ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Ανάλυση απόδοσης με όρια εμπιστοσύνης για αλγόριθμους επιλογής γονιδίων»

ΜΠΑΝΤΗ ΜΑΡΙΑ-ΑΛΙΚΗ
2001030096

Εξεταστική Επιτροπή
Ζερβάκης Μιχάλης (επιβλέπων)
Μπάλας Κωνσταντίνος
Καρυστινός Γεώργιος

Ευχαριστίες

Αρχικά, θέλω να ευχαριστήσω τον καθηγητή μου, κύριο Ζερβάκη για την πολύτιμη βοήθεια και αρωγή του στην εκπλήρωση αυτής της διπλωματικής εργασίας.

Επίσης, θέλω να ευχαριστήσω τον υποψήφιο διδάκτορα Μιχάλη Μπλαζαντωνάκη για τη διαρκή του υποστήριξη, καθ' όλη την περίοδο της εργασίας μέχρι και την εκπλήρωσή της.

Τέλος, ευχαριστώ τα μέλη της εξεταστικής επιτροπής για την παρουσία και την ενασχόλησή τους με το κείμενο της διπλωματικής εργασίας.

Περιεχόμενα

Περιεχόμενα εικόνων- διαγραμμάτων- πινάκων

- **Εικόνες**

- **Διαγράμματα**

- **Πίνακες**

Κεφάλαιο 1: Εισαγωγή

1.1. Περίληψη

Σκοπός της εργασίας μας ήταν η ανάλυση της απόδοσης αλγορίθμων επιλογής γονιδίων με την χρήση ορίων εμπιστοσύνης. Για το σκοπό αυτό υλοποιήσαμε δύο σειρές πειραμάτων:

Αρχικά, επικεντρωθήκαμε σε δεδομένα που σχετίζονταν με τον καρκίνο του μαστού σε σύνολο 25.000 γονιδίων. Στο σύνολο αυτό εφαρμόσαμε 9 διαφορετικές μεθόδους επιλογής γονιδίων και αξιολογήσαμε την ακρίβεια τους χρησιμοποιώντας όρια εμπιστοσύνης. Στα πειράματα αυτά συμμετείχαν 78 ασθενείς οι οποίοι σε κάθε εφαρμογή της μεθόδου χωρίζονταν σε testing και training set, καθώς και ένα ανεξάρτητο σύνολο 19 ασθενών που κατηγοριοποιούταν σε κάθε εφαρμογή.

Η δεύτερη σειρά πειραμάτων σχετιζόταν με την λευχαιμία σε σύνολο 7.000 γονιδίων περίπου. Στο σύνολο αυτό εφαρμόσαμε 3 από τις παραπάνω 9 διαφορετικές μεθόδους και στη συνέχεια αξιολογήσαμε την ακρίβειά τους. Στα πειράματα αυτά συμμετείχαν 38 ασθενείς οι οποίοι κάθε φορά χωρίζονταν σε testing και training set, καθώς και ένα ανεξάρτητο σύνολο 34 ασθενών που κατηγοριοποιούταν σε κάθε εφαρμογή της μεθόδου.

1.2. Σκοπός της εργασίας

Τα δεδομένα της εργασίας μας, βασίζονται σε microarray γενετική έκφραση, η οποία χρησιμοποιείται ευρέως στην ογκολογική έρευνα όταν θέλουμε να προβλέψουμε κάποιο κλινικό αποτέλεσμα, όπως είναι το στιγμιότυπο μιας μετάστασης, ή η ανταπόκριση του ασθενούς σε μια συγκεκριμένη θεραπεία[10].

