγμάτων και για αυτόν τον μικρό αριθμό είναι που θέλουμε να υπάρχει καλύτερη απόδοση. Επιπλέον, όπως αναφέραμε στην παράγραφο 2.1 Περιγραφή προβλήματος,για λιγότερα γονίδια είναι που υπάρχει πρόβλημα ταξινόμησης γραμμικούς ταξινομητές και επομένως εκεί απαιτούμε βελτιστοποίηση μέσω αύξησης της τάξης του πυρήνα.

Ακόμα σε πολλά σημεία , για διευκόλυνση της παρουσίασης των αποτελεσμάτων καταγράψαμε κάθε σύνολο τιμών καθεμιάς από τις

παραμέτρους αξιολόγησης σε έναν συγκριτικό πίνακα τιμών. Κάθε γραμμή του πίνακα μας δείχνει τη διακύμανση των τιμών για μία συγκεκριμένη τιμή RP.Κάθε στήλη αντιστοιχεί σε ένα συγκεκριμένο επίπεδο τιμών της παραμέτρου αξιολόγησης και σε κάθε κελί του πίνακα καταγράφεται ο αριθμός των γονιδίων για τα οποία αντιστοιχεί η συγκεκριμένη τιμή RP και παρατηρήθηκε η συγκεκριμένη τιμή της παραμέτρου αξιολόγησης. Μέσα από την προσεκτική παρατήρηση του πίνακα μπορεί κανείς να βγάλει συμπεράσματα για την τιμή του RP που βελτιστοποιεί τον αλγόριθμο.

Προκειμένου να καθορίσουμε τις βέλτιστες τιμές RP,kernel ακολουθούμε συνοπτικά την παρακάτω διαδικασία:

1.Καταγράφουμε και συγκρίνουμε τις τιμές των παραμέτρων Success Rate, Line Accuracy, Sensitivity , Specificity ( που περιγράψαμε στην παράγραφο 4.3 Αξιολόγηση της απόδοσης του αλγορίθμου KRR)00 για RP:0,05/0,5/100/200/300/400/500/ 600/700/800900/1000/1500 και γραμμικό πυρήνα (πυρήνας τάξης 1). Καταλήγουμε έτσι στις βέλτιστες τιμές RP.(παράγραφος 5.1.1/5.2.1)
2.Για τις παραπάνω βέλτιστες τιμές RP καταγράφουμε και συγκρίνουμε τις τιμές των παραμέτρων Success Rate, Line Accuracy, Sensitivity , Specificity για πυρήνες τάξης:1/2/3/4/6/8/10/12/14/16/18/20/40/45/50. (παράγραφος 5.1.2,5.1.3/5.2.2,5.2.3,5.2.4 )
3.Για τους βέλτιστους πυρήνες καταγράφουμε και συγκρίνουμε τις τιμές των παραμέτρων Success Rate, Line Accuracy, Sensitivity , Specificity για RP:0,05/0,5/100/200/300/400/500/600 /700/800/900/1000/1500 , προκειμένου να προσδιορίσουμε το βέλτιστο RP και για τους βέλτιστους πυρήνες (παράγραφος 5.1.4-5.1.5/5.2.5 )

Η παραπάνω διαδικασία πραγματοποιήθηκε και για τα δύο σύνολα δεδομένων που περιγράψαμε στην παράγραφο 04.1 Περιγραφή των συνόλων δεδομένων(datasets).Κάθε ένα από τα παραπάνω βήματα αποτελεί ουσιαστικά ένα υποκεφάλαιο.

Τα αποτελέσματα παρουσιάζονται υπό μορφή πινάκων και διαγραμμάτων.

Πίνακες χρησιμοποιήσαμε μόνο για την καταγραφή των επιδόσεων του αλγορίθμου για τις διάφορες τιμές RP για γραμμικό πυρήνα. Για κάθε παράμετρο αξιολόγησης αντιστοιχεί και ένας πίνακας. Κάθε γραμμή του πίνακα αντιστοιχεί σε μια συγκεκριμένη τιμή RP , ενώ κάθε στήλη σε ένα από τα επίπεδα τιμών που καταγράφηκε για τη συγκεκριμένη παράμετρο αξιολόγησης. Σε κάθε κελί του πίνακα καταγράφηκε ο αριθμός γονιδίων για τον οποίο παρατηρήσαμε το επίπεδο τιμής της παραμέτρου αξιολόγησης που βρίσκεται τη στήλη που αντιστοιχεί το κελί όταν η τιμή του RP ήταν αυτή που φαίνεται στην αντιστοιχεί για το κελί αυτό γραμμή. Για κάθε παράμετρο αξιολόγησης παρατηρούμε 1,2ή 3 πίνακες, ανάλογα με το μέγεθος των δεδομένων.

Διαγράμματα χρησιμοποιήσαμε για να παρουσιάσουμε τη διαδικασία προσδιορισμού των βέλτιστων τιμών RP αλλά και kernel, μιας και είναι πολύ πιο εύκολο για τον αναγνώστη να κρίνει τα αποτελέσματα μέσα από τις γραφικές. Για κάθε συγκεκριμένο ζεύγος τιμών RP,kernel η επίδοση του αλγορίθμου αντιστοιχεί σε 4 γραφικές παραστάσεις κάθε μία από τις οποίες απεικονίζει τη διακύμανση των τιμών μίας από της παραμέτρους αξιολόγησης απόδοσης (Success Rate, Line Accuracy, Sensitivity , Specificity) συναρτήσει του αριθμού γονιδίων που μειώνουμε σταδιακά. Έτσι συγκρίνοντας τις γραφικές παραστάσεις για διαφορετικές τιμές RP και σταθερό kernel μπορούμε να συμπεράνουμε ποια τιμή του RP βελτιστοποιεί τον αλγόριθμο. Αντίστοιχα συγκρίνοντας τις γραφικές παραστάσεις για διαφορετικές τιμές kernel και σταθερό RP μπορούμε να συμπεράνουμε ποια τιμή του kernel βελτιστοποιεί τον αλγόριθμο. Εκτός από την παρατήρηση των γραφικών για όλα τα γονίδια , ιδιαίτερο βάρος δώσαμε στην σύγκριση των διακυμάνσεων των τιμών των παραμέτρων αξιολόγησης για τα διάφορα ζεύγη τιμών (RP, kernel) για τα τελευταία 200 περίπου γονίδια με διαφορετικές γραφικές.

## *5.1 ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΟ ΠΡΩΤΟ DATASET*

### 5.1.1 Προσδιορισμός βέλτιστου RP για γραμμικό πυρήνα

Για να καθορίσουμε την βέλτιστη τιμή για την μεταβλητή RP δοκιμάσαμε να τρέξουμε τον αλγόριθμο για τις παρακάτω τιμές RP : 0.05 ,0.5 ,100 ,200 ,300

,400 ,500,600,700,800,900,1000,1500 και συγκρίναμε τις επιδόσεις του αλγορίθμου βάσει των τιμών των παραμέτρων SUCCESS RATE, LINE ACCURACY, SENSITIVITY, SPECIFICITY(που περιγράψαμε παραπάνω) .

**Α.Παρατηρήσεις για Success Rate:**

| | **Επίπεδα τιμών Success Rate** | | | |
|---|---|---|---|---|
| **RP** | **78,95** | **84,21** | **89,47** | **94,74** |
| **0,05** | 24188-3288 | 3188-2488 | 2388-688 | 588 |
| **0,5** | 24188-3288 | 3188-2588 | 2488-688 | 588 |
| **100** | 24188-5288 ,5088,4688-4588 | 5188,4988-4888, 4488-4088,3888 | 3988, 3788-788 | 688-288 |
| **200** | 24188-7288 | 7188-5688 | 5588-588 | 488-73 |
| **300** | 24188-8588 | 8388-7188 | 6988-1288 | 1188-68 |
| **400** | | 24188-8488 | 8788-1988 | 1888-73 |
| **500** | | 24188-15890,14290, 14090-13990,12090, 11790-10790,4488, 3988, 3588-3388 | 15690-14390,14190,13890-12190 ,11990-11890,10690 - 4588, 4288-4088, 3888-3788, 3288-2788 | 2688-68 |

*Πίνακας 3. 1. 1: Επίδραση μεταβολής RP(από 0,05 έως 500) στην τιμή του Success Rate για γραμμικό πυρήνα . Κάθε γραμμή του πίνακα καταγράφει τις τιμές του Success Rate κατά το τρέξιμο του αλγορίθμου καθώς σταδιακά μειώνουμε τα γονίδια από 24188 σε 0 για μια συγκεκριμένη τιμή του RP( την οποία σημειώνουμε στην πρώτη στήλη της κάθε γραμμής).Στην πρώτη γραμμή καταγράφουμε τις τιμές που λαμβάνει η μεταβλητή Success Rate για όλα τα τρεξίματα. Έτσι σε κάθε κελί το οποίο αντιστοιχεί σε μια συγκεκριμένη τιμή RP και μια συγκεκριμένη τιμή Success Rate καταγράψαμε το πλήθος των γονιδίων για το οποίο παρατηρήθηκε η συγκεκριμένη τιμή Success Rate.Οι τιμές RP για τις οποίες καταγράφει ο πίνακας τα αποτελέσματα είναι 0,05-500*

| RP | Επίπεδα τιμών Success Rate | | | |
|---|---|---|---|---|
| | **84,21** | **89,47** | **94,74** | **100** |
| **600** | 24188-20290,19490 -19390,19090,18790,5388 ,5088,4788-4588,3588-3488 | 20190-19590, 19290-19190, 18990-18890, 18690-5488 ,5288-5188, 4988-4888, 4488-3688, 3388-2788, 2688,2588-2388,168,118,88 | 2688,2288-178,68,53 | 48-0 |
| **700** | | 24188-12390,12190,11990, 11690-11590,10790-10690, 8288-7388,7088-6988, 5988 , 5688-2088,1888, 1788 ,148,108,78 | 12290,12090,11890-11790,11490-10890, 10590,10290-8388, 7288-7188,6888-6088,5888-5788,1988,1688-168,68 | 23-0 |
| **800** | | 12190-11390,9088-1688, 158-78 | 24188-12290, 11290-9188,1588-168,73,58,25 | 53-0 |

*Πίνακας 3. 1.2: Επίδραση μεταβολής RP(από 600 έως 800) στην τιμή του Success Rate για γραμμικό πυρήνα. Κάθε γραμμή του πίνακα καταγράφει τις τιμές του Success Rate κατά το τρέξιμο του αλγορίθμου καθώς σταδιακά μειώνουμε τα γονίδια από 24188 σε 0 για μια συγκεκριμένη τιμή του RP( την οποία σημειώνουμε στην πρώτη στήλη της κάθε γραμμής).Στην πρώτη γραμμή καταγράφουμε τις τιμές που λαμβάνει η μεταβλητή Success Rate για όλα τα τρεξίματα. Έτσι σε κάθε κελί το οποίο αντιστοιχεί σε μια συγκεκριμένη τιμή RP και μια συγκεκριμένη τιμή Success Rate καταγράψαμε το πλήθος των γονιδίων για το οποίο παρατηρήθηκε η συγκεκριμένη τιμή Success Rate.Οι τιμές RP για τις οποίες καταγράφει ο πίνακας τα αποτελέσματα είναι 500-800*

Στους πίνακες 3.1.1, 3.1.2 φαίνεται η διακύμανση των τιμών της παραμέτρου Success Rate για διάφορες τιμές της μεταβλητής RP. Οι τιμές αυτές αφορούν το σύνολο των επαναλήψεων

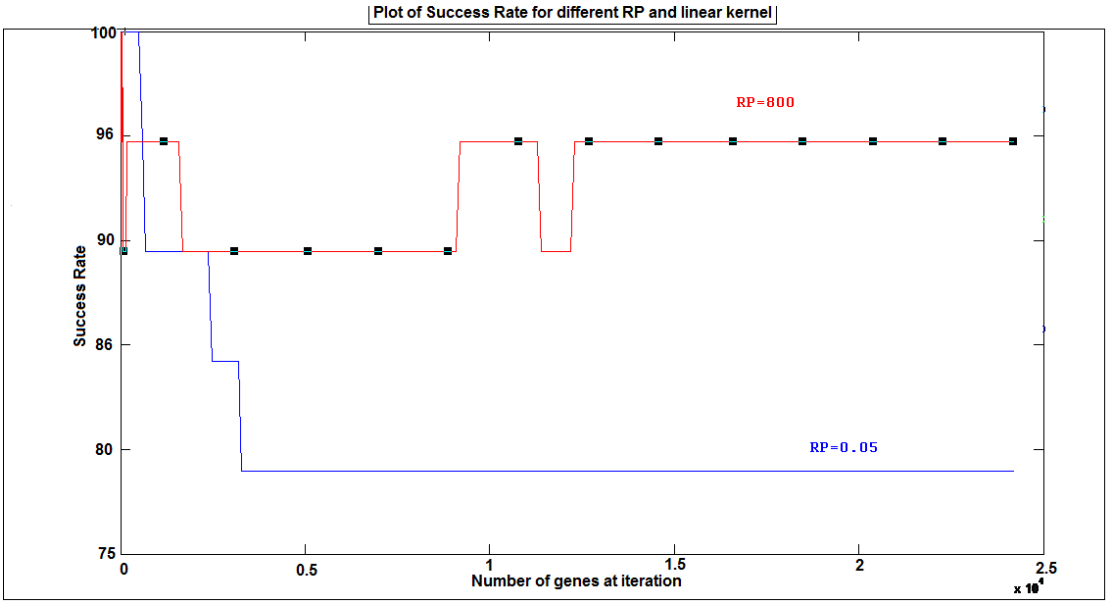
Πιο συγκεκριμένα παρατηρούμε τα ακόλουθα:

*-Για μικρές τιμές RP(0.5,0.05,100):* Η τιμή του Success Rate παραμένει σχεδόν αμετάβλητη. Το σύνολο τιμών είναι αρκετά ευσταθές .Δεν παρατηρούμε ,δηλαδή απότομες μεταβολές στις τιμές της παραμέτρου Success Rate.Η παράμετρος Success Rate παίρνει τη μέγιστη τιμή της –την τιμή 100- για αριθμό γονιδίων μικρότερο ή ίσο του 600.

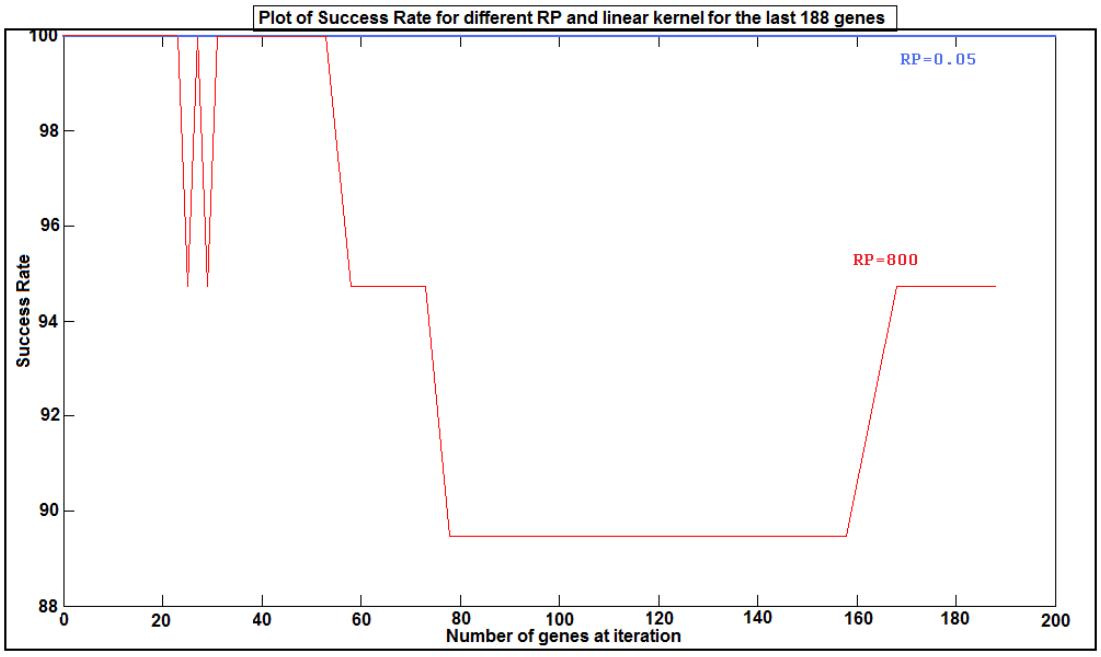
*-Για 200<=RP<=800:*Παρατηρούμε αύξηση στην τιμή της παραμέτρου Success Rate καθώς επίσης και ευστάθεια στις μεταβολές της τιμής της παραμέτρου .
*-Για 900<=RP<=1000:*Παρατηρούμε μείωση του συνόλου τιμών της παραμέτρου Success Rate κατά την αύξηση του RP.Επιπλέον των σύνολο τιμών είναι ασταθές. Παρατηρούμε δηλαδή απότομες μεταβολές στην τιμή της παραμέτρου Success Rate για κάθε συγκεκριμένη τιμή του RP.

Υπενθυμίζουμε ότι οι παραπάνω παρατηρήσεις βασίζονται στο σύνολο των επαναλήψεων. Παρατηρήσαμε , δηλαδή, την μεταβολή των τιμών του Success Rate για κάθε RP καθώς μειώναμε τα γονίδια από τα 24188-0.
**Συμπερασματικά το βέλτιστο RP βάσει της τιμής του Success Rate στο σύνολο των επαναλήψεων είναι το RP=800.Αν όμως μας ενδιαφέρουν μόνο οι επαναλήψεις για τις τελευταίες 3-4 εκατοντάδες γονιδίων τότε οι επιδόσεις είναι ελάχιστα καλύτερες για RP=0.05.**

**Παρακάτω βλέπουμε τις καλύτερες παρατηρήσεις για RP=800 και RP=0.05**




Διάγραμμα 2. 1*: απεικόνιση της διακύμανσης των τιμών του Success Rate καθώς αφαιρούμε σταδιακά γονίδια για τιμές RP=800,0.05 για το σύνολο των γονιδίων*




*Διάγραμμα 2. 2: Η διακύμανση των τιμών του Success Rate καθώς αφαιρούμε σταδιακά γονίδια για τιμές RP=800,0.05 για τις οποίες καταγράφηκαν οι καλύτερες τιμές Success Rate για τα τελευταία 188 γονίδια*

### Β.Παρατηρήσεις για Line Accuracy

**--Για το σύνολο των επαναλήψεων :24188-0 γονίδια**

| | Επίπεδα τιμών Line Accuracy | | | |
|---|---|---|---|---|
| **RP** | **96,15** | **97,44** | **98,72** | **100** |
| **0,05-200** | | | | 188-0 |
| **300** | | | 188-148, 118-41 | 133-128,39-0 |
| **400** | | 168,148,118,98-53,41 | 188-178,158, 138-128,108, 48-43,15 | 39-17 ,13-0 |
| **500** | 168-58 | 188-78,53-41 | 39-25,15 | 27-17, 13-0 |

*Πίνακας 3. 2. 1: Επίδραση μεταβολής RP(από 0,05 έως 500) στην τιμή του Line Accuracy για γραμμικό πυρήνα. Κάθε γραμμή του πίνακα καταγράφει τις τιμές του Line Accuracy κατά το τρέξιμο του αλγορίθμου καθώς σταδιακά μειώνουμε τα γονίδια από 24188 σε 0 για μια συγκεκριμένη τιμή του RP( την οποία σημειώνουμε στην πρώτη στήλη της κάθε γραμμής).Στην πρώτη γραμμή καταγράφουμε τις τιμές που λαμβάνει η μεταβλητή Line Accuracy για όλα τα τρεξίματα. Έτσι σε κάθε κελί το οποίο αντιστοιχεί σε μια συγκεκριμένη τιμή RP και μια συγκεκριμένη τιμή Line Accuracy καταγράψαμε το πλήθος των γονιδίων για το οποίο παρατηρήθηκε η συγκεκριμένη τιμή Line Accuracy. Οι τιμές RP για τις οποίες καταγράφει ο πίνακας τα αποτελέσματα είναι 500-1000*

| | Επίπεδα τιμών Line Accuracy | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **RP** | **88,46** | **92,31** | **93,59** | **94,87** | **96,15** | **97,44** | **98,72** | **100** |
| **600** | | | | | 188-39 | 37 | 35-31,17,13 | 29-19, 15, 11,0 |
| **700** | | | | | 188-37 | | 35-31,23-21, 11 | 29-25, 19-13, 10-0 |
| **800** | | | | 188-178 | 158-37 | 35,19 | 33-29,25-21, 11,6 | 27,17, 10-7,5-0 |
| **900** | | | | 188-128,63,37 | 118-68,58-39,35-33 | 31,27,21-19 | 29,25-23, 17-13,10,6 | 11,9-7, 5-0 |
| **1000** | | | 41 | 188-118,63,53-48,39-33 | 108-68,58,43 | 31-27,23-17 | 25,15-9,6 | 8-7,5-0 |
| **1500** | 158, 43-41 | 168,108,53 ,39-37,33-29,5,8,7 | 188-178,128-118,98-83,73-63,48,35,27,23 | 148-138, 58 | 21,11 | 19-15,10 | 13,9-6 | 5-0 |

*Πίνακας 3. 2. 2: Επίδραση μεταβολής RP(από 600 έως 1500) στην τιμή του Line Accuracy για γραμμικό πυρήνα. Κάθε γραμμή του πίνακα καταγράφει τις τιμές του Line Accuracy κατά το τρέξιμο του αλγορίθμου καθώς σταδιακά μειώνουμε τα γονίδια από 24188 σε 0 για μια συγκεκριμένη τιμή του RP( την οποία σημειώνουμε στην πρώτη στήλη της κάθε γραμμής).Στην πρώτη γραμμή καταγράφουμε τις τιμές που λάμβανε η μεταβλητή Line Accuracy για όλα τα τρεξίματα. Έτσι σε κάθε κελί το οποί αντιστοιχεί σε μια συγκεκριμένη τιμή RP και μια συγκεκριμένη τιμή Line Accuracy καταγράψαμε το πλήθος των γονιδίων για το οποίο παρατηρήθηκε η συγκεκριμένη τιμή Line Accuracy. Οι τιμές RP για τις οποίες καταγράφει ο πίνακας τα αποτελέσματα είναι 500-1500*