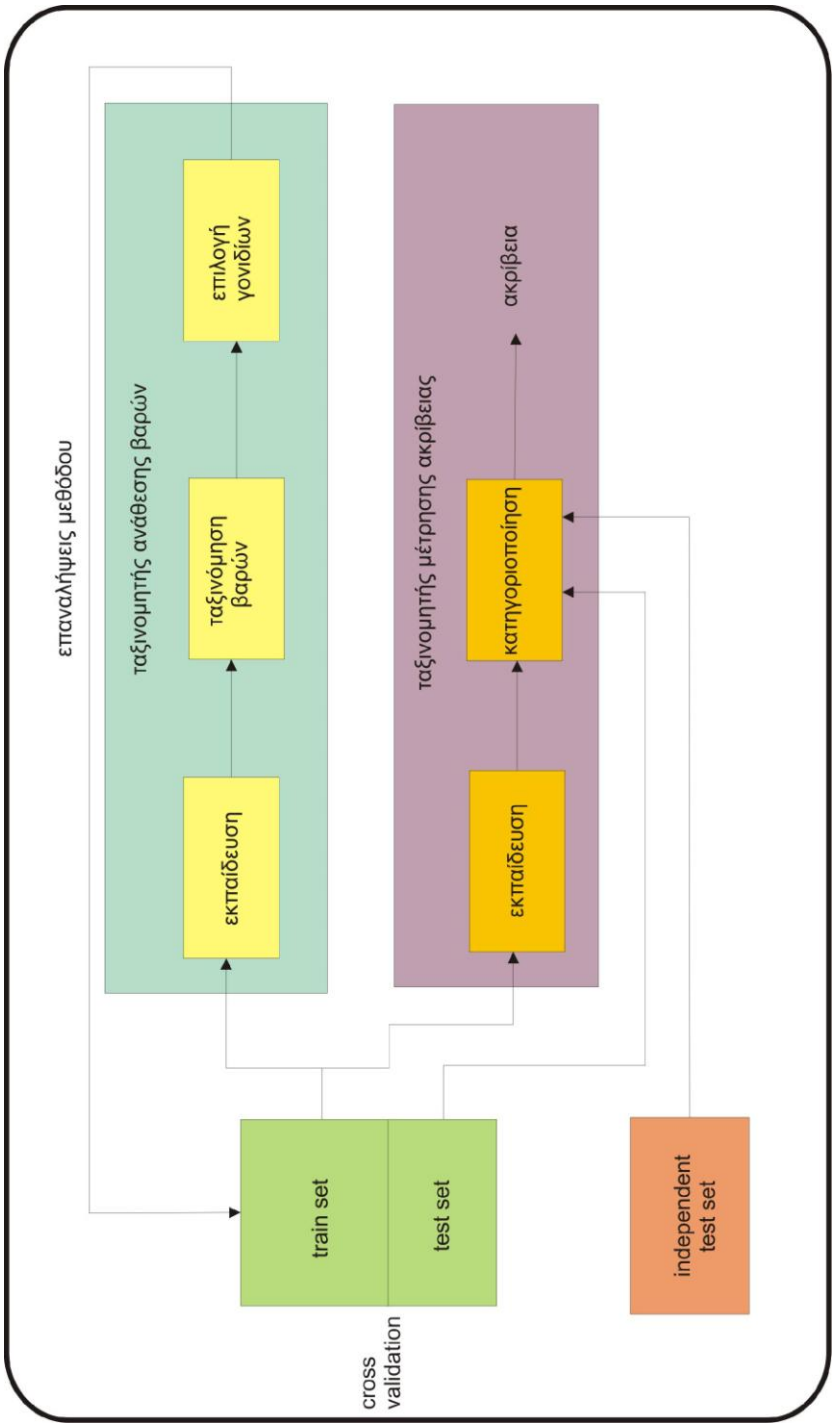
Τα microarray πειράματα συνδέονται κυρίως με εφαρμογές όπου έχουμε μικρό αριθμό περιπτώσεων (ασθενείς), αλλά μεγάλο αριθμό χαρακτηριστικών (γονίδια). Συνήθως, ο αριθμός των χαρακτηριστικών είναι τάξεις μεγέθους μεγαλύτερος από τον αριθμό των περιπτώσεων, όπως συμβαίνει και στην εργασίας μας όπου μελετάμε 78 περιπτώσεις με 25.000 χαρακτηριστικά στο ένα σύνολο δεδομένων και 38 περιπτώσεις με 7.000 χαρακτηριστικά στο δεύτερο σύνολο δεδομένων. Για κάθε περίπτωση, η πληροφορία που έχουμε αποτελείται από την γενετική έκφραση που προέρχεται από την microarray ανάλυση, καθώς και από μια ετικέτα η οποία μας δείχνει την κατηγορία στην οποία εντάσσεται η συγκεκριμένη περίπτωση (υγιής ή ασθενής).

Ένα μοντέλο πρόβλεψης, μπορεί να εκπαιδευτεί με τα δεδομένα που έχουμε και να προβλέψει στη συνέχεια το κλινικό αποτέλεσμα για μελλοντικές περιπτώσεις ασθενών. Συνήθως η αξιολόγηση ενός τέτοιου μοντέλου γίνεται με βάση το λάθος πρόβλεψης, παράμετρο που μελετούμε στην εργασία μας , εφαρμόζοντας επιπλέον και τα όρια εμπιστοσύνης αυτής.

Εφ ’όσον ο αριθμός των χαρακτηριστικών υπερβαίνει κατά πολύ των αριθμό των περιπτώσεων, θα πρέπει να επιλέξουμε ένα σύνολο γονιδίων το οποίο θα είναι το ελάχιστο δυνατό το οποίο θα μπορεί να προβλέψει το κλινικό αποτέλεσμα με την μέγιστη δυνατή ακρίβεια. Η διαδικασία αυτή είναι η διαδικασία «επιλογής χαρακτηριστικών»(feature selection) και υλοποιείται με 9 διαφορετικές μεθόδους, η απόδοση των οποίων αξιολογείται κάθε φορά.

Η κάθε μέθοδος εφαρμόζεται και σε κάθε επανάληψή της, κατηγοριοποιεί το train set και το test set που έχουν προκύψει από την διαδικασία του cross validation, καθώς και το independent test set. Στην ίδια επανάληψη υλοποιεί την διαδικασία επιλογής γονιδίων από την οποία προκύπτουν τα νέα train set, test set και independent test set που αποτελούνται από τους ίδιους ασθενείς αλλά με μικρότερο αριθμό γονιδίων. Τα νέα αυτά σύνολα, ανατροφοδοτούνται στην μέθοδο στην επόμενη επανάληψη, η οποία τα κατηγοριοποιεί και η διαδικασία επαναλαμβάνεται μέχρι να εξαλειφθούν όλα τα γονίδια.

Έτσι, στην κάθε επανάληψη, η μέθοδος χρησιμοποιεί δύο ταξινομητές, έναν για να κατηγοριοποιήσει τα δεδομένα και έναν για να κάνει επιλογή χαρακτηριστικών. Σε κάποιες μεθόδους αυτοί οι δύο ταξινομητές μπορεί να είναι και του ίδιου τύπου. Τα παραπάνω συνοψίζονται στην παρακάτω εικόνα:



Εικόνα 1:Σχηματικό διάγραμμα εφαρμογής μεθόδου για τον υπολογισμό του τελικού πίνακα

1.3 The Recursive Feature Elimination Approach (RFE)

Η προσέγγιση αυτή είναι η κεντρική ιδέα κάθε wrapper μεθόδου. Οι wrapper μέθοδοι εστιάζουν στην αλληλεπίδραση των γονιδίων μεταξύ τους και όχι στα εγγενή χαρακτηριστικά τους όπως οι μέθοδοι φίλτρου. Επιπλέον, τα βάρη των γονιδίων δεν παραμένουν σταθερά, αλλά επαναπροσδιορίζονται σε κάθε επανάληψη της μεθόδου και ενημερώνονται δυναμικά μέσω της προσέγγισης RFE.

Ο ταξινομητής ο οποίος χρησιμοποιείται για να υλοποιήσουμε την επιλογή χαρακτηριστικών όπως αναφέρθηκε παραπάνω, ορίζει ένα βάρος για κάθε χαρακτηριστικό (γονίδιο) και στη συνέχεια γίνεται ταξινόμηση των χαρακτηριστικών με βάση τις απόλυτες τιμές των βαρών τους.