Παρατηρούμε με τη βοήθεια και των πινάκων 3.2.1 ,3.2.2 ότι η τιμή της παραμέτρου Line Accuracy μειώνεται λίγο καθώς αυξάνουμε την τιμή του RP.Για μικρά διαστήματα(λίγες επαναλήψεις) παρατηρούμε πιο μεγάλη πτώση της τιμής. **Για RP:0.05-200 η τιμή του Line Accuracy είναι η βέλτιστη (100) για όλες τις επαναλήψεις.**

**--Για τις τελευταίες εκατοντάδες γονιδίων :188-0 γονίδια**

| | Επίπεδα τιμών *Line Accuracy* | | | |
|---|---|---|---|---|
| **RP** | **96,15** | **97,44** | **98,72** | **100** |
| **0,05** | | | | 188-0 |
| **0,5** | | | | 188-0 |
| **100** | | | | 188-0 |
| **200** | | | | 188-0 |
| **300** | | | 188-148, 118-41 | 133-128, 39-0 |
| **400** | | 168,148,118,98-53,41 | 188-178, 158, 138-128,108, 48-43,15 | 39-17 ,13-0 |
| **500** | 168-58 | 188-78,53-41 | 39-25,15 | 27-17, 13-0 |

*Πίνακας 3. 3. 1 Επίδραση μεταβολής RP(από 0,05 έως 500) στην τιμή του Line Accuracy για γραμμικό πυρήνα για τα τελευταία 188 γονίδια . Κάθε γραμμή του πίνακα καταγράφει τις τιμές του Line Accuracy κατά το τρέξιμο του αλγορίθμου καθώς σταδιακά μειώνουμε τα γονίδια από 188 σε 0 για μια συγκεκριμένη τιμή του RP( την οποία σημειώνουμε στην πρώτη στήλη της κάθε γραμμής).Στην πρώτη γραμμή καταγράφουμε τις τιμές που λαμβάνει η μεταβλητή Line Accuracy για όλα τα τρεξίματα. Έτσι σε κάθε κελί το οποίο αντιστοιχεί σε μια συγκεκριμένη τιμή RP και μια συγκεκριμένη τιμή Line Accuracy καταγράψαμε το πλήθος των γονιδίων για το οποίο παρατηρήθηκε η συγκεκριμένη τιμή Line Accuracy. Οι τιμές RP για τις οποίες καταγράφει ο πίνακας τα αποτελέσματα είναι 0,05-50*

| | Επίπεδα τιμών Line Accuracy | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **RP** | **88,46** | **92,31** | **93,59** | **94,87** | **96,15** | **97,44** | **98,72** | **100** |
| **600** | | | | | 188-39 | 37 | 35-31,17 ,13 | 29-19, 15, 11,0 |
| **700** | | | | | 188-37 | | 35-31,23 -21, 11 | 29-25, 19-13, 10-0 |
| **800** | | | | 188-178 | 158-37 | 35,19 | 33-29,25 -21, 11,6 | 27,17, 10-7, 5-0 |
| **900** | | | | 188-128,63,37 | 118-68,58-39,35-33 | 31,27,21-19 | 29,25-23, 17-13,10 ,6 | 11,9-7, 5-0 |
| **1000** | | | 41 | 188-118, 63,53-48, 39-33 | 108-68,58,43 | 31-27,23-17 | 25,15-9,6 | 8-7,5-0 |
| **1500** | 158, 43-41 | 168,108,53 ,39-37,33-29,5,8,7 | 188-178,128-118,98-83,73-63,48,35,27,23 | 148-138, 58 | 21,11 | 19-15,10 | 13,9-6 | 5-0 |

*Πίνακας 3. 3. 2: Επίδραση μεταβολής RP(από 600 έως 1500) στην τιμή του Line Accuracy για γραμμικό πυρήνα για τα τελευταία 188 γονίδια .Κάθε γραμμή του πίνακα καταγράφει τις τιμές του Line Accuracy κατά το τρέξιμο του αλγορίθμου καθώς σταδιακά μειώνουμε τα γονίδια από 188 σε για μια συγκεκριμένη τιμή του RP( την οποία σημειώνουμε στην πρώτη στήλη της κάθε γραμμής).Στην πρώτη γραμμή καταγράφουμε τις τιμές που λαμβάνει η μεταβλητή Line Accuracy για όλα τα τρεξίματα. Έτσι σε κάθε κελί το οποίο αντιστοιχεί σε μια συγκεκριμένη τιμή RP και μια συγκεκριμένη τιμή Line Accuracy καταγράψαμε το πλήθος των γονιδίων για το οποίο παρατηρήθηκε η συγκεκριμένη τιμή Line Accuracy. Οι τιμές RP για τις οποίες καταγράφει ο πίνακας τα αποτελέσματα είναι 500-1500*

Παρατηρούμε ότι η τιμή της παραμέτρου Line Accuracy μειώνεται καθώς αυξάνουμε την τιμή του RP μετά την τιμή RP=200. **Ακόμα για RP:0.05-200, η τιμή του Line Accuracy είναι η βέλτιστη (100) για όλες τις επαναλήψεις. Συμπερασματικά ,βάσει της τιμής του Line Accuracy η βέλτιστη τιμή για το RP είναι RP:0.05-200**

### Γ. Παρατηρήσεις για Sensitivity

**--<u>Για το σύνολο των επαναλήψεων :24188-0 γονίδια</u>**

| | Επίπεδα τιμών Sensitivity | | |
|---|---|---|---|
| **RP** | **0,8333** | **0,9167** | **100,00** |
| **0,05-0,5** | | 24188-2588 | 2488-0 |
| **100** | | 24188-4088,3888 | 3988,3788-0 |
| **200** | | 24188-5688 | 5488 |
| **300** | | 24188-7188 | 7088-0 |
| **400** | | 24188-8888 | 8788-0 |
| **500** | | 24188-15890,14290,14090-13990, 12090, 11790-10790, 4488, 3988, 3588-3388,68 | 15790-14490, 4190, 13890-12390, 11990 11890,10690- 4588 , 4388-4088, 3788-3688,3288-0,188-73,63-0 |
| **600** | | 24188-20290,19490-19440, 19090, 18790,5388,5088, 4888-2788,2588-2388,168-158 ,118-88,53 | 20190-19590 ,19290 -19190,18990-18890, 18690-5488 ,5288 -5188,2688, 2288-178, 148-128,83-58,48-0 |
| **700** | | 8288-7388,7088-6988,5688 -2088, 1888-1788,148,108-78, 73,68,53,25 | 24188-8388,7288-7188, 6888-5788, 1988, 1688 .188-158,138-118,73, 63-58,48-27,23-0 |
| **800** | | 12188-11388,9088-1688, 158-78, 68-58,29,25 | 24188-12288,11288-9188, 1588 -168,73, 53-31,27, 23-0 |
| **900** | | 16788-15588,14988-14888, 14188, 13688-1588,388, 188, 168-41,33-27, 19 | 24188-16888,15488-15088, 14788-14288, 14088-13788, 1488-488, 288,178,39-35,25-21, 17-0 |
| **1000** | | 24188-1588,588-388, 188-53, 43, 31-17 | 1488-688,288, 48,39-33,15-0 |
| **1500** | 4488-788,188,108, 83,58-25 | 24188-4588,688-288,178-118, 98-88,78-63,23-17,13 | 15,11-0 |

*Πίνακας 3. 4. 1: Επίδραση μεταβολής RP(από 0,05 έως 1500) στην τιμή του Sensitivity για γραμμικό πυρήνα. Κάθε γραμμή του πίνακα καταγράφει τις τιμές του Sensitivity κατά το τρέξιμο του αλγορίθμου καθώς σταδιακά μειώνουμε τα γονίδια από 24188 σε 0 για μια συγκεκριμένη τιμή του RP( την οποία σημειώνουμε στην πρώτη στήλη της κάθε γραμμής).Στην πρώτη γραμμή καταγράφουμε τις τιμές που λαμβάνει η μεταβλητή Sensitivity για όλα τα τρεξίματα. Έτσι σε κάθε κελί το οποίο αντιστοιχεί σε μια συγκεκριμένη τιμή RP και μια συγκεκριμένη τιμή Sensitivity καταγράψαμε το πλήθος των γονιδίων για το οποίο παρατηρήθηκε η συγκεκριμένη τιμή Sensitivity*

Εξετάζοντας τον Πίνακα 3.4.1 που καταγράφει την διακύμανση των τιμών της παραμέτρου Sensitivity καθώς αλλάζουμε την τιμή της μεταβλητής RP για το σύνολο των επαναλήψεων, παρατηρούμε τα ακόλουθα:

*-Για πολύ μικρές τιμές RP(0.5,0.05):* Συγκρίνοντας τα σύνολα τιμών της παραμέτρου Sensitivity για τις δύο αυτές τιμές του RP δεν παρατηρούμε μεγάλες διαφορές.

*-Για 100<=RP<=800:*Καθώς αυξάνουμε την τιμή του RP από 100 έως 400, οι τιμές της παραμέτρου Sensitivity αυξάνονται στο σύνολο των επαναλήψεων. Για RP>400 οι τιμές της παραμέτρου Sensitivity παρουσιάζουν για αρκετές επαναλήψεις απότομες διακυμάνσεις, σε αντίθεση με το σύνολο τιμών της παραμέτρου Sensitivity για RP<400.Για το λόγο αυτό θεωρούμε ότι ο αλγόριθμος βελτιστοποιείται ως προς την τιμή της παραμέτρου Sensitivity για RP=400.

*-Για 900<=RP<=1500:*Οι τιμές της παραμέτρου Sensitivity μειώνονται, για το σύνολο των επαναλήψεων, καθώς αυξάνουμε την τιμή της μεταβλητής RP από 900 έως 1500.

Ακόμα οι τιμές της παραμέτρου Sensitivity παρουσιάζουν για αρκετές επαναλήψεις απότομες διακυμάνσεις,

Τελικά **την βέλτιστη τιμή του RP για το σύνολο των επαναλήψεων την παρατηρούμε για RP=400**

**--<u>Για τις τελευταίες εκατοντάδες γονιδίων :188-0 γονίδια</u>**

| | Επίπεδα τιμών Sensitivity | | |
|---|---|---|---|
| **RP** | **0,8333** | **0,9167** | **100,00** |
| **0,05-400** | | | 188-0 |
| **500** | | 68 | 188-73,63-0 |
| **600** | | 168-158 ,118-88,53 | 188-178,148-128,83-58,48-0 |
| **700** | | 148,108-78,73,68,53,25 | 188-158,138-118,73, 63-58, 48-27,23-0 |
| **800** | | 158-78,68-58,29,25 | 188-168,73, 53-31,27,23-0 |
| **900** | | 188,168-41,33-27,19 | 178,39-35,25-21,17-0 |
| **1000** | | 188-53,43,31-17 | 48,39-33,15-0 |
| **1500** | 4488-788,188 ,108, 83,58-25 | 178-118,98-88,78-63,23-17,13 | 15,11-0 |

*Πίνακας 3. 5. 1: Επίδραση μεταβολής RP(από 0,05 έως 1500) στην τιμή του Sensitivity για γραμμικό πυρήνα για τα τελευταία 188 γονίδια .Κάθε γραμμή του πίνακα καταγράφει τις τιμές του Sensitivity κατά το τρέξιμο του αλγορίθμου καθώς σταδιακά μειώνουμε τα γονίδια από 188 σε 0 για μια συγκεκριμένη τιμή του RP( την οποία σημειώνουμε στην πρώτη στήλη της κάθε γραμμής).Στην πρώτη γραμμή καταγράφουμε τις τιμές που λαμβάνει η μεταβλητή Sensitivity για όλα τα τρεξίματα. Έτσι σε κάθε κελί το οποίο αντιστοιχεί σε μια συγκεκριμένη τιμή RP και μια συγκεκριμένη τιμή Sensitivity καταγράψαμε το πλήθος των γονιδίων για το οποίο παρατηρήθηκε η συγκεκριμένη τιμή Sensitivity*

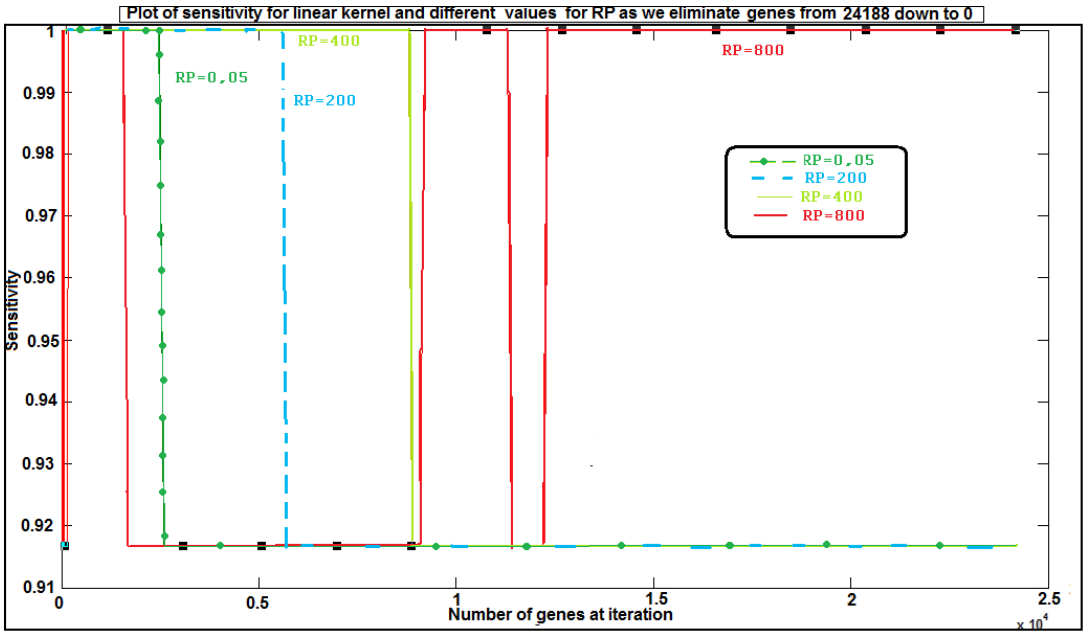
Εξετάζοντας τον παραπάνω πίνακα **3.5.1** παρατηρούμε τα ακόλουθα:

*<u>-Για 0.5<=RP<=400:</u>*Η τιμή της παραμέτρου Sensitivity είναι η βέλτιστη(ίση με 100) για όλες τις επαναλήψεις καθώς μειώνουμε τα γονίδια από 188 μέχρι και 0.

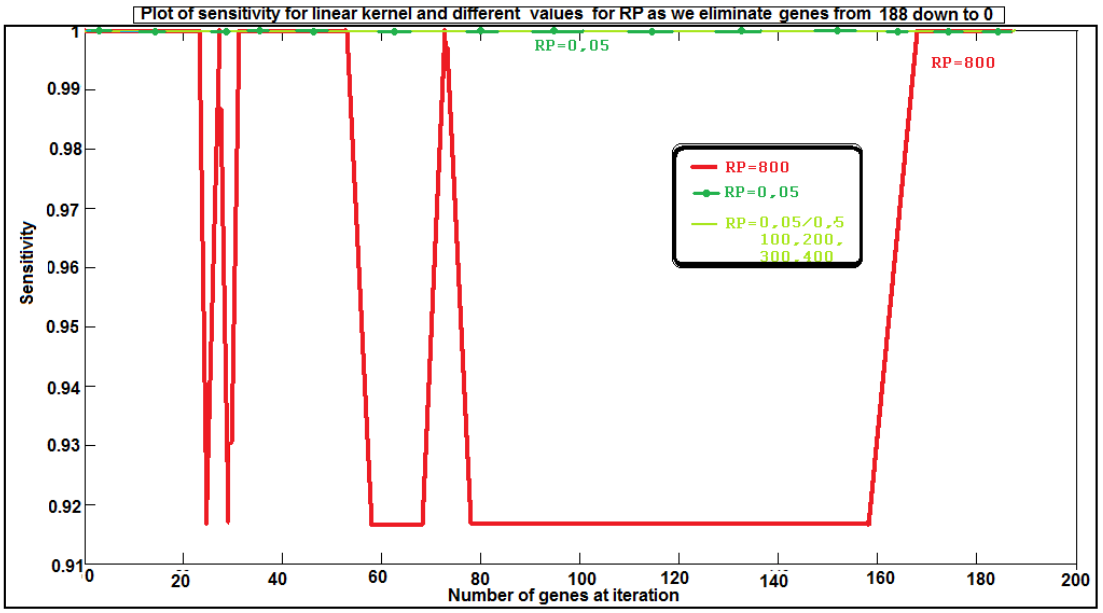
*<u>- Για 400<=RP<=1500:</u>* Η τιμή της παραμέτρου Sensitivity είναι η βέλτιστη(ίση με 100) για όλες τις επαναλήψεις καθώς μειώνουμε τα γονίδια από 188 μέχρι και 0. Εξαίρεση αποτελούν ελάχιστες επαναλήψεις κατά τις οποίες η τιμή της παραμέτρου sensitivity μειώνεται ελάχιστα.

Τελικά **την βέλτιστη τιμή του RP την παρατηρούμε για RP=400.**

Στις γραφικές : *Διάγραμμα 2.3*, *Διάγραμμα 2.4* απεικονίζεται το παραπάνω συμπέρασμα




*Διάγραμμα 2. 3: Διακύμανση των τιμών του Sensitivity συναρτήσει του αριθμού γονιδίων για τιμές RP=800,200,400,0.05 για τις οποίες καταγράφηκαν οι καλύτερες τιμές Sensitivity. Αφορά όλα τα γονίδια*




*Διάγραμμα 2. 4: Διακύμανση των τιμών του Sensitivity καθώς αφαιρούμε σταδιακά από 188 σε 0 για τιμές RP=800,200,400,0.05 για τις οποίες καταγράφηκαν οι καλύτερες τιμές Sensitivity*

**Δ. Παρατηρήσεις για Specificity:**

**--<u>Για το σύνολο των επαναλήψεων :24188-0 γονίδια</u>**

| | Επίπεδα τιμών Specificity | | | |
|---|---|---|---|---|
| **RP** | **0,5714** | **0,7143** | **0,8571** | **100,00** |
| **0,05** | 24188-3288,2488 | 3188-2588,2388-688 | 588 | 488-0 |
| **0,5** | 24188-3288 | 3188-688 | 588 | 488-0 |
| **100** | 24188-5288, 5088 ,4688 -4588 | 5188,4988-4788,4488-788 | 688-288 | 188-0 |
| **200** | 24188-7288 | 7188-588 | 488-148,118-93, 83 -73 | 138-128,88,68-0 |
| **300** | 24188-8588 | 8488-1288 | 1188-68 | 63-0 |

*Πίνακας 3. 6. 1: Επίδραση μεταβολής RP(από 0,05 έως 300) στην τιμή του Specificity για γραμμικό πυρήνα. Κάθε γραμμή του πίνακα καταγράφει τις τιμές του Specificity κατά το τρέξιμο του αλγορίθμου καθώς σταδιακά μειώνουμε τα γονίδια από 24188 σε 0 για μια συγκεκριμένη τιμή του RP( την οποία σημειώνουμε στην πρώτη στήλη της κάθε γραμμής).Στην πρώτη γραμμή καταγράφουμε τις τιμές που λαμβάνει η μεταβλητή Specificity για όλα τα τρεξίματα. Έτσι σε κάθε κελί το οποίο αντιστοιχεί σε μια συγκεκριμένη τιμή RP και μια συγκεκριμένη τιμή Specificity καταγράψαμε το πλήθος των γονιδίων για το οποίο παρατηρήθηκε η συγκεκριμένη τιμή Specificity. Οι τιμές RP για τις οποίες καταγράφει ο πίνακας τα αποτελέσματα είναι 0,05-300.*

| | Επίπεδα τιμών Specificity | | |
|---|---|---|---|
| **RP** | **0,7143** | **0,8571** | **100,00** |
| **400** | 24188-1988 | 1888-73 | 68-0 |
| **500** | 24188-2788 | 2688-73 | 68-0 |
| **600** | 24188-4988,4788-4588,3588-3488 | 4888,4488-3688, 3388-188, 168-158, 118-88, 53 | 188-178,148-128 ,83-58, 48-0 |
| **700** | 24188,12388,12188,11988,11688-11588, 10788 -10688, 10488-10388, 5988 | 12288,12088,11888 -11788, 11488-10888, 10588 ,10288 -6088 , 5888-73 | 68-0 |
| **800-1500** | | 24188-73 | 68-0 |

*Πίνακας 3. 6. 2: Επίδραση μεταβολής RP(από 400 έως 1500) στην τιμή του Specificity για γραμμικό πυρήνα. Κάθε γραμμή του πίνακα καταγράφει τις τιμές του Specificity κατά το*

*τρέξιμο του αλγορίθμου καθώς σταδιακά μειώνουμε τα γονίδια από 24188 σε 0 για μια συγκεκριμένη τιμή του RP( την οποία σημειώνουμε στην πρώτη στήλη της κάθε γραμμής).Στην πρώτη γραμμή καταγράφουμε τις τιμές που λαμβάνει η μεταβλητή Specificity για όλα τα τρεξίματα. Έτσι σε κάθε κελί το οποίο αντιστοιχεί σε μια συγκεκριμένη τιμή RP και μια συγκεκριμένη τιμή Specificity καταγράψαμε το πλήθος των γονιδίων για το οποίο παρατηρήθηκε η συγκεκριμένη τιμή Specificity. Οι τιμές RP για τις οποίες καταγράφει ο πίνακας τα αποτελέσματα είναι 400-1500..*

Εξετάζοντας τους πίνακες 3.6.1,3.6.2 παρατηρούμε τα ακόλουθα

*-Για 0.05<=RP<=800:*Καθώς αυξάνουμε την τιμή του RP από 100 έως 800, οι τιμές της παραμέτρου Specificity αυξάνονται στο σύνολο των επαναλήψεων. Για τιμές RP=100,600,700 οι τιμές της παραμέτρου Specificity παρουσιάζουν για αρκετές επαναλήψεις απότομες διακυμάνσεις

*-Για 800<RP<=1500:* οι τιμές της παραμέτρου Specificity δεν αυξάνονται άλλο.

**Τελικά βέλτιστη περιοχή τιμών RP(για το σύνολο των επαναλήψεων) θεωρούμε τις τιμές 800-1500.**

**--Για τις τελευταίες εκατοντάδες γονιδίων :188-0 γονίδια**

| RP\ Specificity | 0,8571 | 100,00 |
|---|---|---|
| **0,05-0,5** | | 488-0 |
| **100** | | 188-0 |
| **200** | 188-148,118-93,83-73 | 138-128,88,68-0 |
| **300** | 188-68 | 63-0 |
| **400-500** | 188-73 | 68-0 |
| **600** | 168-158,118-88,53 | 188-178,148-128,83-58,48-0 |
| **700-800** | 188-73 | 68-0 |
| **900-1000** | 188-68 | 63-0 |
| **1500** | 188-88 | 83-0 |

*Πίνακας 3. 7. 1 Επίδραση μεταβολής RP(από 0,05 έως 1500) στην τιμή του Specificity για γραμμικό πυρήνα για τα τελευταία 188 γονίδια. Κάθε γραμμή του πίνακα καταγράφει τις τιμές του Specificity κατά το τρέξιμο του αλγορίθμου καθώς σταδιακά μειώνουμε τα γονίδια από 188 σε 0 για μια συγκεκριμένη τιμή του RP( την οποία σημειώνουμε στην πρώτη στήλη της*

*κάθε γραμμής).Στην πρώτη γραμμή καταγράφουμε τις τιμές που λαμβάνει η μεταβλητή Specificity για όλα τα τρεξίματα. Έτσι σε κάθε κελί το οποίο αντιστοιχεί σε μια συγκεκριμένη τιμή RP και μια συγκεκριμένη τιμή Specificity καταγράψαμε το πλήθος των γονιδίων για το οποίο παρατηρήθηκε η συγκεκριμένη τιμή Specificity*

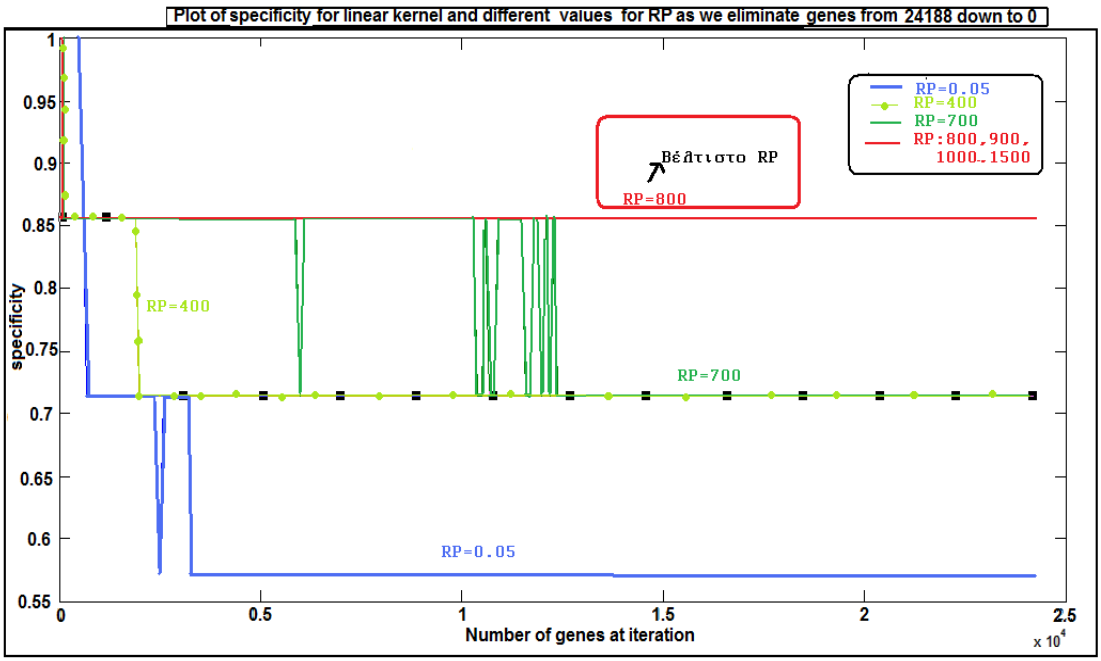
Εξετάζοντας τον παραπάνω πίνακα παρατηρούμε τα ακόλουθα

*-Για πολύ μικρές τιμές RP(0.5,0.05,100):*Η τιμή της παραμέτρου Specificity είναι η βέλτιστη(ίση με 100) για όλες τις επαναλήψεις καθώς μειώνουμε τα γονίδια από 188 μέχρι και 0

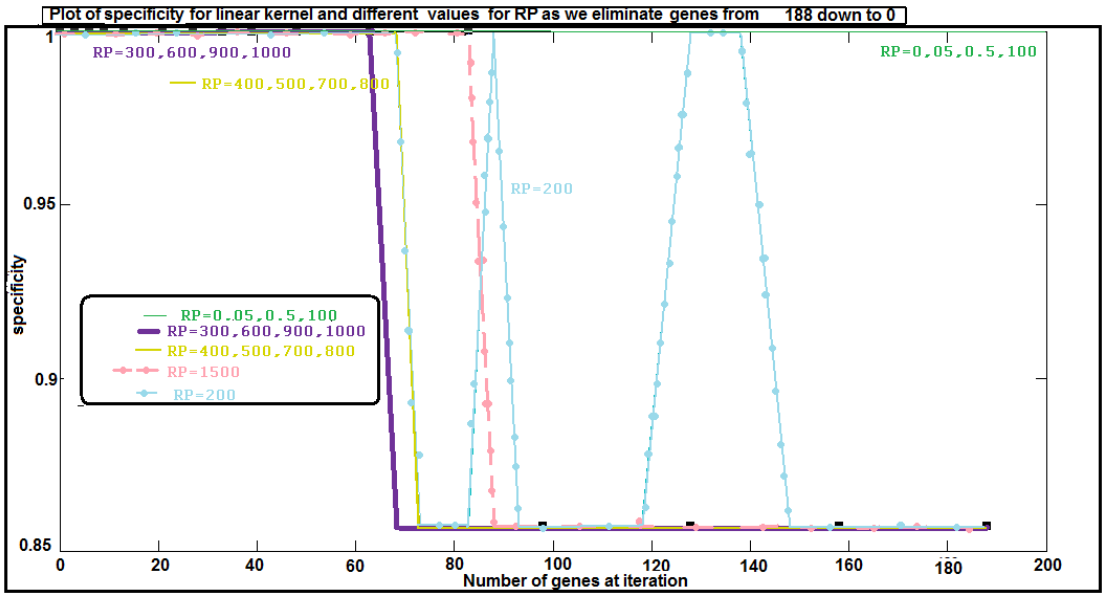
*- Για 200<=RP<=400:* Καθώς αυξάνουμε την τιμή του RP από 200 έως 400, οι τιμές της παραμέτρου Specificity μειώνεται .

*- Για 500<=RP<=1500:* Καθώς αυξάνουμε την τιμή του RP από 200 έως 400, οι τιμές της παραμέτρου Specificity μένουν σχεδόν σταθερές.

**Συμπερασματικά .Βάσει της τιμής του Specificity η βέλτιστη τιμή για το RP είναι RP:800(αν μας ενδιαφέρει το σύνολο των επαναλήψεων) και οι τιμές RP =0,05 /0,5/100 να μας ενδιαφέρουν οι τελευταίες εκατοντάδες γονιδίων. Αυτό απεικονίζεται και στις παρακάτω γραφικές παραστάσεις:**




*Διάγραμμα 2. 5: Διακύμανση των τιμών του Specificity καθώς αφαιρούμε σταδιακά από 24188 σε 0 για διάφορες τιμές RP .Φαίνεται καθαρά ότι για RP=800 καταγράφηκαν οι καλύτερες τιμές Specificity*




*Διάγραμμα 2. 6: Διακύμανση των τιμών του Specificity καθώς αφαιρούμε σταδιακά από 188 σε 0 για διάφορες τιμές RP .*

**Τελικά συμπεραίνουμε ότι:**

**1.Βάσει της βελτιστοποίησης της παραμέτρου Success Rate βέλτιστη τιμή για τη μεταβλητή RP είναι** RP:400,600,800(με βέλτιστο το 800) **(αν μας ενδιαφέρει το σύνολο των επαναλήψεων) και οι τιμή RP = 0,05 να μας ενδιαφέρουν οι τελευταίες εκατοντάδες γονιδίων.** ***(Διαγράμματα :2.1,2.2, Πίνακες :3.1-3.2)***

**2.Βάσει της βελτιστοποίησης της παραμέτρου Line Accuracy βέλτιστη τιμή για τη μεταβλητή RP είναι RP:0.05-200** ***(Διαγράμματα :2.3,2.4, Πίνακες :3.3-3.4)***

**3.Βάσει της βελτιστοποίησης της παραμέτρου Sensitivity βέλτιστη τιμή για τη μεταβλητή RP είναι RP:0.05-400** ***(Διαγράμματα :2.5,2.6, Πίνακες :3.5-3.6)***

**4.Βάσει της βελτιστοποίησης της παραμέτρου Specificity βέλτιστη τιμή για τη μεταβλητή RP είναι RP:800(αν μας ενδιαφέρει το σύνολο των επαναλήψεων) και οι τιμές RP = 0,05 /0,5/100 να μας ενδιαφέρουν οι τελευταίες εκατοντάδες γονιδίων.** ***(Διαγράμματα :2.7,2.8, Πίνακες :3.7-3.8)***

### 5.1.2 Προσδιορισμός βέλτιστου πυρήνα για δεδομένο RP=800

Υπενθυμίζουμε ότι θα δοκιμάσουμε διάφορους πολυωνυμικούς πυρήνες των οποίων τις επιδόσεις και θα συγκρίνουμε. Πιο συγκεκριμένα συγκρίνουμε τις επιδόσεις για γραμμικό πυρήνα, πολυωνυμικούς πυρήνες τάξεως 2,3,4,6,8,10, 12,14,16, 18,20, 40, 45,50.

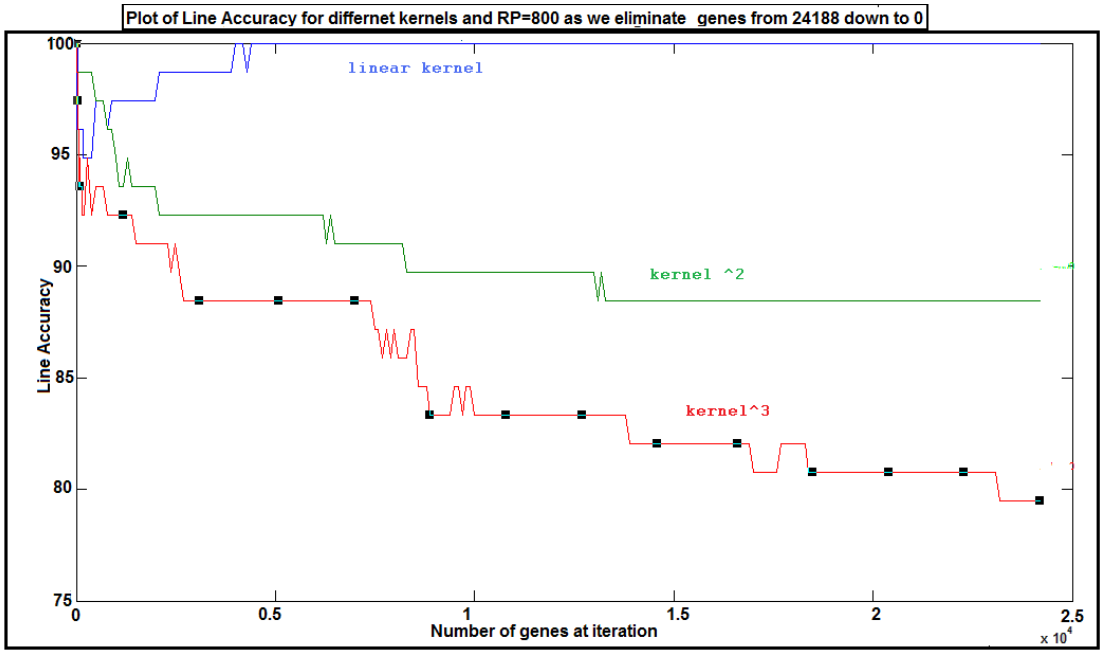
Επειδή διαπιστώσαμε ότι καλύτερες τιμές για το RP είναι RP=800 και RP=0,05 επιλέξαμε να τρέξουμε τον αλγόριθμο για τις δύο αυτές τιμές RP αλλάζοντας τον πυρήνα , προκειμένου να βρούμε τον βέλτιστο.

Θα παρατηρήσουμε και πάλι κατά πόσο ο κάθε πυρήνας βελτιστοποιεί τις Παραμέτρους :Success Rate, Line Accuracy, Sensitivity,Specificity για να καθορίσουμε τον πυρήνα που βελτιστοποιεί περισσότερο τον αλγόριθμο. Η εκτίμηση της βελτιστοποίησης των παραμέτρων αυτών θα γίνεται και για σύνολο των επαναλήψεων (24188 έως 0 γονίδια) αλλά και εστιάζοντας στις τελευταίες εκατοντάδες γονίδια.

**A.Παρατηρήσεις για επίδραση πυρήνων στην παράμετρο Line Accuracy**

***--Στο σύνολο των επαναλήψεων: (24188 έως 0 γονίδια)***

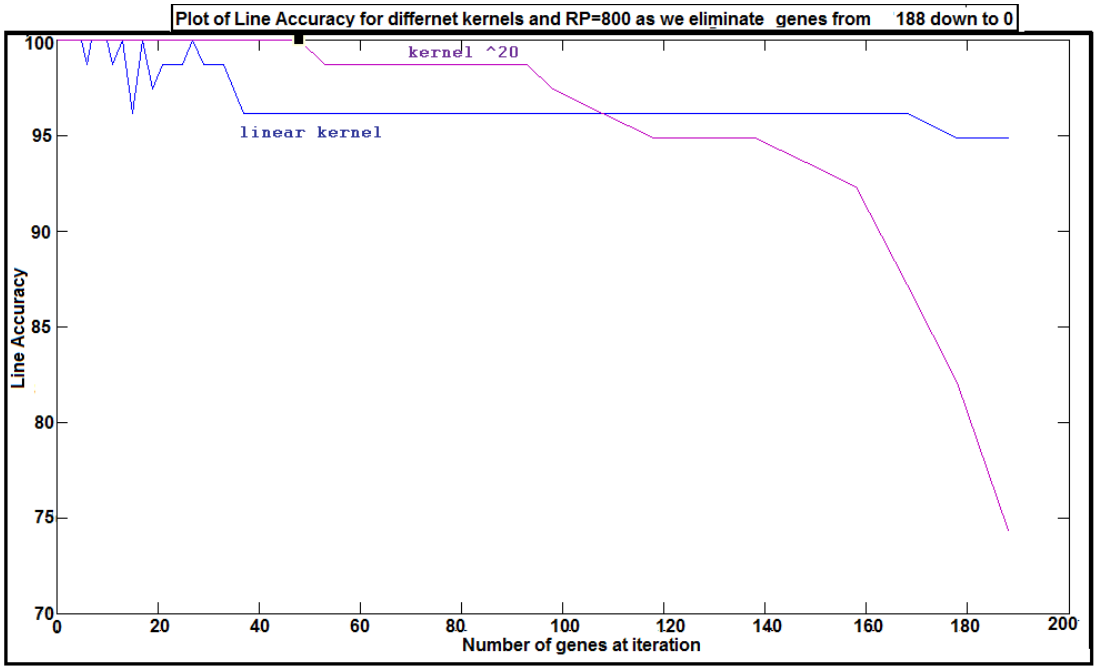
Στο παρακάτω διάγραμμα (Διάγραμμα 2.7) παρουσιάζεται ενδεικτικά η διακύμανση των τιμών της παραμέτρου Line Accuracy στο σύνολο των επαναλήψεων για πυρήνες τάξης 1(γραμμικός πυρήνας), 2 και 3.



*Διάγραμμα 2. 7: Διακύμανση των τιμών του Line Accuracy καθώς αφαιρούμε σταδιακά από 24188 σε 0 για RP=800 και διαφορετικούς πυρήνες .Κάθε μία από τις γραφικές αντιστοιχεί σε διαφορετική τάξη του πολυωνυμικού πυρήνα .Το διάγραμμα απεικονίζει τις γραφικές για πολυωνυμικούς πυρήνες τάξεων:1,2,3 για τους οποίους καταγράφηκαν οι καλύτερες τιμές Line Accuracy .*

Παρατηρούμε ότι για το σύνολο των επαναλήψεων η τιμή της παραμέτρου Line accuracy χειροτερεύει καθώς αυξάνουμε την τάξη του πολυωνυμικού πυρήνα. Αυτό ήταν και αναμενόμενο, μιας και για χώρους πολλών διαστάσεων(λόγω των πολλών γονιδίων) είναι εξηγήσιμο ο γραμμικός πυρήνας να έχει καλύτερες επιδόσεις. Αντίθετα βελτιώσεις περιμένουμε όταν τα γονίδια μειωθούν πολύ(τελευταία εκατοντάδα γονιδίων).Έτσι από εδώ και κάτω δε θα εξετάσουμε κατά πόσο αυξάνεται η τιμή των υπόλοιπων παραμέτρων για όλες τις επαναλήψεις, αλλά θα εστιάσουμε στις τελευταίες 2 εκατοντάδες γονιδίων.

***--Για 188 έως 0 γονίδια:***




*Διάγραμμα 2. 8: Διακύμανση των τιμών του Line Accuracy καθώς αφαιρούμε γονίδια σταδιακά από 188 σε 0 για RP=800 για διαφορετικούς πυρήνες .Κάθε μία από τις γραφικές αντιστοιχεί σε διαφορετική τάξη του πολυωνυμικού πυρήνα .Το διάγραμμα απεικονίζει τις γραφικές για πολυωνυμικούς πυρήνες τάξεων:1,20 για τους οποίους καταγράφηκαν οι καλύτερες τιμές Line Accuracy*

Στο παραπάνω διάγραμμα παρουσιάζεται ενδεικτικά η διακύμανση των τιμών της παραμέτρου Line Accuracy για τα τελευταία 188 γονίδια για πυρήνες τάξης 1(γραμμικός πυρήνας), 2 και 3.

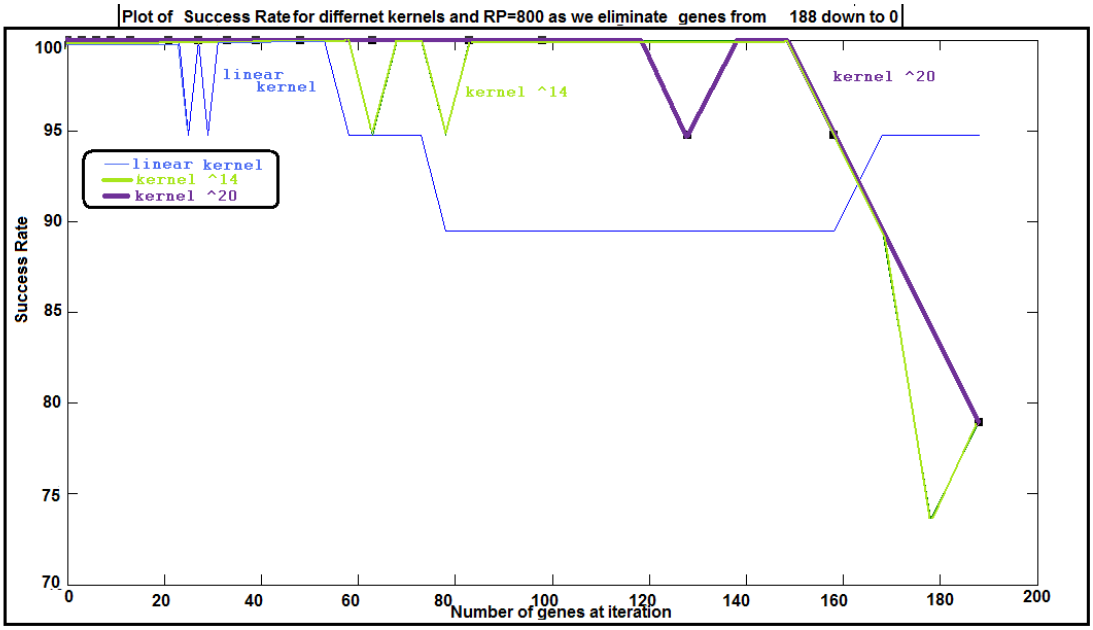
Παρατηρήσαμε ότι καθώς αυξάνουμε την τάξη του πολυωνυμικού πυρήνα αυξάνεται και η τιμή της παραμέτρου Line Accuracy.Πιο συγκεκριμένα από τους πυρήνες τάξης 1-20 παρατηρούμε βελτίωση των τιμών του Line Accuracy.Από εδώ και έπειτα παρατηρούμε χειροτέρευση του Line Accuracy. Δηλαδή L.A(kernel _1)< L.A(kernel _2)< L.A(kernel _4)< L.A(kernel _6)<…<L.A(kernel _20)> L.A(kernel _40)> L.A(kernel _45)> L.A(kernel _50)

***Συμπέρασμα: Ο πυρήνας που βελτιστοποιεί τον αλγόριθμο ως προς την παράμετρο Line Accuracy είναι ο πολυωνυμικός πυρήνας τάξης 20 .Αξίζει ακόμα να παρατηρήσουμε ότι ο πολυωνυμικός πυρήνας τάξης 20 παρουσιάζει ευστάθεια στις μεταβολές των τιμών του συγκριτικά με άλλους πυρήνες***

**Β. Παρατηρήσεις για επίδραση πυρήνων στην παράμετρο Success Rate**

***--<u>Για 188 έως 0 γονίδια:</u>***

Στο παρακάτω διάγραμμα (Διάγραμμα 2.9) παρουσιάζεται ενδεικτικά η διακύμανση των τιμών της παραμέτρου Success Rate για τα τελευταία 188 γονίδια για πυρήνες τάξης 1(γραμμικός πυρήνας), 14 και 20