Στη συνέχεια, το χαρακτηριστικό με το μικρότερο κατ' απόλυτη τιμή βάρος εξαλείφεται και η διαδικασία συνεχίζεται επαναληπτικά. Στο σημείο αυτό πρέπει να αναφέρουμε ότι αν παραπάνω από ένα χαρακτηριστικά έχουν την ίδια απόλυτη τιμή βάρους, τότε εξαλείφονται όλα μαζί. Σημειώνουμε ότι σε μια τέτοια προσέγγιση τα βάρη ενημερώνονται δυναμικά. Έτσι ένα γονίδιο που σε μια επανάληψη έχει χαμηλό βάρος, μπορεί να έχει υψηλή τιμή βάρους σε μία επόμενη επανάληψη.

Ο Guyon και οι συνεργάτες του [5] εφάρμοσαν μια τέτοια προσέγγιση σε συνδυασμό με έναν SVM ταξινομητή και κατάφεραν να προσδιορίσουν ένα συνδυασμό 7 γονιδίων τα οποία διαχωρίζουν πλήρως τους δύο τύπους λευχαιμίας (ALL and AML) .Σημειώνουμε ότι ο Golub και οι συνεργάτες του [2] κατέληξαν στο ίδιο αποτέλεσμα χρησιμοποιώντας μια μέθοδο φίλτρου με 50 γονίδια. Αυτή η βελτιωμένη απόδοση της wrapper προσέγγισης, άνοιξε νέους ερευνητικούς δρόμους στο πεδίο της microarray analysis.

Η RFE μέθοδος συνοψίζεται στα ακόλουθα βήματα:

1. Έστω m είναι ο αρχικός αριθμός χαρακτηριστικών(γονιδίων)
2. Όσο $(m \geq 0)$
3. Ενημέρωσε το διάνυσμα βαρών $\boldsymbol{w}$ χρησιμοποιώντας έναν ταξινομητή.
4. Ταξινόμησε τα γονίδια με βάση τις απόλυτες τιμές του διανύσματος

$\boldsymbol{w}$.

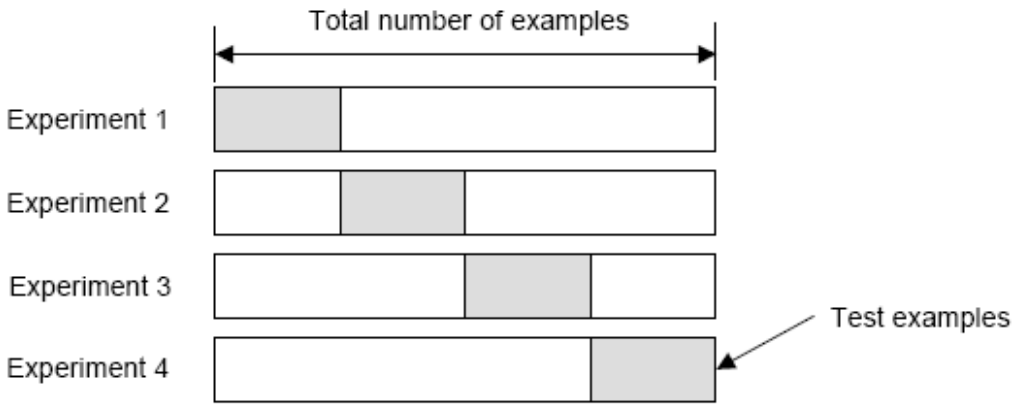
5. Αφαίρεσε το χαρακτηριστικό με το μικρότερο κατ' απόλυτη τιμή βάρος $(m \leftarrow m-1)$. Περισσότερα από ένα χαρακτηριστικά μπορεί να αφαιρεθούν σε κάθε επανάληψη.
6. Εκτίμησε την ακρίβεια του ταξινομητή στα εναπομείναντα m χαρακτηριστικά χρησιμοποιώντας έναν ταξινομητή.
7. Τέλος while loop
8. Βγάλε σαν markers genes το σύνολο των χαρακτηριστικών με το οποίο ο ταξινομητής πετυχαίνει την μεγαλύτερη ακρίβεια.