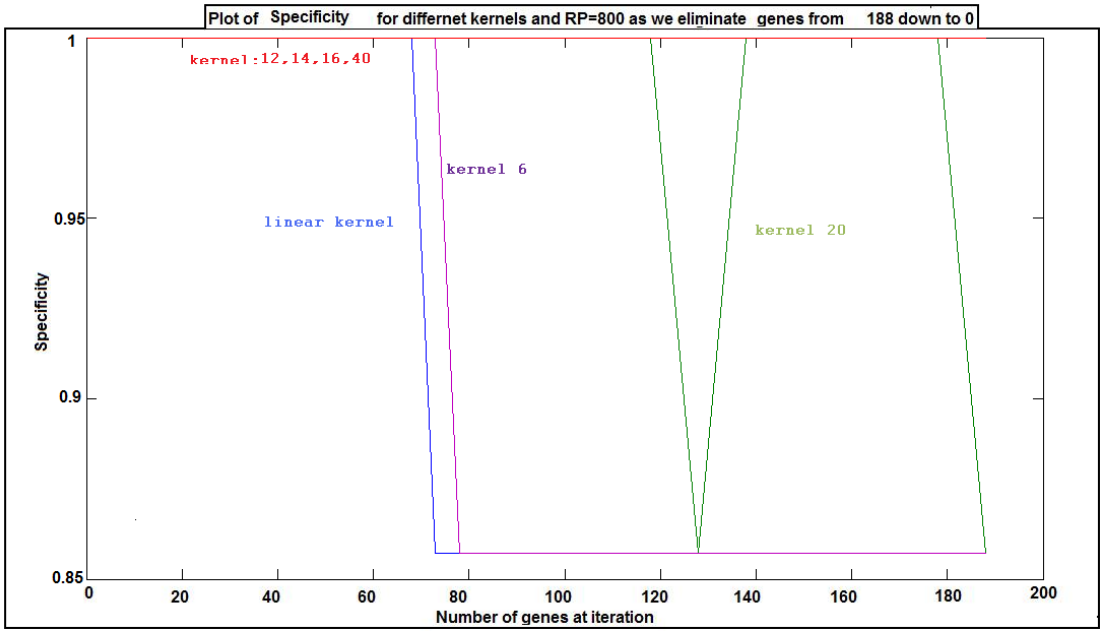


*Διάγραμμα 2. 9: Διακύμανση των τιμών του Success Rate καθώς αφαιρούμε σταδιακά από 188 σε 0 για πολυωνυμικούς πυρήνες τάξεων:1,14,20 . Για τους πυρήνες αυτούς καταγράφηκαν οι καλύτερες τιμές Success Rate*

Παρατηρήσαμε και εδώ ότι η παράμετρος Success Rate Βελτιώνεται καθώς αυξάνουμε την τάξη του πολυωνυμικού πυρήνα. Η βελτίωση παρατηρείται και εδώ μέχρι και για τάξη πυρήνα ίση με 20.Από εκεί και πάνω έχουμε χειροτέρευση. Πολλοί βέβαια από τους πυρήνες δίνουν μικρή μόνο βελτίωση και αρκετοί έχουν αστάθειες. **Τελικά συμπεράναμε ότι μεγαλύτερη βελτίωση παρατηρούμε για τους πυρήνες τάξης 14 και 20, με τον πυρήνα τάξης 20 να παρουσιάζει λιγότερες αστάθειες.**

### Γ. Παρατηρήσεις για επίδραση πυρήνων στην παράμετρο Specific

***--Για 188 έως 0 γονίδια:***




Διάγραμμα 2. 10:Διακύμανση των τιμών του Specificity καθώς αφαιρούμε σταδιακά γονίδια από 188 σε 0 για πολυωνυμικούς πυρήνες τάξεων:1,6,12,14,16,40 . Για τους πυρήνες αυτούς *καταγράφηκαν οι καλύτερες τιμές Specificity*

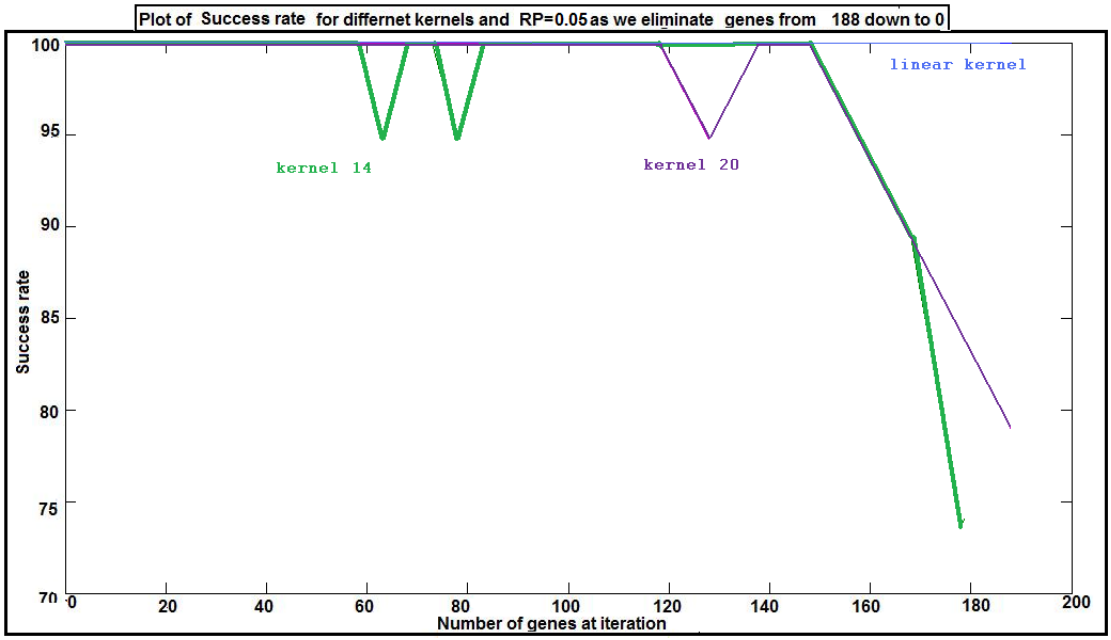
Στο παραπάνω διάγραμμα (Διάγραμμα 2.10) παρουσιάζεται ενδεικτικά η διακύμανση των τιμών της παραμέτρου Specificity για τα τελευταία 188 γονίδια για πυρήνες τάξης 1(γραμμικός πυρήνας), 12,14 ,16 και 40 μιας και παρατηρήσαμε ότι γι αυτούς ο αλγόριθμος παρουσιάζει τις καλύτερες τιμές Specificity.

Παρατηρήσαμε ότι οι πυρήνες που βελτιστοποιούν τον αλγόριθμο ως προς το Specificity είναι οι πυρήνες τάξης 12,14,16,40.Σημαντική βελτίωση παρατηρούμε για όλους τους υπόλοιπους πυρήνες, αν και παρατηρούμε επίσης πιο μεγάλη αστάθεια καθώς και αρκετές επαναλήψεις στις οποίες η τιμή πέφτει κάτω από τις επιδόσεις του γραμμικού πυρήνα

**ΤΕΛΙΚΗ ΑΞΙΟΛΟΓΗΣΗ: Συμπερασματικά ,καλύτερους πυρήνες θεωρούμε τους πυρήνες τάξης 14 και 20 ,λαμβάνοντας υπόψη τις επιδράσεις στο σύνολο των παραμέτρων του αλγορίθμου.**

### 5.1.3 Προσδιορισμός βέλτιστου πυρήνα για RP=0.05

#### A. Παρατηρήσεις για επίδραση της τάξης του πυρήνα στο Success Rate

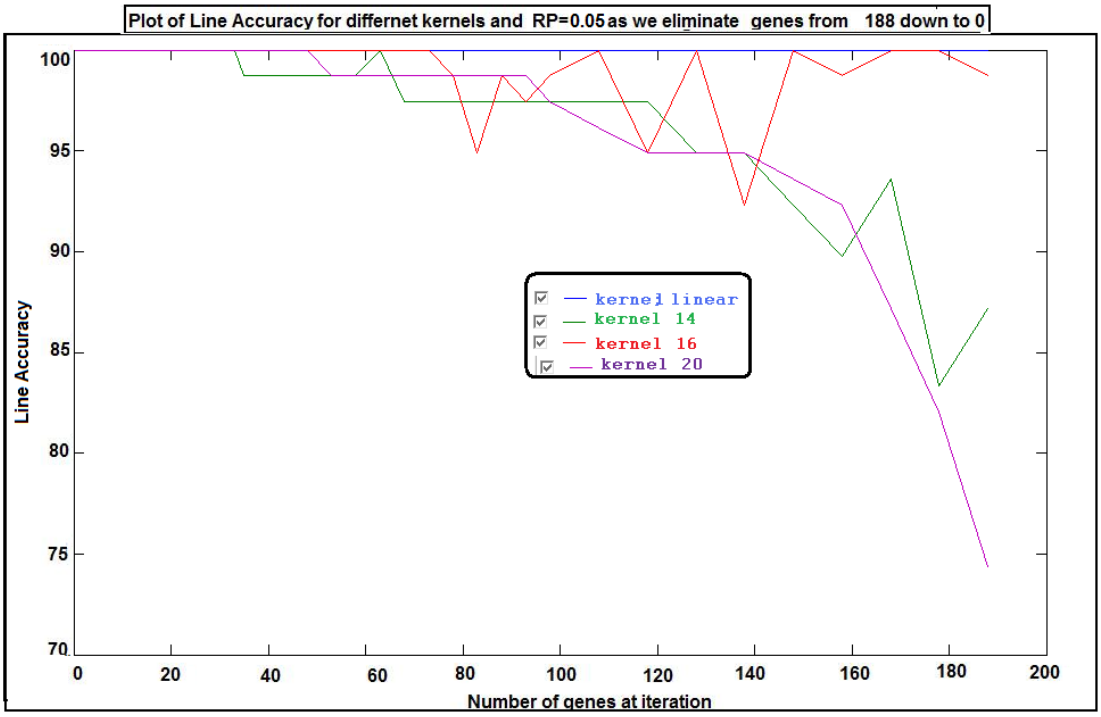



*Διάγραμμα 2. 11:* Διακύμανση *των τιμών του Success Rate καθώς αφαιρούμε σταδιακά γονίδια από 188 σε 0 για πολυωνυμικούς πυρήνες τάξεων:1,14,20.Για τους πυρήνες αυτούς καταγράφηκαν οι καλύτερες τιμές Success Rate*

Το Διάγραμμα 2.11 εικονίζει τη διακύμανση των τιμών της παραμέτρου Success Rate για τα τελευταία 188 γονίδια και για πυρήνες τάξης 1(γραμμικός πυρήνας),14,20.Τα παρακάτω συμπεράσματα δεν αφορούν τις παρατηρήσεις πάνω στο διάγραμμα αλλά συνολικά για τους πυρήνες που τρέξαμε τον αλγόριθμο. Απεικονίσαμε ενδεικτικά τους καλύτερους. Πιο συγκεκριμένα

παρατηρήσαμε ότι βέλτιστες είναι οι επιδόσεις του αλγορίθμου για γραμμικό πυρήνα. Ακόμα από τους υπόλοιπους πυρήνες καλύτεροι είναι οι kernel 20 και kernel 14.Υπάρχει βελτίωση των επιδόσεων του αλγορίθμου καθώς ανεβαίνει η τάξη του πυρήνα (χωρίς να συγκρίνουμε βέβαια με το γραμμικό πυρήνα).Παρατηρούμε όμως και πολλές αστάθειες στις τιμές σχεδόν σε όλους τους υπόλοιπους πυρήνες.

### Β. Παρατηρήσεις για επίδραση πυρήνων στην παράμετρο Line Accuracy



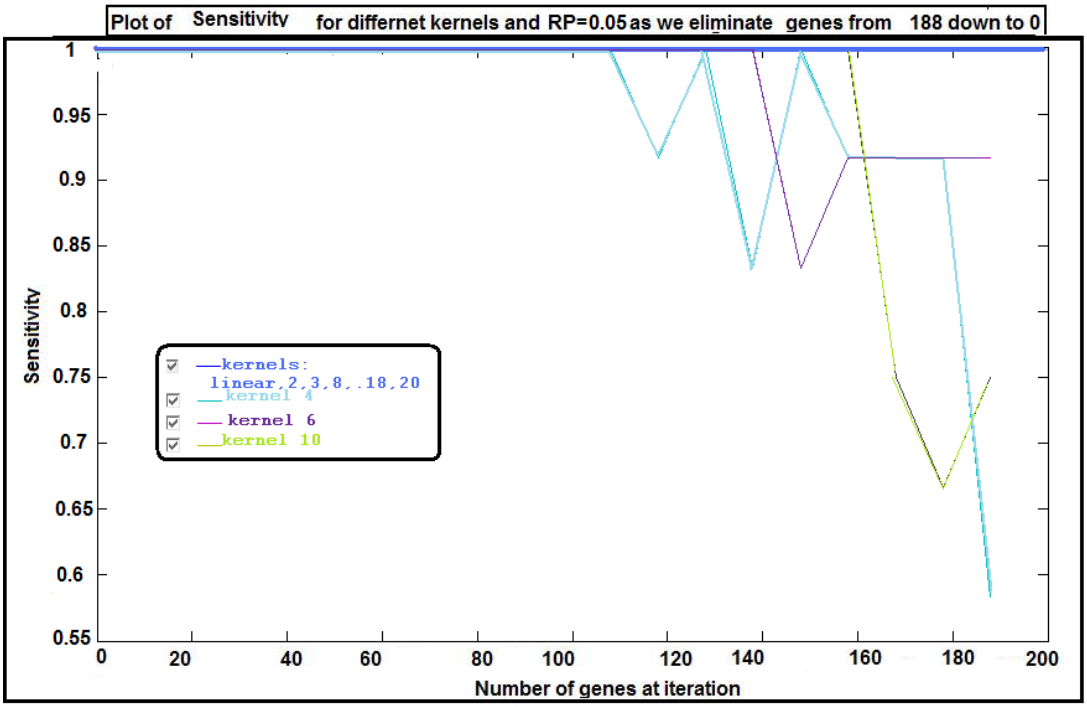

*Διάγραμμα 2. 12:* Διακύμανση *των τιμών του Line Accuracy καθώς αφαιρούμε σταδιακά γονίδια από 188 σε 0 για πολυωνυμικούς πυρήνες τάξεων:1,14,16,20.Για τους πυρήνες αυτούς καταγράφηκαν οι καλύτερες τιμές Line Accuracy*

Το Διάγραμμα 2.12 απεικονίζει τη διακύμανση των τιμών Line Accuracy για τους πυρήνες για τους οποίους παρατηρήσαμε ότι ο αλγόριθμος εμφανίζει καλύτερη τιμή.

Και εδώ βέλτιστος είναι ο γραμμικός πυρήνας. Γενικά παρατηρώ μείωση των επιδόσεων με την αύξηση της τάξης με εξαιρέσεις. Οι περισσότεροι πυρήνες παρουσιάζουν μεγάλες αστάθειες. Αν αφαιρέσουμε από τις συγκρίσεις μας

τον γραμμικό πυρήνα καλύτερες είναι και πάλι οι επιδόσεις για kernel 14, 20 και λόγω τιμής και λόγω ευστάθειας.

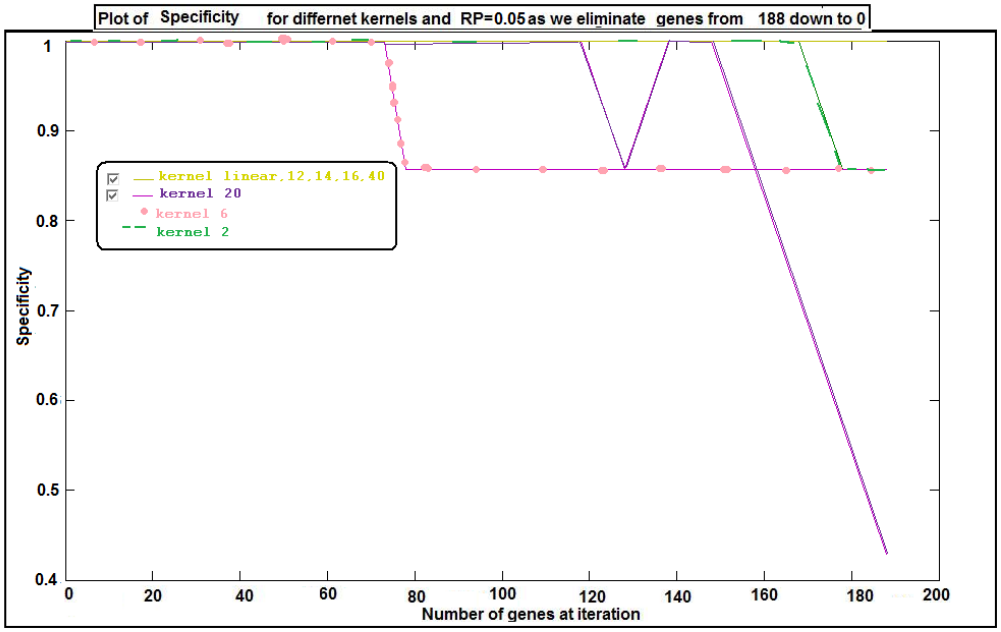
### Γ. Παρατηρήσεις για επίδραση πυρήνων στην παράμετρο Sensitivity




*Διάγραμμα 2. 13: Διακύμανση των τιμών του Sensitivity καθώς αφαιρούμε σταδιακά γονίδια από 188 σε 0 για πολυωνυμικούς πυρήνες τάξεων: 2,3 ,8,18,20. Για τους πυρήνες αυτούς καταγράφηκαν οι καλύτερες τιμές Sensitivity*

Οι πυρήνες με τις καλύτερες επιδόσεις είναι πολυωνυμικοί πυρήνες τάξης 1,2,3,8,18,20.Στους υπόλοιπους(αυτούς που δεν απεικονίζουμε) παρατηρούμε μεγάλες αστάθειες.

### Δ. Παρατηρήσεις για επίδραση πυρήνων στην παράμετρο Specificity:




*Διάγραμμα 2. 14: Διακύμανση των τιμών του Specificity καθώς αφαιρούμε σταδιακά γονίδια από 188 σε 0 για διάφορους πυρήνες .Για πολυωνυμικούς πυρήνες τάξεων: 2,6,12,14,16, 20, 40 καταγράφηκαν οι καλύτερες τιμές Specificity*

Όπως παρατηρούμε στο Διάγραμμα 2.14 βέλτιστες τιμές για το Specificity πετυχαίνουμε για πυρήνες :γραμμικό , τάξης 12, 14 , 16, 40 .Καλές είναι οι επιδόσεις και για τους υπόλοιπους απεικονιζόμενους πυρήνες , ενώ γι αυτούς που δεν απεικονίζουμε παρατηρήσαμε αστάθειες.

Τελικά συμπεραίνουμε ότι και για RP=0,05 οι πυρήνες τάξεις 14, 20 συνεπάγονται καλές επιδόσεις του αλγορίθμου. Παρατήρησα όμως συνολικά ότι βελτιστοποίηση όλων των παραμέτρων του έχω για γραμμικό πυρήνα.

Συμπέρασμα: Για πολύ μικρές τιμές RP όπως 0,05 παρατηρούμε ότι η αύξηση της τάξης του πυρήνα δεν συνεπάγεται βελτίωση των επιδόσεων. Μας αρκεί δηλαδή ένας γραμμικός πυρήνας.

Για μεγάλες τιμές RP (από 100 και πάνω) παρατηρούμε ότι η αύξηση της τάξης του πυρήνα συνεπάγεται βελτίωση των επιδόσεων του αλγορίθμου με βέλτιστη τιμή να είναι η τιμή 800.

Η παρατήρηση αυτή είναι απόλυτα εξηγήσιμη. Όπως παρατηρούμε στις σχέσεις που εκφράζουν τη διαδικασία της εκπαίδευσης και της ταξινόμησης του αλγορίθμου μας

*Εκπαίδευση-εύρεση w.* $W=\frac{1}{2l}\sum_{t=1}^{T} a_t x_t = (K+lI)^{-1} yx$

*Πρόβλεψη κλάσεων για νέα δείγματα:* $y_{new} = wx_{new} = (K+lI)^{-1} yx \cdot x_{new}$

$= (K+lI)^{-1} yK = y'(K+lI)^{-1} K$

Για πολύ μικρές τιμές του l(=RP) το w δεν εξαρτάται από την τιμή του πυρήνα, ενώ για μεγάλες τιμές της σταθερά κανονικοποίησης l(=RP) για τον καθορισμό του w σημαντικότερο ρόλο παίζει ο πυρήνας και όχι το L.Αυτό συμβαίνει λόγω της αντιστροφής.

Βάσει των παραπάνω παρατηρήσεων καταλήγουμε ότι η χρήση γραμμικού πυρήνα προτιμάται για πολύ μικρές τιμές RP.Ακόμα ο γραμμικός πυρήνας είναι προτιμότερος για μεγάλο αριθμό γονιδίων (πολλές χιλιάδες) , καθώς ο χώρος μας στην περίπτωση αυτή είναι πολλών διαστάσεων και ο διαχωρισμός πολύ πιο δύσκολος με πολυωνυμικούς πυρήνες. Αυτός είναι και ο λόγος που οι όποιες βελτιώσεις με χρήση πολυωνυμικών πυρήνων παρατηρούνται μόνο για τις τελευταίες εκατοντάδες γονιδίων .Αντίθετα , πολυωνυμικούς πυρήνες χρησιμοποιούμε για πιο μεγάλες τιμές RP και για τις λιγότερα γονίδια.(188 και κάτω)

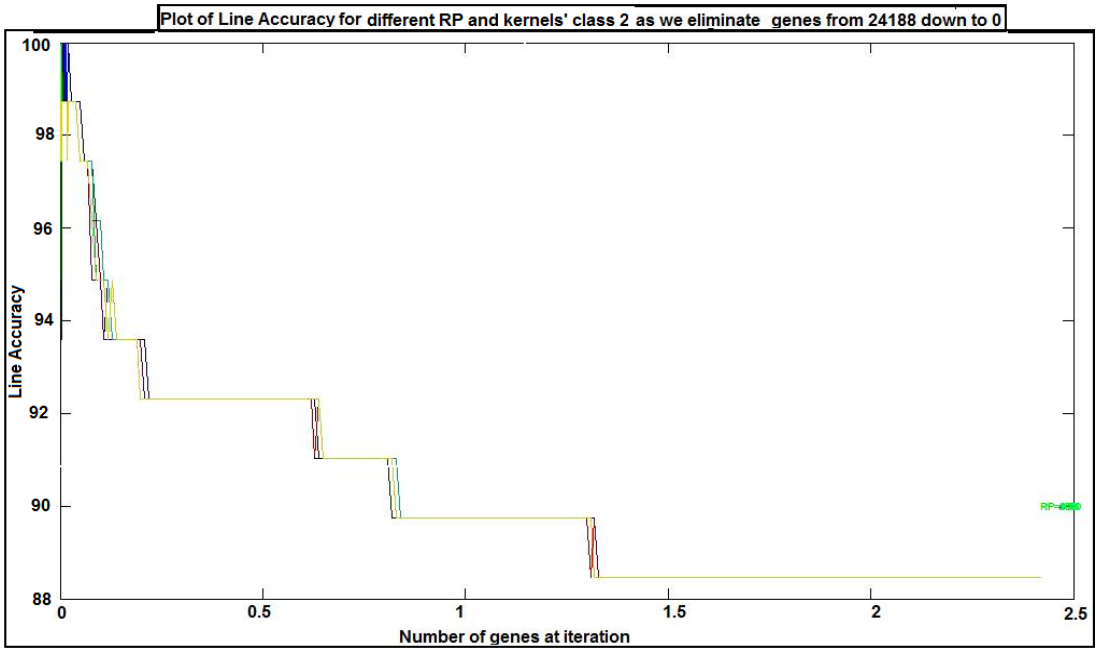
### 5.1.4 Προσδιορισμός βέλτιστου RP για ιδανικούς πυρήνες τάξης

Και στις δύο περιπτώσεις(πυρήνας τάξης 14 , πυρήνας τάξης 20) παρατηρήσαμε ότι η μεταβολή στις τιμές των παραμέτρων (Success Rate, Line Accuracy, Sensitivity, Specificity) καθώς μεταβάλλαμε την τιμή του rp είναι ελάχιστη(πυρήνας τάξης 20) ή μηδενική (πυρήνας τάξης 14) .

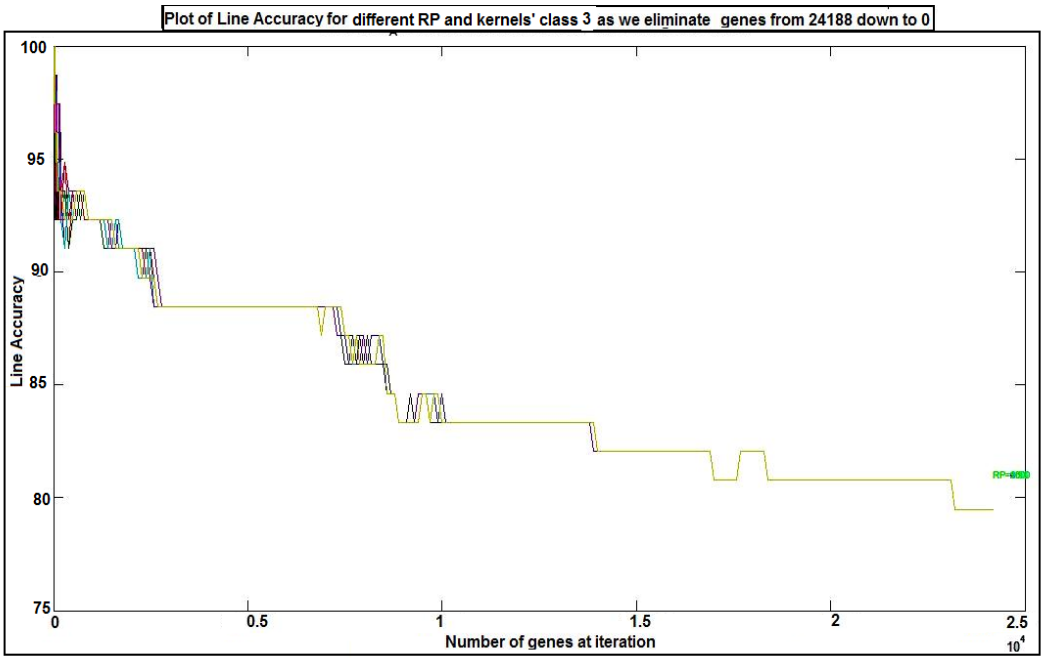
Για τους πυρήνες αυτούς η μεταβολή της τιμής της σταθερά κανονικοποίησης δεν επηρεάζει ουσιαστικά τον αλγόριθμο.

Παρατηρούμε δηλαδή ότι η Μεταβολή του RP έχει κατά βάση επίδραση για γραμμικό πυρήνα και για πυρήνες μικρής τάξης 2, 3 .

Το Διάγραμμα 2.15 και Διάγραμμα 2.16 απεικονίζουν ενδεικτικά οι διακυμάνσεις των τιμών της Line Accuracy για δύο πυρήνες μικρής τάξης 2(Διάγραμμα 2.15 ) , 3(Διάγραμμα 2.16) για διαφορετικές τιμές RP.Παρατηρούμε πως ακόμα και εδώ οι μεταβολές είναι ελάχιστες.




*Διάγραμμα 2. 15:Διακύμανση των τιμών του Line Accuracy καθώς αφαιρούμε σταδιακά γονίδια από 24188 σε 0 για πολυωνυμικό πυρήνα τάξης 2 και διάφορες τιμές RP.Παρατηρούμε πολύ μικρή μεταβολή στη γραφική που αναπαριστά τη Διακύμανση των τιμών του Line Accuracy κατά την αλλαγή της παραμέτρου Κανονικοποίησης RP.*

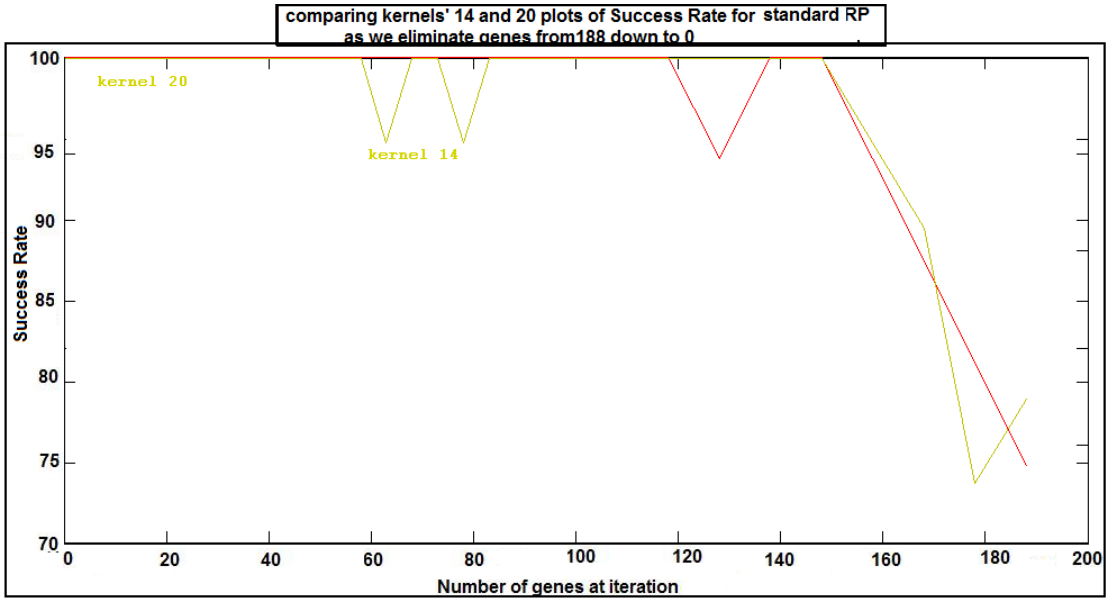



*Διάγραμμα 2. 16: Διακύμανση των τιμών του Line Accuracy καθώς αφαιρούμε Σταδιακά γονίδια από 188 σε 0 για πυρήνα τάξης 2 και διάφορες τιμές RP.Παρατηρούμε πολύ μικρή μεταβολή στη γραφική που αναπαριστά τη Διακύμανση των τιμών του Line Accuracy κατά την αλλαγή της παραμέτρου Κανονικοποίησης RP.*

### 5.1.5 Σύγκριση επιδόσεων για δύο καλύτερους πυρήνες

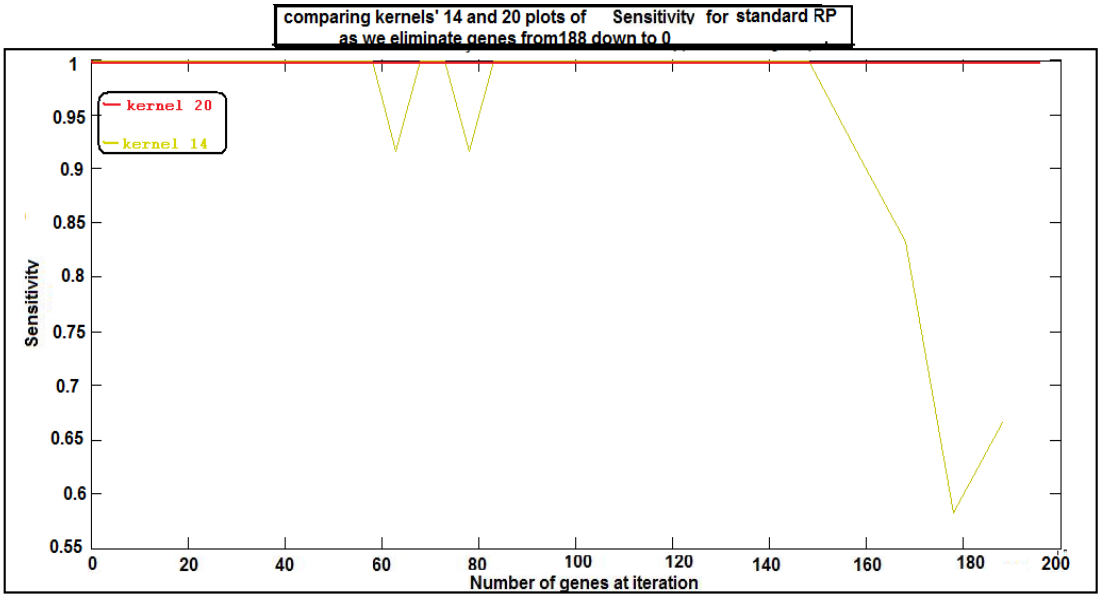
Παραθέτοντας τις γραφικές παραστάσεις που απεικονίζουν τις διακυμάνσεις των τιμών των παραμέτρων Line Accuracy, Success Rate, Sensitivity, Specificity θα συγκρίνουμε τους δύο καλύτερους πυρήνες: τον πολυωνυμικό πυρήνα τάξης 14 με τον πολυωνυμικό πυρήνα τάξης 20. Υπενθυμίζουμε από τις προηγούμενες παρατηρήσεις ότι για πυρήνες μεγάλης τάξης (όχι 2,3) η γραφική παράσταση που παρουσιάζει τις αλλαγές των τιμών των παραπάνω παραμέτρων καθώς σβήνουμε σταδιακά υποσύνολα γονιδίων είναι αμετάβλητη όσο και αν μεταβάλουμε την τιμή της σταθεράς κανονικοποιήσης RP.Αυτή η παρατήρηση ισχύει και για τους πυρήνες για τους οποίους αλγόριθμους εμφανίζει τις καλύτερες επιδόσεις (τους πολυωνυμικού πυρήνες τάξης 14, 20).Γι αυτό στην παρακάτω σύγκριση οι γραφικές που θα παρατηρήσουμε για τους πυρήνες τάξης 14, 20 είναι οι ίδιες για όλα τα RP που δοκιμάσαμε.

## A. Success Rate




*Διάγραμμα 2. 17: Διακύμανση των τιμών του Success Rate καθώς αφαιρούμε σταδιακά γονίδια από 188 σε 0 για πυρήνες τάξης 14 ,20 για μία σταθερό RP.Για κάθε έναν από τους πυρήνες (14,20) η γραφική των τιμών του Success Rate καθώς μειώνουμε τα γονίδια είναι η ίδια για οποιαδήποτε από τις τιμές της σταθερά κανονικοποιήσης RP:0,05 /0,5 / 100 /200/300/400/500/ 600/700/800900/1000/1500.Συνεπώς τα συμπεράσματα από τη σύγκριση των δύο πυρήνων ισχύουν αυτούσια για όλα τις παραπάνω τιμές RP .*

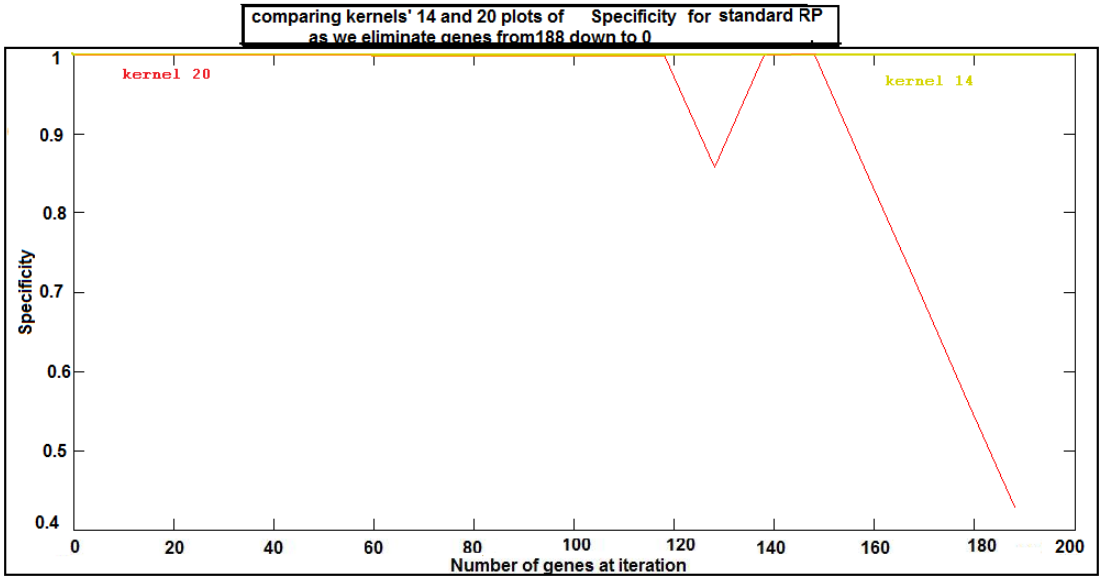
## B.Sensitivity




*Διάγραμμα 2. 18: Διακύμανση των τιμών του Sensitivity καθώς αφαιρούμε σταδιακά γονίδια από 188 σε 0 για πυρήνες τάξης 14,20 για σταθερό RP.Για κάθε έναν από τους πυρήνες (14,20) η γραφική των τιμών του Sensitivity καθώς μειώνουμε τα γονίδια είναι η ίδια για*

*οποιαδήποτε από τις τιμές της σταθερά κανονικοποιήσης RP:0,05 /0,5 / 100 /200/300/400/500/ 600/700/800900/1000/1500.Συνεπώς τα συμπεράσματα από τη σύγκριση των δύο πυρήνων ισχύουν αυτούσια για όλα τις παραπάνω τιμές RP .*

## Γ. Specificity



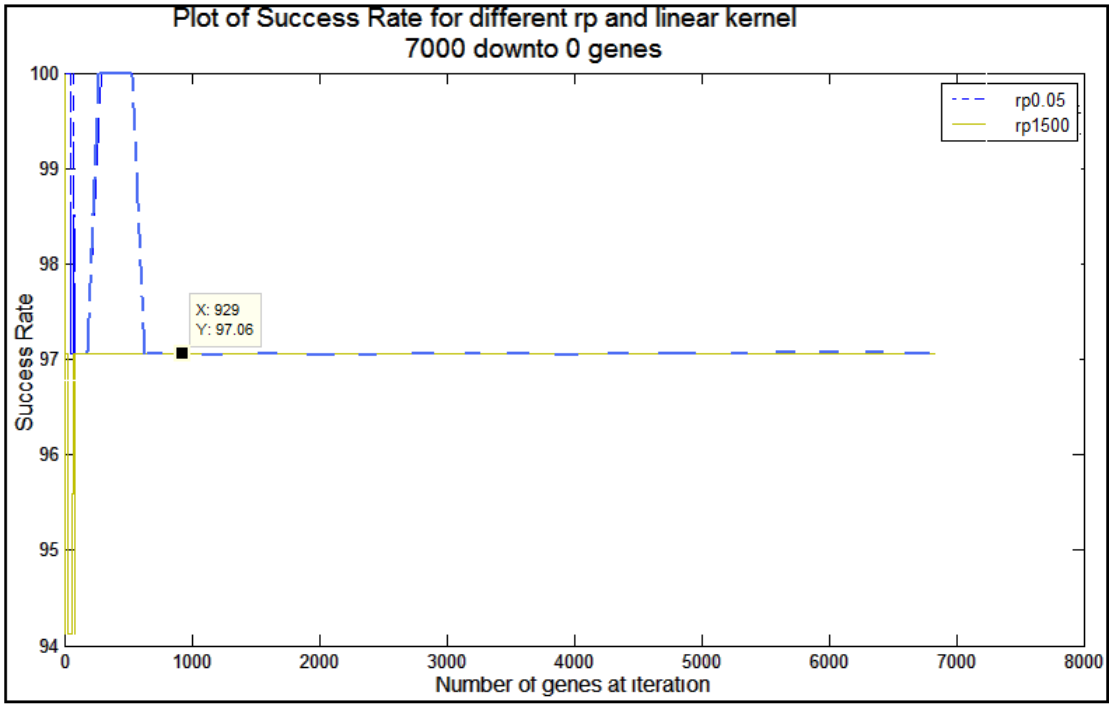

*Διάγραμμα 2. 19: Διακύμανση των τιμών του Specificity καθώς αφαιρούμε σταδιακά γονίδια από 188 σε 0 για πυρήνες τάξης14,20 για σταθερό RP.Για κάθε έναν από τους πυρήνες (14,20) η γραφική των τιμών του Specificity καθώς μειώνουμε τα γονίδια είναι η ίδια για οποιαδήποτε από τις τιμές της σταθερά κανονικοποιήσης RP:0,05 /0,5 /100 /200/300/400/500/ 600/700/800900/1000/1500.Συνεπώς τα συμπεράσματα από τη σύγκριση των δύο πυρήνων ισχύουν αυτούσια για όλα τις παραπάνω τιμές RP .*

Όπως διαπιστώνουμε εύκολα από τη σύγκριση των διαγραμμάτων 2.17,2.18,2.19,2.20 **για το σύνολο των παραμέτρων βελτιστοποίηση έχουμε με τον πυρήνα τάξης 20.**

## *5.2 Δεύτερο Dataset*

### 5.2.1 Προσδιορισμός βέλτιστης τιμής του RP για γραμμικό πυρήνα

### A. Επίδραση στην παράμετρο Success Rate

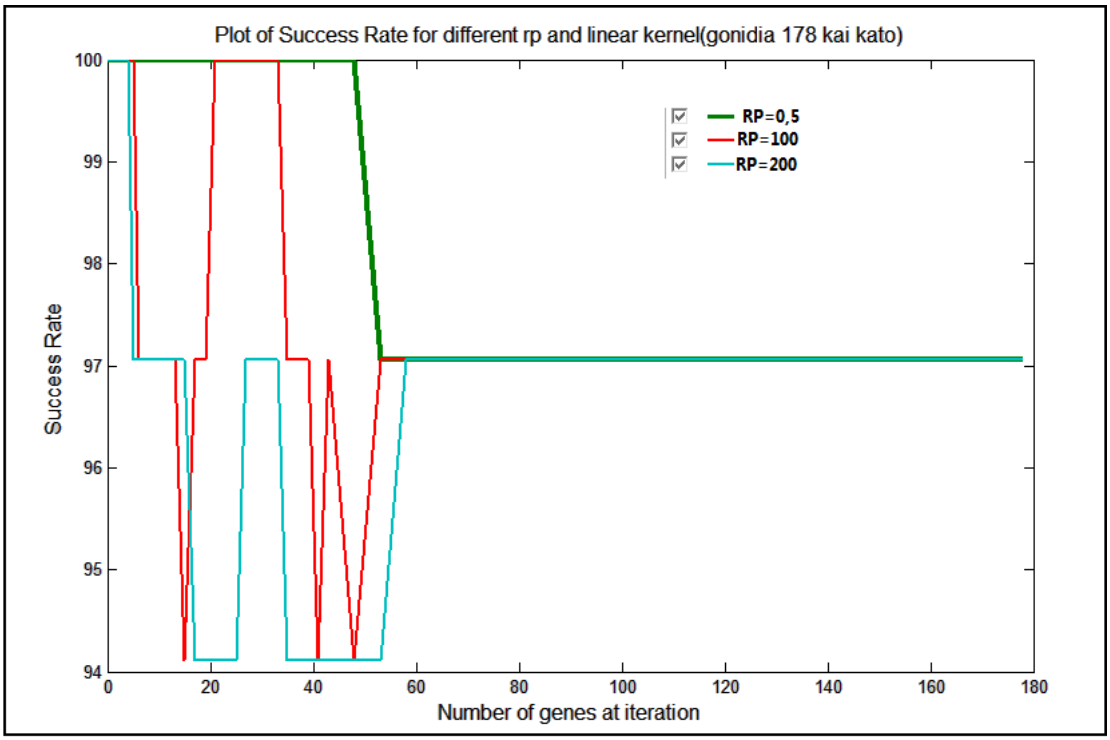



*Διάγραμμα 2. 20: Διακύμανση των τιμών του Success Rate καθώς αφαιρούμε σταδιακά γονίδια από 7000 σε 0 για γραμμικό πυρήνα για RP=0.05,1500.*

Στο Διάγραμμα 2.20 η γραμμή που απεικονίζει τις διακυμάνσεις τιμών της παραμέτρου Success Rate για RP=1500 αντιπροσωπεύει και τις τιμές RP=0,5,100:100:1000 , για τις οποίες παρατηρήσαμε ότι έχουν σχεδόν ίδιες γραφικές.

Παρατηρήσαμε ότι οι τιμές του Success Rate είναι στην τιμή 97,06 για το μεγαλύτερο διάστημα των επαναλήψεων και για όλες τις τιμές RP που δοκιμάσαμε μέχρι να φτάσουμε στα 629 γονίδια. Για 529-288 γονίδια η τιμή της παραμέτρου γίνεται μέγιστη (ίση με 100) μόνο για την περίπτωση του RP=0,05, ενώ για τα υπόλοιπα RP η τιμή της παραμέτρου παραμένει σταθερή μέχρι και τα 78 γονίδια οπότε και μειώνεται. Μετά τα 78 γονίδια παρατηρούμε υπάρχει γενικά αστάθεια στις μεταβολές των τιμών της παραμέτρου για όλα τα RP.Θα μπορούσαμε γενικά να πούμε ότι για RP=0.05έχουμε καλύτερες επιδόσεις. Η διαφοροποίηση(βελτίωση για RP=0,05) παρατηρείται από τα 629 γονίδια και κάτω.

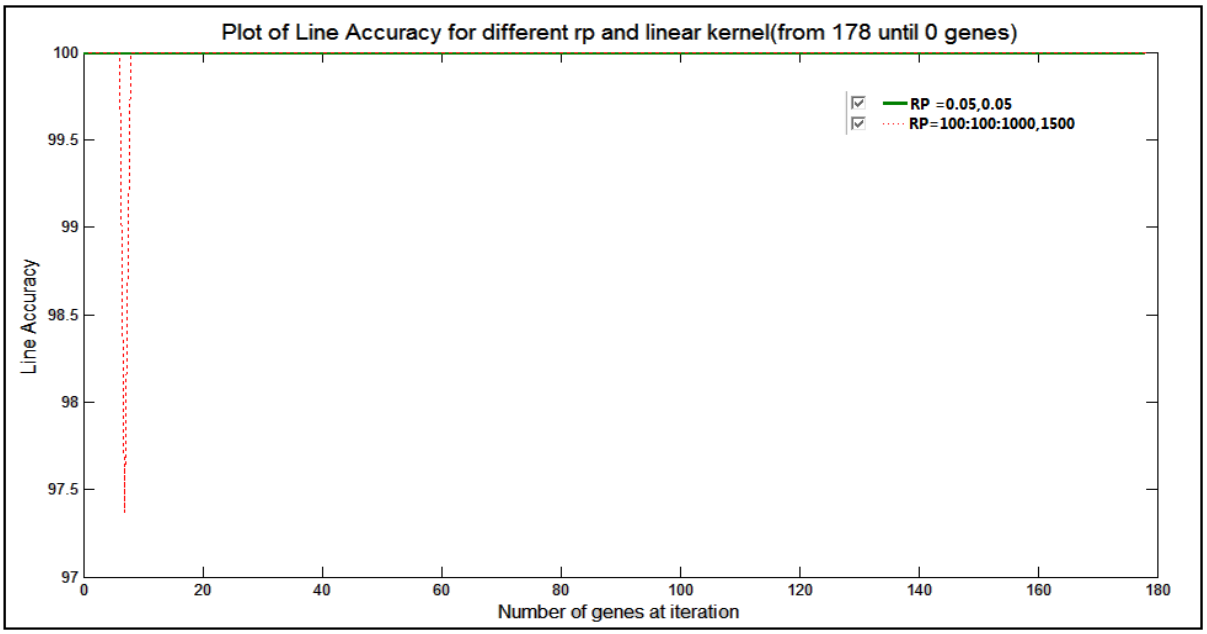
Γι αυτό και στην παρακάτω γραφική θα μελετήσουμε ιδιαίτερα τις μεταβολές στις τιμές του Success Rate για διάφορα RP για τα τελευταία 178 γονίδια.




*Διάγραμμα 2. 21: Διακύμανση των τιμών του Success Rate καθώς αφαιρούμε σταδιακά γονίδια από 178 σε 0 για γραμμικό πυρήνα και RP=0.05,100,1500.*

Παρατηρήσαμε τελικά ότι η ιδανική τιμή RP είναι 0,5 .Για RP=0,05 οι τιμές της παραμέτρου είναι σχεδόν ίδιες, αλλά εμφανίζουν πολλές απότομες διακυμάνσεις. Απότομες διακυμάνσεις παρατηρούνται για όλα τα υπόλοιπα RP.Επίσης ισχύει η παρατήρηση πως για όλα τα υπόλοιπα RP έχουμε χειροτέρευση των τιμών της παραμέτρου.

## Β. Επίδραση RP στην παράμετρο Line accuracy.



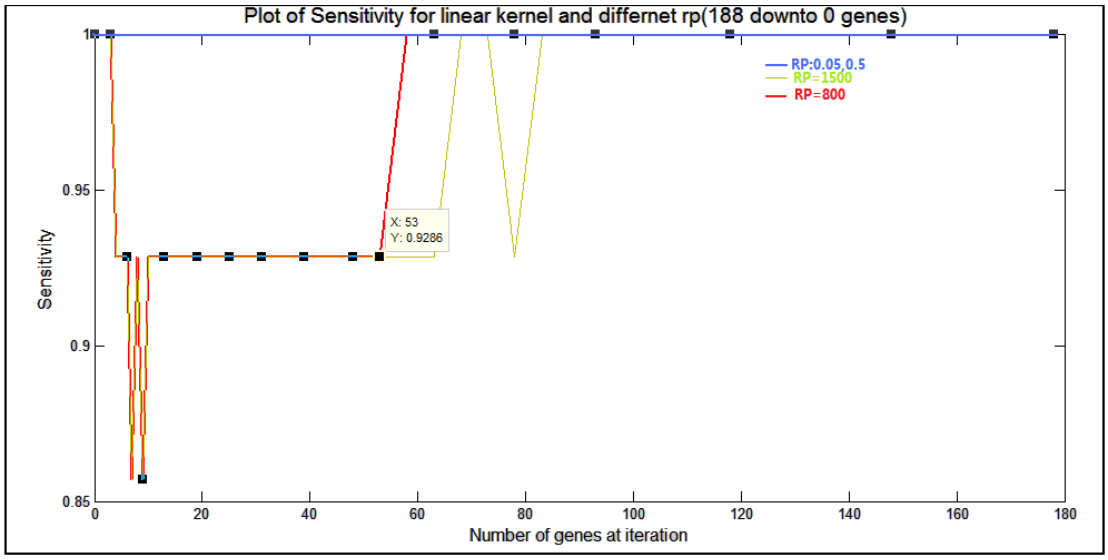

*Διάγραμμα 2. 22: Διακύμανση των τιμών του Line Accuracy καθώς αφαιρούμε σταδιακά γονίδια από 178 σε 0 για γραμμικό πυρήνα και RP=0.05,100:100:1000,1500.*

Παρατηρώντας τις διακυμάνσεις της τιμής της παραμέτρου Line Accuracy στο σύνολο των επαναλήψεων για όλα τα RP διαπιστώσαμε ότι

-η τιμή της παραμέτρου είναι 100 σε όλο το διάστημα των επαναλήψεων για RP=0.05,0.5 και

-η τιμή της παραμέτρου είναι 100 σχεδόν σε όλο το διάστημα των επαναλήψεων για RP=100:100:1000,1500 και 97,37 για αριθμό γονίδιων ίσο με 7 .

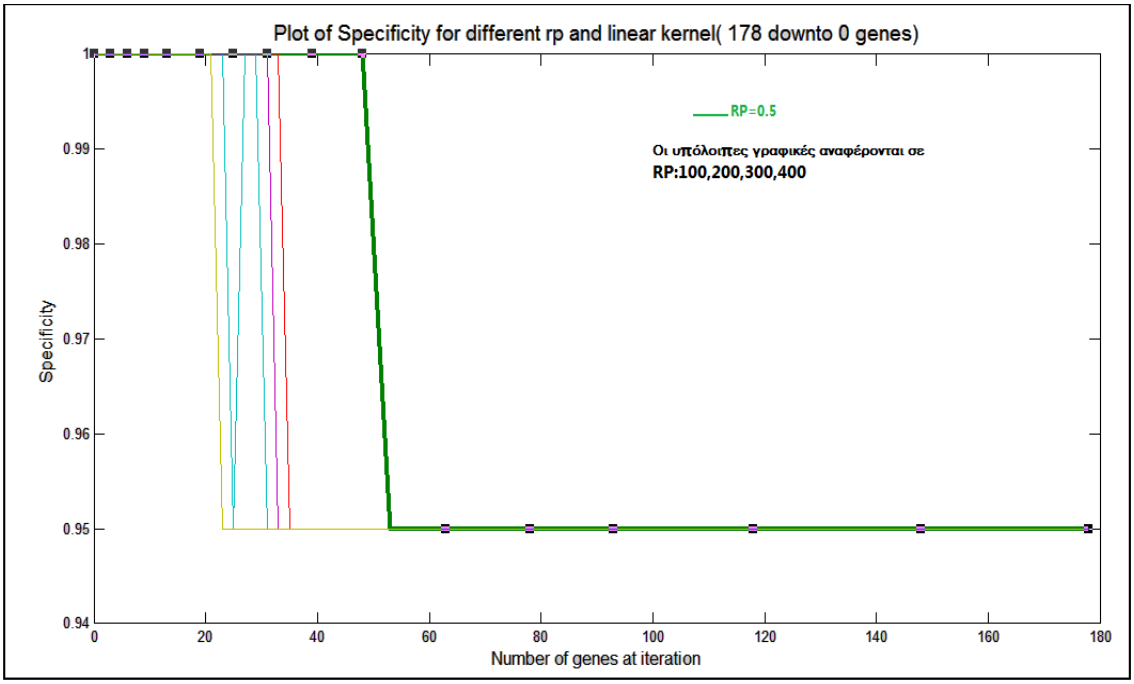
## Γ. Επίδραση RP στην παράμετρο Sensitivity




*Διάγραμμα 2. 23: Διακύμανση των τιμών του Sensitivity καθώς αφαιρούμε σταδιακά γονίδια από 188 σε 0 για γραμμικό πυρήνα και RP=0.5,0.05,800,1500.*

Και πάλι για RP=0.05,0.5 έχουμε τις βέλτιστες τιμές (100 σε όλο το διάστημα των επαναλήψεων).Για όλα τα άλλα RP παρατηρείται χειροτέρευση των τιμών και απότομες μεταβολές από τα 83 γονίδια και κάτω. Πριν φτάσουμε , ωστόσο στα 83 γονίδια η τιμή της Sensitivity είναι η βέλτιστη(100) για όλα τα RP.

### Δ. Επίδραση RP στην παράμετρο Specificity




*Διάγραμμα 2. 24: Διακύμανση των τιμών του Specificity καθώ ςαφαιρούμε σταδιακά γονίδια από 188 σε 0 για γραμμικό πυρήνα και RP=0.5,100,200,300,400*

-Για RP=0.5 έχουμε τις βέλτιστες και με μεγαλύτερη ευστάθεια τιμές στην παράμετρο Specificity.Πιο συγκεκριμένα η παράμετρος Specificity έχει τιμή 0.95 σε όλο το διάστημα των επαναλήψεων μέχρι και 53 γονίδια και 100 μέχρι να φτάσουμε στα 0 γονίδια.

-Για όλα τα άλλα RP παρατηρείται χειροτέρευση των τιμών και απότομες μεταβολές από τα 53 γονίδια και κάτω. Πριν φτάσουμε , ωστόσο στα 53 γονίδια η τιμή της Specificity είναι 0,95για όλα τα RP.

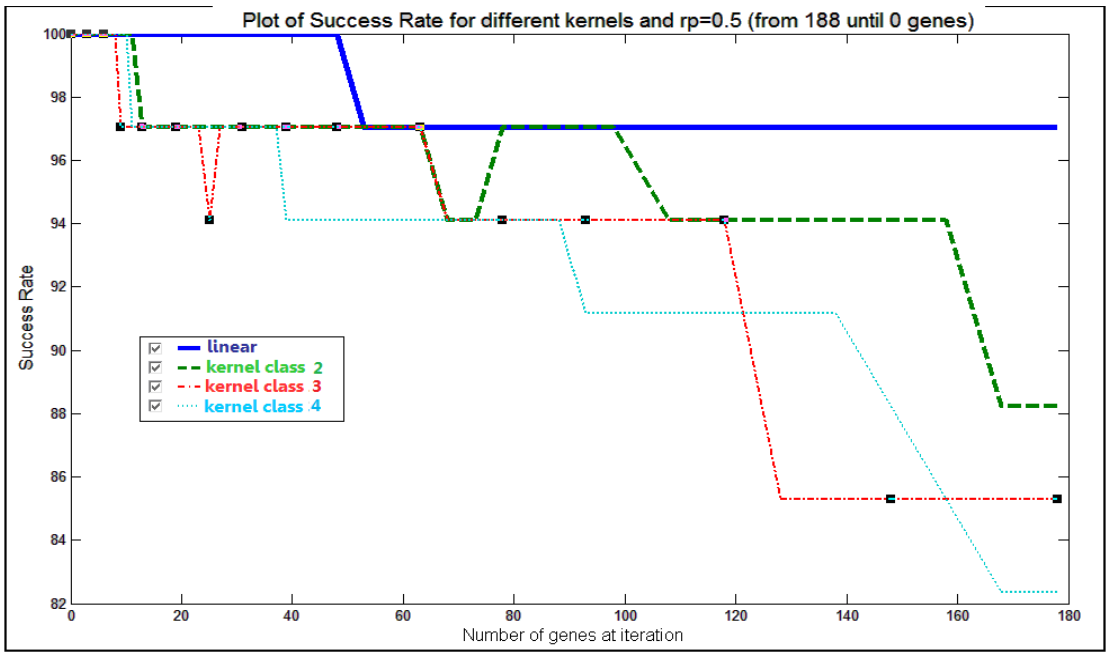
Τελικά συμπεραίνουμε ότι η ιδανική τιμή του RP είναι αρκετά μικρή , ίση με 0.5, για γραμμικό πυρήνα.

Στη συνέχεια με δεδομένο αυτό το RP, RP=0.5 θα δοκιμάσουμε την εναλλαγή της τάξης των πυρήνων ,ώστε μετά από σύγκριση της επίδρασης στις παραμέτρους Success Rate, Line Accuracy,Sensitivity, Specificity να καταλήξουμε στον καλύτερο /καλύτερους πυρήνες.

### 5.2.2 Προσδιορισμός βέλτιστου πυρήνα για τιμή του RP=0.5

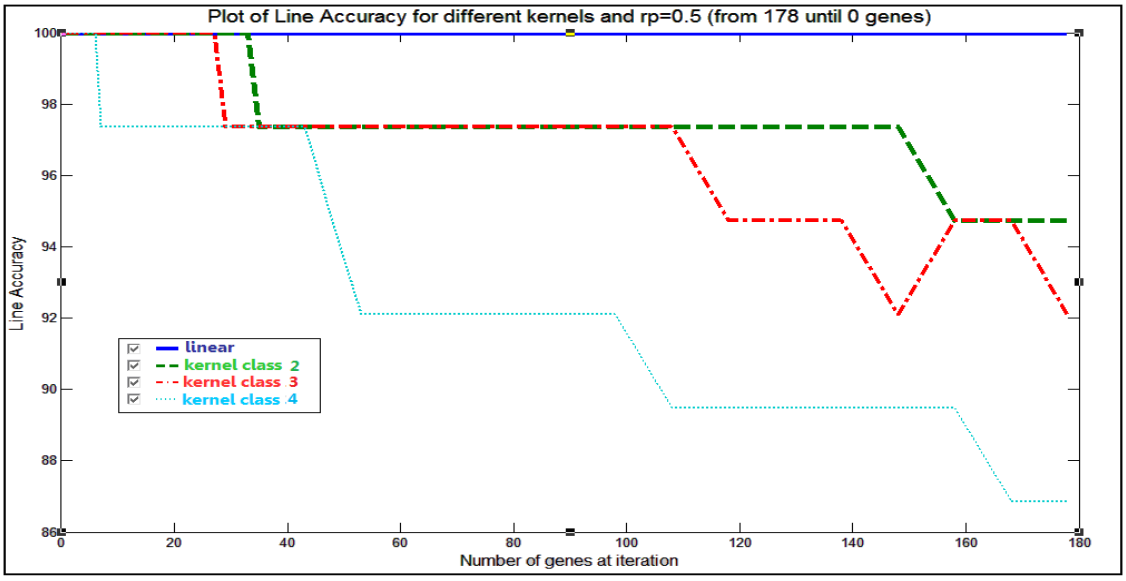
Παρατήρησα ότι γενικά καλύτερες είναι οι επιδόσεις για γραμμικό πυρήνα. Στα τελευταία περίπου 100 γονίδια για μερικές παραμέτρους (Sensitivity, Specificity) γίνεται καλύτερη η επίδοση για πυρήνα τάξης 3.Ακόμα είναι σημαντικό να σημειώσουμε ότι μετά την τάξη 3 όσο ανεβάζουμε την τάξη του πυρήνα τόσο χειροτερεύει η επίδοση.Οι παρατηρήσεις αυτές φαίνονται στα παρακάτω διαγράμματα:

**A. Επίδραση τάξης πυρήνα στην παράμετρο Success Rate**




*Διάγραμμα 2. 25: Διακύμανση των τιμών του Success Rate καθώς αφαιρούμε σταδιακά γονίδια από 188 σε 0 για RP=0.5 και διάφορους πυρήνες .Κάθε μία από τις παραπάνω γραφικές αντιστοιχεί σε πυρήνα διαφορετική τάξης. Απεικονίζονται οι γραφικές για πολυωνυμικούς πυρήνες τάξης 1,2,3,4*

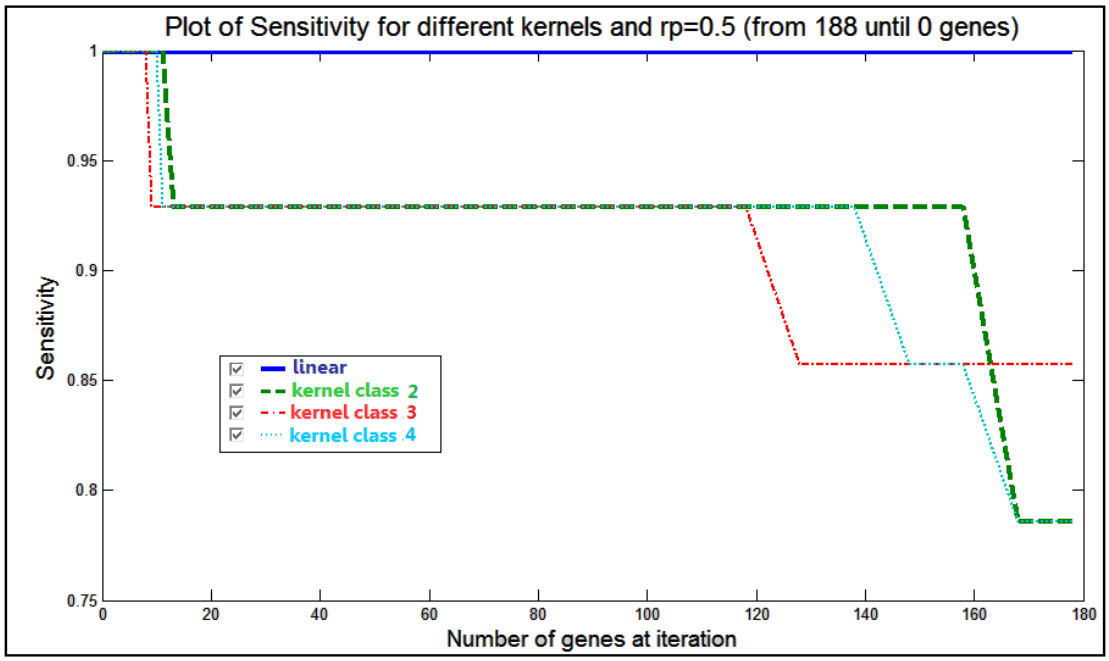
## Β. Επίδραση τάξης πυρήνα στην παράμετρο Line accuracy.




*Διάγραμμα 2. 26: Διακύμανση των τιμών του Line Accuracy καθώς αφαιρούμε σταδιακά γονίδια από 178 σε 0 για RP=0.5 και διάφορους πυρήνες .Κάθε μία από τις παραπάνω γραφικές αντιστοιχεί σε πυρήνα διαφορετική τάξης. Απεικονίζονται οι γραφικές για πολυωνυμικούς πυρήνες τάξης 1,2,3,4*

Στο Διάγραμμα 2.26 παρατηρούμε ότι η τιμή της παραμέτρου Line accuracy μειώνεται καθώς αυξάνουμε την τάξη του πυρήνα.

## Γ. Επίδραση τάξης πυρήνα στην παράμετρο Sensitivity

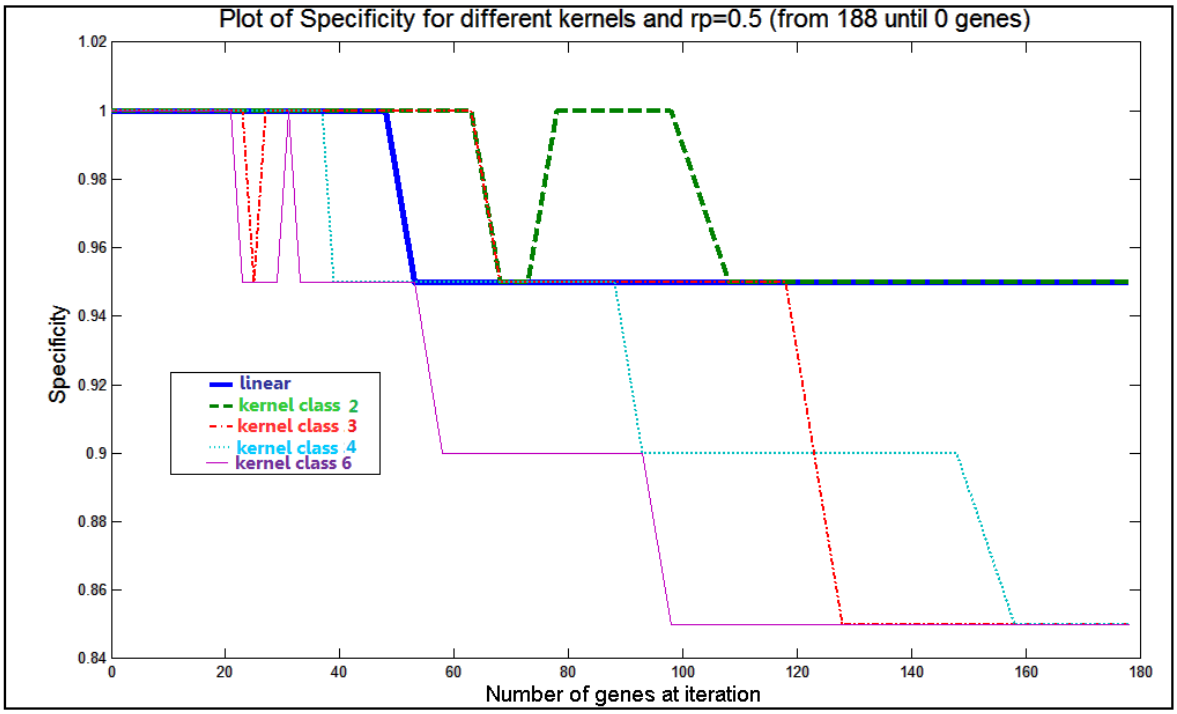



*Διάγραμμα 2. 27 Διακύμανση των τιμών του Sensitivity καθώς αφαιρούμε σταδιακά γονίδια από 178 σε 0 για RP=0.5 και διάφορους πυρήνες .Κάθε μία από τις παραπάνω γραφικές*

*αντιστοιχεί σε πυρήνα διαφορετική τάξης. Απεικονίζονται οι γραφικές για πολυωνυμικούς πυρήνες τάξης 1,2,3,4*

Στο Διάγραμμα 2.27 παρατηρούμε ότι η τιμή της παραμέτρου Sensitivity μειώνεται καθώς αυξάνουμε την τάξη του πυρήνα .Οι πυρήνες 1,2,3,4 απεικονίζουν ενδεικτικά όσα ισχύουν και για τους υπόλοιπους πυρήνες. Εξαίρεση αποτελεί ο πυρήνας τάξης 4 για τον οποίο παρατηρούμε βελτίωση έναντι του πυρήνα τάξης 3.

### Δ. Επίδραση τάξης πυρήνα στην παράμετρο Specificity



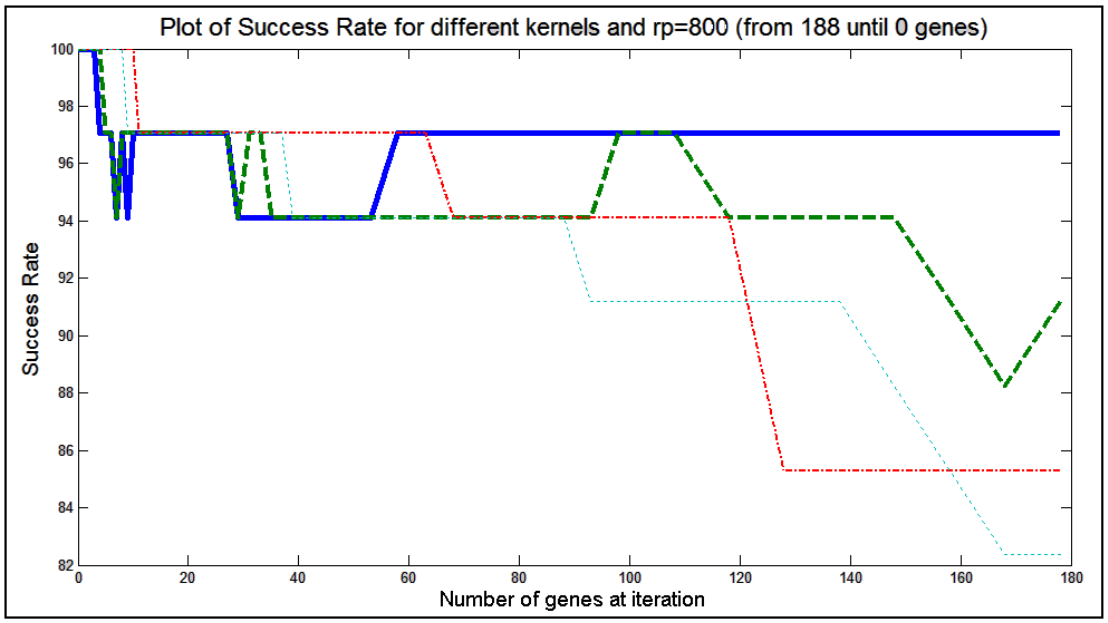

*Διάγραμμα 2. 28: Διακύμανση των τιμών του Specificity καθώς αφαιρούμε σταδιακά γονίδια από 178 σε 0 για RP=0.5 και διάφορους πυρήνες .Κάθε μία από τις παραπάνω γραφικές αντιστοιχεί σε πυρήνα διαφορετική τάξης. Απεικονίζονται οι γραφικές για πολυωνυμικούς πυρήνες τάξης 1,2,3,4*

Στο Διάγραμμα 2.28 παρατηρούμε ότι η τιμή της παραμέτρου Specificity μειώνεται καθώς αυξάνουμε την τάξη του πυρήνα από τον πυρήνα τάξης 4 και πάνω. .Οι πυρήνες 1,2,3,4,6 απεικονίζουν ενδεικτικά όσα ισχύουν και για τους υπόλοιπους πυρήνες. Εξαίρεση αποτελεί ο πυρήνας τάξης 2,3 για τους οποίους παρατηρούμε βελτίωση έναντι του πυρήνα τάξης 1(γραμμικός).

### 5.2.3 Προσδιορισμός βέλτιστου πυρήνα για δεδομένο RP=800

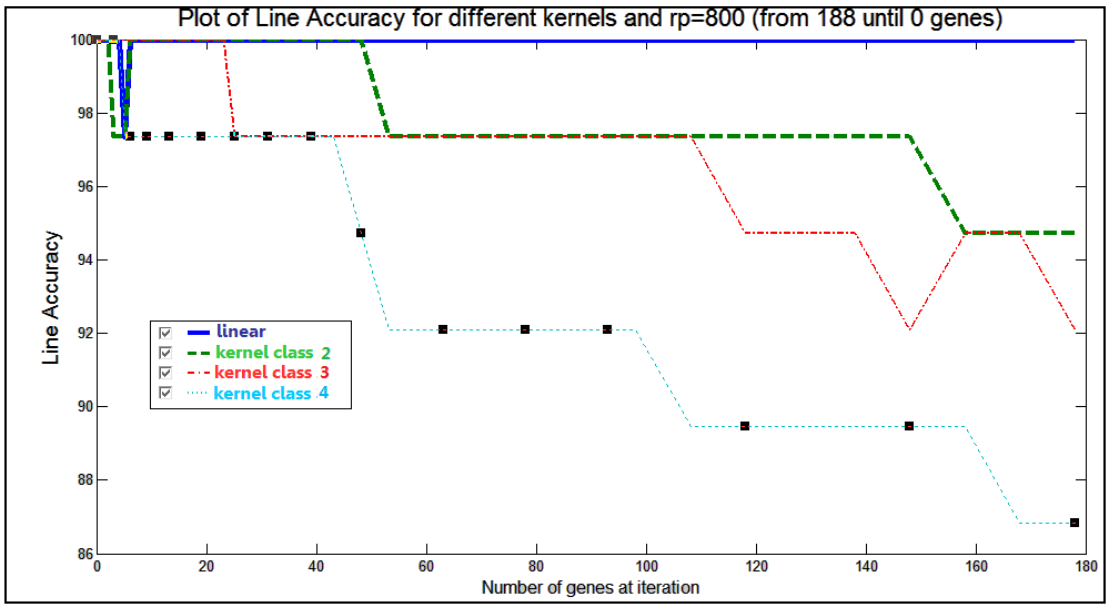
Παρατήρησα ότι γενικά καλύτερες είναι οι επιδόσεις για γραμμικό πυρήνα. Στα τελευταία περίπου 100 γονίδια για μερικές παραμέτρους (Sensitivity, Specificity) γίνεται καλύτερη η επίδοση για πυρήνα τάξης 3.Ακόμα είναι σημαντικό να σημειώσουμε ότι μετά την τάξη 3 όσο ανεβάζουμε την τάξη του πυρήνα τόσο χειροτερεύει η επίδοση.

**Α. Επίδραση της τάξης του πυρήνα στην παράμετρο Success Rate**



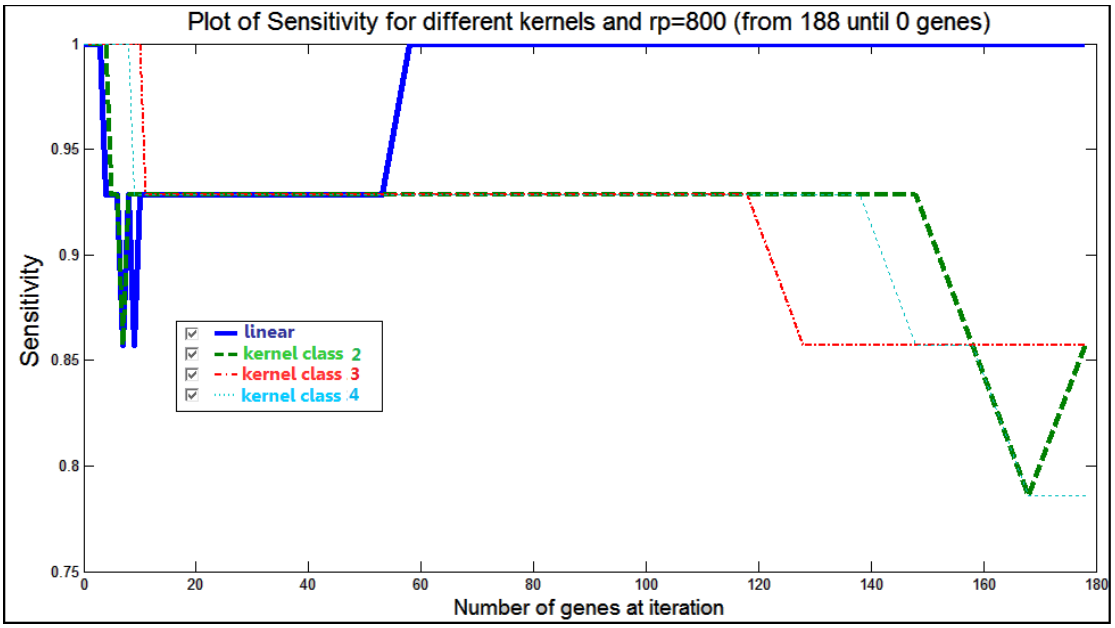

*Διάγραμμα 2. 29: Διακύμανση των τιμών του Success Rate καθώς αφαιρούμε σταδιακά γονίδια από 188 σε 0 για RP=800 και διάφορους πυρήνες. Κάθε μία από τις παραπάνω γραφικές αντιστοιχεί σε πυρήνα διαφορετική τάξης. Απεικονίζονται οι γραφικές για πολυωνυμικούς πυρήνες τάξης 1,2,3,4*

## Β. Επίδραση τάξης πυρήνα στην παράμετρο Line accuracy.



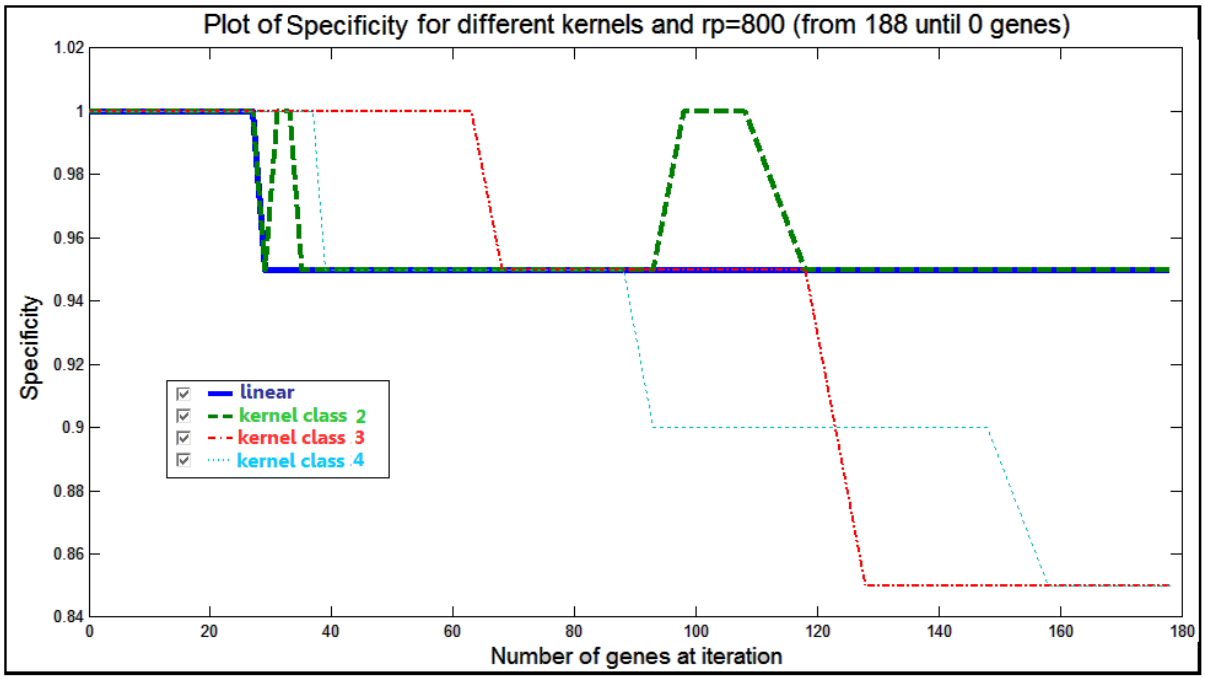

*Διάγραμμα 2. 30:Διακύμανση των τιμών του Line Accuracy καθώς αφαιρούμε σταδιακά γονίδια από 178 σε 0 για RP=800 και διάφορους πυρήνες .Κάθε μία από τις παραπάνω γραφικές αντιστοιχεί σε πυρήνα διαφορετική τάξης. Απεικονίζονται οι γραφικές για πολυωνυμικούς πυρήνες τάξης 1,2,3,4*

## Γ. Επίδραση τάξης πυρήνα στην παράμετρο Sensitivity




*Διάγραμμα 2. 31: Διακύμανση των τιμών του Sensitivity καθώς αφαιρούμε σταδιακά γονίδια από 178 σε 0 για RP=800 και διάφορους πυρήνες.Κάθε μία από τις παραπάνω γραφικές αντιστοιχεί σε πυρήνα διαφορετική τάξης. Απεικονίζονται οι γραφικές για πολυωνυμικούς πυρήνες τάξης 1,2,3,4*

## Δ. Επίδραση τάξης πυρήνα στην παράμετρο Specificity

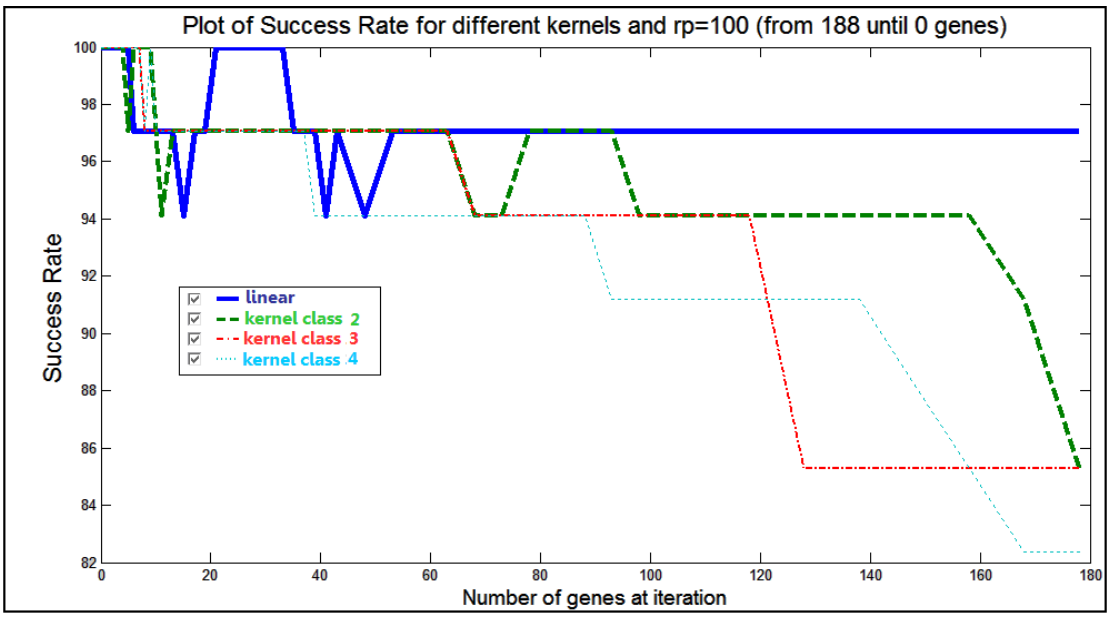



*Διάγραμμα 2. 32: Διακύμανση των τιμών του Specificity καθώς αφαιρούμε σταδιακά γονίδια από 178 σε 0 για RP=800 και διάφορους πυρήνες .Κάθε μία από τις παραπάνω γραφικές αντιστοιχεί σε πυρήνα διαφορετική τάξης. Απεικονίζονται οι γραφικές για πολυωνυμικούς πυρήνες τάξης 1,2,3,4*

Στο παραπάνω διάγραμμα παρατηρούμε ότι η τιμή της παραμέτρου Specificity μειώνεται καθώς αυξάνουμε την τάξη του πυρήνα από τον πυρήνα τάξης 4 και πάνω. Οι πυρήνες 1,2,3,4,6 απεικονίζουν ενδεικτικά όσα ισχύουν και για τους υπόλοιπους πυρήνες. Εξαίρεση αποτελεί ο πυρήνας τάξης 2,3 για τους οποίους παρατηρούμε βελτίωση έναντι του πυρήνα τάξης 1(γραμμικός).

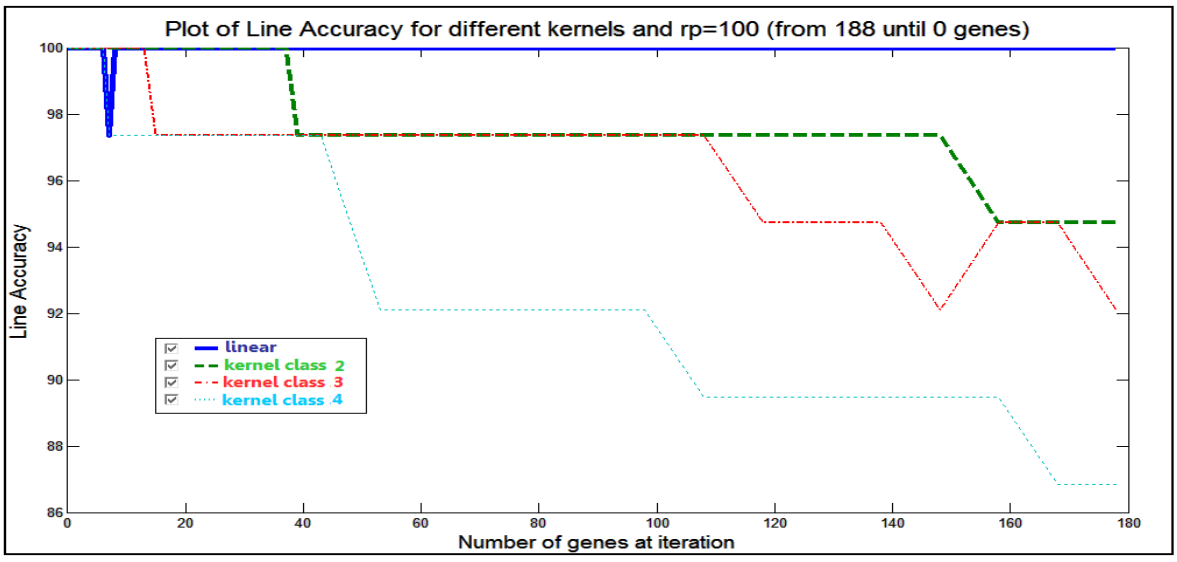
### 5.2.4 Προσδιορισμός βέλτιστου πυρήνα για δεδομένο RP=100

## Α. Επίδραση της τάξης του πυρήνα στην παράμετρο Success Rate



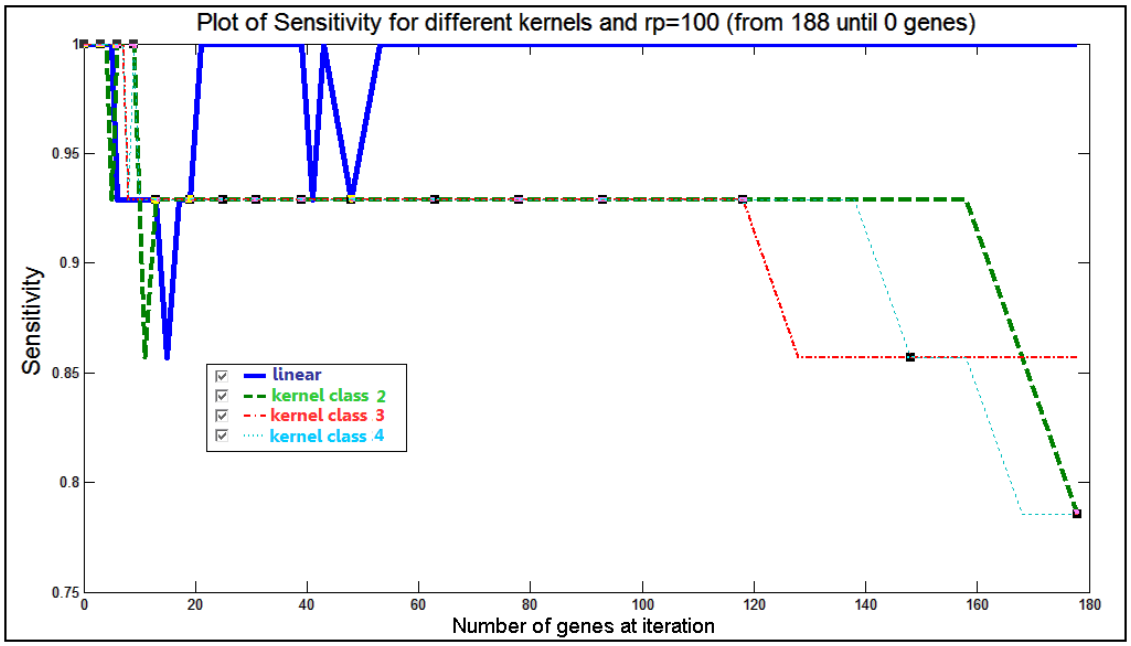

*Διάγραμμα 2. 33: Διακύμανση των τιμών του Success Rate καθώς αφαιρούμε σταδιακά γονίδια από 188 σε 0 για RP=100 και διάφορους πυρήνες .Κάθε μία από τις παραπάνω γραφικές αντιστοιχεί σε πυρήνα διαφορετική τάξης. Απεικονίζονται οι γραφικές για πολυωνυμικούς πυρήνες τάξης 1,2,3,4*

## Β. Επίδραση τάξης πυρήνα στην παράμετρο Line accuracy.



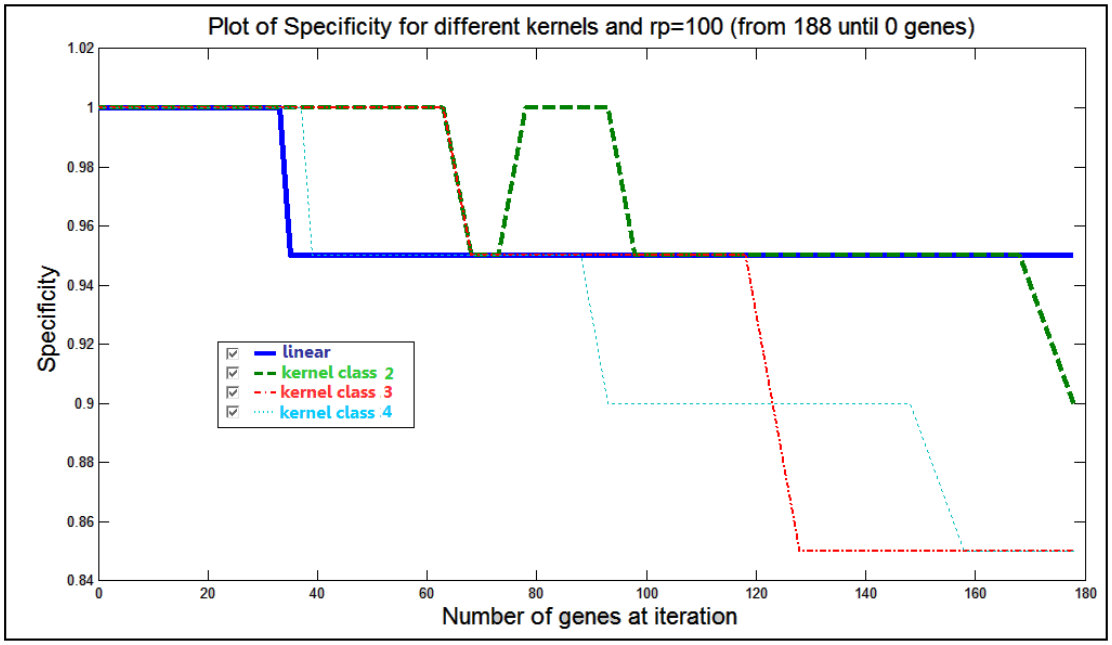

*Διάγραμμα 2. 34:Διακύμανση των τιμών του Line Accuracy καθώς αφαιρούμε σταδιακά γονίδια από 178 σε 0 για RP=100 και διάφορους πυρήνες .Κάθε μία από τις παραπάνω γραφικές αντιστοιχεί σε πυρήνα διαφορετική τάξης. Απεικονίζονται οι γραφικές για πολυωνυμικούς πυρήνες τάξης 1,2,3,4*

## Γ. Επίδραση τάξης πυρήνα στην παράμετρο Sensitivity




*Διάγραμμα 2. 35:Διακύμανση των τιμών του Sensitivity καθώς αφαιρούμε σταδιακά γονίδια από 178 σε 0 για RP=100 και διάφορους πυρήνες.Κάθε μία από τις παραπάνω γραφικές αντιστοιχεί σε πυρήνα διαφορετική τάξης. Απεικονίζονται οι γραφικές για πολυωνυμικούς πυρήνες τάξης 1,2,3,4*

## Δ. Επίδραση τάξης πυρήνα στην παράμετρο Specificity

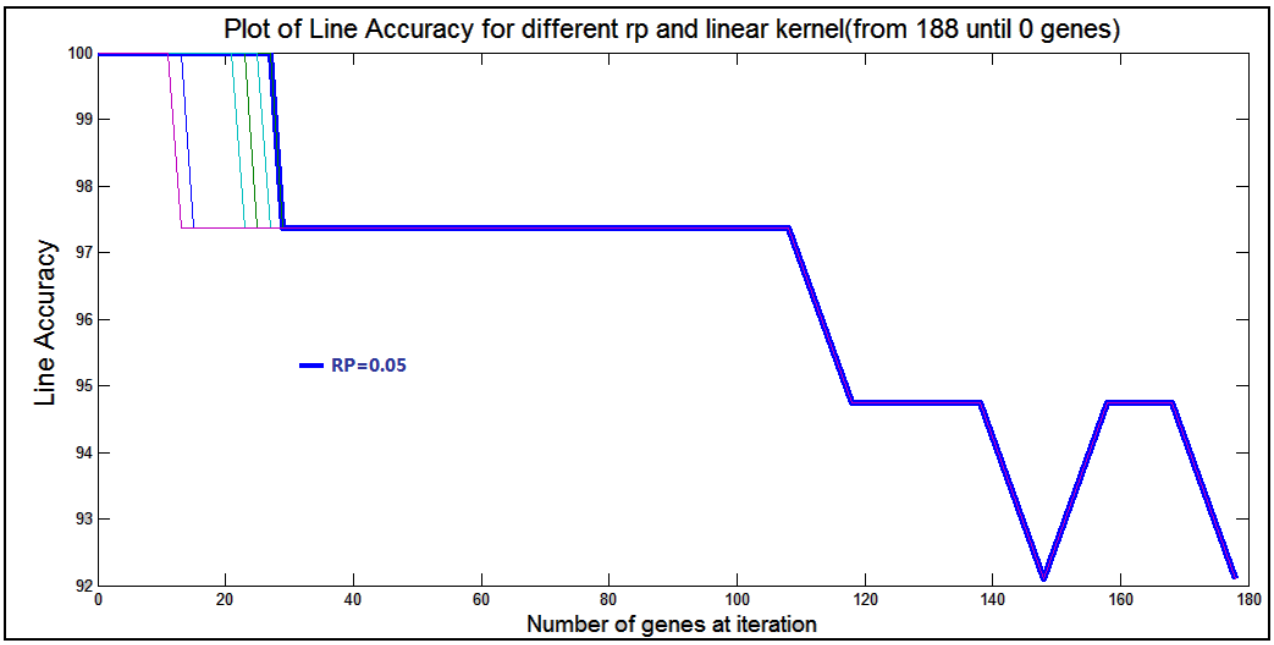



*Διάγραμμα 2. 36:Διακύμανση των τιμών του Specificity καθώ ςαφαιρούμε σταδιακά γονίδια από 178 σε 0 για RP=100 και διάφορους πυρήνες .Κάθε μία από τις παραπάνω γραφικές αντιστοιχεί σε πυρήνα διαφορετική τάξης. Απεικονίζονται οι γραφικές για πολυωνυμικούς πυρήνες τάξης 1,2,3,4*

Στο παραπάνω διάγραμμα παρατηρούμε ότι η τιμή της παραμέτρου Specificity μειώνεται καθώς αυξάνουμε την τάξη του πυρήνα από τον πυρήνα τάξης 4 και πάνω.Οι πυρήνες 1,2,3,4,6 απεικονίζουν ενδεικτικά όσα ισχύουν και για τους υπόλοιπους πυρήνες. Εξαίρεση αποτελεί ο πυρήνας τάξης 2,3 για τους οποίους παρατηρούμε βελτίωση έναντι του πυρήνα τάξης 1(γραμμικός).

### 5.2.5 Προσδιορισμός βέλτιστου RP για πυρήνα τάξης 3

Παρατηρήσαμε ότι για RP=0,05 έχουμε τις καλύτερες επιδόσεις , ενώ καθώς αυξάνουμε την τιμή του RP όλες οι παράμετροι σύγκρισής χειροτερεύουν. Ενδεικτικά παραθέτουμε το παρακάτω διάγραμμα.




*Διάγραμμα 2. 37: Διακύμανση των τιμών του Line Accuracy καθώς αφαιρούμε σταδιακά γονίδια από 178 σε 0 για γραμμικό πυρήνα και διάφορες τιμές RP .Κάθε μία από τις παραπάνω γραφικές αντιστοιχεί σε διαφορετική τιμή της σταθεράς κανονικοποίησης RP.Απεικονίζονται οι γραφικές για διάφορα RP από τα οποία βέλτιστο είναι το RP=0.05*

**Τελικά συμπεραίνουμε ότι για το δεύτερο dataset βέλτιστες επιδόσεις παρουσιάζει ο αλγόριθμος για γραμμικό πυρήνα και RP=0,05/0,5.**

# Κεφάλαιο6:Συμπεράσματα

## *6.1 Πρώτο Dataset*

- **Συμπεράσματα για την επίδραση της παραμέτρου κανονικοποιήσης RP στην απόδοση του αλγορίθμου KRR:** Καλύτερες τιμές για την παράμετρο RP για το πρώτο dataset που ελέγξαμε είναι οι τιμές 0,05,400,800.Οι πολύ μικρές τιμές RP είναι ιδανικές για τις τελευταίες εκατοντάδες γονιδίων , ενώ οι τιμές 400 και 800 είναι καλύτερες στο σύνολο. Υπενθυμίζουμε ότι η παράμετρος RP είναι αυτή που θέτει όριο στην τιμή του w βοηθώντας μας να καταπολεμήσουμε το πρόβλημα της συγγραμμικότητας των δεδομένων. Ιδιαίτερα έντονο είναι το τελευταίο πρόβλημα σε δεδομένα πολλών διαστάσεων (όπως τα δικά μας).Ακόμα όσο μεγαλύτερο είναι το RP , τόσο αυστηρότερο είναι το όριο για το W και τόσο καλύτερα καταπολεμούμε το πρόβλημα αυτό. Δεν είναι , λοιπόν , ανεξήγητο το γεγονός ότι για περισσότερα γονίδια ιδανικό είναι μεγαλύτερο RP από ότι για πολύ λιγότερα.
- **Συμπεράσματα για την επίδραση του πυρήνα στην απόδοση του αλγορίθμου KRR:** Σε σχέση με τη χρήση πυρήνα καταλήξαμε στο συμπέρασμα ότι οι πολυωνυμικοί πυρήνες 14 και 20 είναι οι καλύτεροι για το πρώτο dataset .Αξίζει να σημειώσουμε εδώ ότι οι βελτιώσεις με τη χρήση πυρήνων αρχίζουν να φαίνονται μόνο για τις τελευταίες εκατοντάδες γονιδίων. Αυτό είναι εξηγήσιμο, μιας και όταν ο χώρος μας είναι πολύ μεγάλων διαστάσεων το πιθανότερο είναι ο καλύτερος διαχωρισμός να επιτυγχάνεται με μια ευθεία. Αντίθετα όσο τα δεδομένα μειώνονται σε μερικές εκατοντάδες , ο χώρος μικραίνει, και ο διαχωρισμός του μπορεί να μην είναι εφικτός χωρίς τη χρήση υπερεπιπέδου. Επειδή τα δεδομένα μας συνεχίζουν να είναι πολλά και μη γραμμικώς διαχωρίσιμα , καλύτερες είναι οι επιδόσεις πολυωνυμικών πυρήνων μεγαλύτερης τάξης(πχ 14 ,20).Από ένα σημείο , όμως και μετά, η αύξηση της τάξης δημιουργεί υπερεπίπεδο ακατάλληλο για τον καλό διαχωρισμό των δεδομένων. Για τάξη πυρήνα 40 και πάνω οι επιδόσεις του αλγορίθμου μειώνονται.

## *6.2 Δεύτερο Dataset*

- **Συμπεράσματα για την επίδραση της παραμέτρου κανονικοποιήσης RP στην απόδοση του αλγορίθμου KRR:** Η τιμή της σταθερά κανονικοποιήσης για την οποία καταγράψαμε καλύτερη απόδοση του αλγορίθμου είναι αρκετά μικρή (ίση με 0.5/0.05).Το γεγονός αυτό συνάδει και με τα συμπεράσματα που καταγράψαμε για το πρώτο dataset ,όταν μειώσαμε αρκετά τα γονίδια. Το δεύτερο σύνολο δεδομένων είναι κατά πολύ μικρότερο του πρώτου και όπως εξηγήσαμε και στα συμπεράσματα που αφορούσαν το πρώτο dataset , για πιο μικρά σύνολα δεδομένων είναι κατάλληλη σταθερά κανονικοποιήσης μικρότερης τιμής.
- **Συμπεράσματα για την επίδραση του πυρήνα στην απόδοση του αλγορίθμου KRR:** Παρατηρήσαμε ότι για το δεύτερο dataset βέλτιστες επιδόσεις παρουσιάζει ο αλγόριθμος για γραμμικό πυρήνα πράγμα το οποίο εξηγείται και πάλι (όπως εξηγήσαμε στο πρώτο dataset) από κατά πολύ το μικρότερων διαστάσεων δεύτερο dataset.

Σε γενικές γραμμές θα μπορούσαμε να πούμε ότι οι απόδοση του αλγορίθμου ήταν ικανοποιητική σε ότι αφορά τις τιμές των εξεταζόμενων παραμέτρων. Επιπλέον η επίδραση του πυρήνα απλούστευσε κατά πολύ αρκετά πολύπλοκους υπολογισμούς που προέκυπταν από το μεγάλο όγκο δεδομένων.

# Κεφάλαιο7:Μελλοντική επέκταση

- Η περαιτέρω έρευνα στον αλγόριθμο Kernel Ridge Regression και η εφαρμογή της μεθόδου σε περισσότερα σύνολα δεδομένων, μπορεί να αποφέρει βελτιώσεις στη μέθοδο, αλλά και να καταδείξει τυχόν καταλληλότητα της μεθόδου για την πρόβλεψη καρκίνου συγκεκριμένων ειδών καρκίνου.
- Η μέθοδος Kernel Ridge Regression παρουσιάζει αρκετά κοινά χαρακτηριστικά με τον αλγόριθμο SVM .Θα είχε ιδιαίτερο ενδιαφέρον η σύγκριση των δύο αλγορίθμων.
- Μιας και η βελτιστοποίηση του Kernel Ridge Regression εξαρτάται σε μεγάλο βαθμό από τον προσδιορισμό της ιδανικής σταθεράς κανονικοποιήσης, θα μπορούσαν να μελετηθούν περισσότερες μέθοδοι προσδιορισμού της σταθεράς αυτής.

# Βιβλιογραφία


**[1]** Watson JD, Crick FH. "*Molecular Structure of Nucleic Acids. A Structure for Deoxyribose Nucleic Acid*." Nature ,1953 April 25,Vol.171;737

**[2]** Lockhart DJ, Winzeler EA. "*Genomics, gene expression and DNA arrays*.", Nature,2000 Jun 15,405(6788):827- 36. Review.

**[3]** R.O.Duda, P.E.Hart and D.G.Stork, "*Pattern Classification*, 2nd ed." NewYork Wiley, 2001".

**[4]** Momiao Xiong, Wuju Li, Jinying Zhao, Li Jin and Eric Boerwinkle "*Feature (Gene) Selection in Gene Expression-Based Tumor Classification"* Molecular Genetics and Metabolism **73,** 239–247 (2001)

**[5]** Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES, "*Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*." Science. 1999 Oct 15;286 (5439):531-7

**[6]** M.Brownetal., "Knowledge-based analysis of microarray gene expression data by using support vector machines," National Academy Sciences, vol. 97, 2000, pp. 262–267

**[7]** Gavin C. Cawley, Nicola L. C. Talbot and Olivier Chapelle," Estimating Predictive Variances with Kernel Ridge Regression"

**[8]** Cristianini, N., Shawe-Taylor, J."An Introduction to Support Vector Machines (and other kernel-based learning methods)." Cambridge University Press, Cambridge, U.K. ,2000

**[9]** Hoerl, A.E., Kennard, R.W "Ridge regression: Biased estimation for nonorthogonal problems." Technometrics 12,1970, 55–67

**[10]** Schoelkopf, B., Smola, A.J."Learning with kernels - support vector machines, regularization, optimization and beyond." MIT Press, Cambridge, 2002

**[11]** Guy ,"Linear Regression and Least Squares Estimation", Lebanon ,April 25, 2007 (http://www.stat.purdue.edu/~lebanon/notes/linReg.pdf)

**[12]** Alexei Pozdnoukhov "The analysis of Kernel Ridge regression Learning Algorithm",IDIAP Research report, November 2002

**[13]** Sona Mardikyan and Eyüp Çetin ,"Efficient Choice of Biasing Constant for Ridge Regression", Int. J. Contemp. Math. Sciences, 2008, Vol. 3, no. 11, 527 - 536

**[14]** D.A. Belsley, K. Edwin and R.E. Welsch, "Regression Diagnostics: Identifying Influential Data and Sources of Collinearity" Wiley, New York, 1980.

**[15]** D.M. Hawkins and X. Yin, “A faster algorithm for ridge regression, Computational Statistics & Data Analysis” 40 ,2002, 253-262.

**[16]** T.S. Lee, “Optimum ridge parameter selection”, Applied Statistics, 36(1) ,1988, 112-118.

**[17]** Fan Li and Yiming Yang,Language Technology Institute, “Analysis of recursive gene selection approaches from microarray data”, *bioinformatics 2005, Vol. 21, no. 19, pages 3741–3747*

**[18]** Isabelle Guyon, Jason Weston,Stephen Barnhill, Vladimir Vapnik, “Gene Selection for Cancer Classification using Support Vector Machines” Machine Learning, , 2002 46, 389–422

**[19]** Saunders, C., Gammerman, A., Vovk, V ,”Ridge regression learning algorithm in dual variables.” In: Proc., 15th Int. Conf. on Machine Learning, Madison,1998, p.515–521

**[20]** *Cawley, G. C., N. L.C. Talbot and O. Chapelle:* ,”Estimating Predictive Variances with Kernel Ridge Regression. Machine Learning Challenges” ,Max Plank Institute for Biological Cybernetics,First PASCAL Machine Learning Challenges Workshop ,2005,p. 56-77. (Eds.)

**[21]** A.Bjoerkstroem, R.Sundberg ,” A Generalized View on Continuum Regression”, Board of the Foundation of the Scandinavian Journal of Statistics,1999

**[22]** T. Poggio, F. Girosi,” Regularization Algorithms for Learning that are Equivalent to Multilayer Networks”, *Science*, New Series, 1990, Vol. 247, No. 4945., pp. 978-982.

**[23]** T. Sounapoglou ,”Algorithm fusion for selection genetic markers from big genetic databases”,2006

**[24]** Kounelakis M., Zervakis M., Kotsiakis X. ,“The impact of Microarray Technology in Brain Cancer”, A Taktak & A. C. Fisher (ed): A Multi-Perspective View on Outcome Prediction in Cancer, Elsevier Publishing Company, Section 13, 2006

**[25]** Blazadonakis M., Zervakis M.,Kounelakis M.,Beganzoli E., Lamma N. “ Support Vector Machines and Neural Networks as Marker Selectors for Cancer Gene Analysis” 3rd IEEE International Conference on Intelligent Systems (IS), pp.626-630, London, U.K. 4-6/9/ 2006

**[26]** Blazadonakis M., Zervakis M., “Recursive Feature Elimination by Support Vector Machine and Neural Networks”, Artificial Intelligence in Medicine, Sept. 2006

**[27]** Αρβανίτη Ευτυχία, «Ποσοτική Ανάλυση Της Χωρικής Κατανομής Των Εντάσεων Δυναμικών Οπτικών Σημάτων: Εφαρμογή Στην Εκτίμηση Του Ρίσκου Μετεξέλιξης Νεοπλασμάτων Σε Καρκίνο»,2008

**[28]** Moraitis Anastasios ,«Design and implentation on microprocessor remote control fuzzy logic controller for the management of food for the aquaculture industry »,2006

# Λεξιλόγιο-Όροι-Συντομογραφίες

$a_t$ **:** συντελεστής Lagrange.Δυική μεταβλητή του w.

**feature space:** ο χώρος χαρακτηριστικών(γονιδίων για την περίπτωσή μας)

**K:**kernel:η συνάρτηση πυρήνας.

**KRR**-Kernel Ridge Regression: Μέθοδος παλινδρόμησης κορυφής με χρήση πυρήνα-κανονικοποιημένη μέθοδος των ελαχίστων τετραγώνων.

**I**: σταθερά κανονικοποίησης

**pattern**: πρότυπο-δείγμα-ο ασθενής (στην περίπτωσή μας)

**RR**-Ridge Regression: Κανονικοποιημένη παλινδρόμηση ή παλινδρόμηση κορυφής.

**RP:**Ridge Parameter: Συμβολίζουμε με RP την σταθερά κανονικοποίησης της RR.Σε επίπεδο υλοποίησης η σταθερά κανονικοποίησης συμβολίζεται με RP,ενώ σε όλες τις εξισώσεις την συμβολίζουμε με I.

**SVM-Support Vector Machines:** Μηχανές διανυσμάτων υποστήριξης.

**w:**Διάνυσμα βαρών .Το μέγεθός του ισούται με το πλήθος των χαρακτηριστικών. Δείχνει κατά πόσο κάθε χαρακτηριστικό συμβάλει στη διαμόρφωση της κλάσης εξόδου. Καθορίζει την κλίση της ευθείας ή το υπερεπιπέδου διαχωρισμού