Πίνακας 1: Αλγοριθμική αναπαράσταση της RFE διαδικασίας. Σημειώνουμε ότι στο βήμα 3, μπορεί να χρησιμοποιηθεί οποιοσδήποτε ταξινομητής για να υλοποιήσει την εκτίμηση των βαρών.

1.4.K-fold cross validation

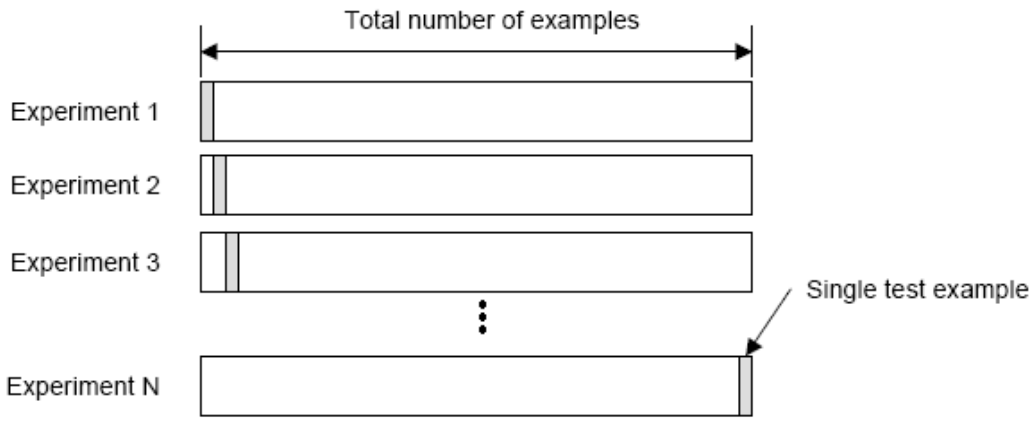
Η μέθοδος που χρησιμοποιήσαμε για να δημιουργήσουμε τα training και testing set είναι μία παραλλαγή της k-fold cross validation μεθόδου.

Στην k-fold cross validation, τα δεδομένα χωρίζονται σε k υποσύνολα και η όλη διαδικασία επαναλαμβάνεται k φορές. Κάθε φορά, ένα από τα k υποσύνολα είναι το test set και τα υπόλοιπα k-1 αποτελούν όλα μαζί το train set[11], όπως φαίνεται και παρακάτω:



Εικόνα 2: Σχηματική αναπαράσταση της k-fold cross validation μεθόδου[11]

Αν επιλέξουμε το k έτσι ώστε να είναι ίσο με το μέγεθος δείγματος, τότε προκύπτουν test sets μεγέθους ενός δείγματος και η μέθοδος αυτή λέγεται leave-one-out cross validation. Στην εικόνα παρακάτω φαίνεται σχηματικά η μέθοδος αυτή:



Εικόνα3: Σχηματική αναπαράσταση της leave-one-out cross validation μεθόδου[11]

Το πλεονέκτημα αυτής της μεθόδου είναι ότι μετράει σχετικά λίγο το πώς χωρίζονται τα δεδομένα. Κάθε δείγμα συμμετέχει μία μόνο φορά στο test set και συμμετέχει k-1 φορές ακριβώς στο train set. Επιπλέον, βλέπουμε ότι όλα τα δείγματα του συνόλου μας συμμετέχουν τελικά και στο train αλλά και στο test set, αλλά κάθε φορά επιλέγουμε το k προσπαθώντας να βρούμε την χρυσή τομή μεταξύ ακρίβειας, διασποράς και υπολογιστικού κόστους[11].

Επιλέγοντας την τιμή του k, συναντούμε κάθε φορά πλεονεκτήματα και μειονεκτήματα που πρέπει να λάβουμε υπ' όψιν ώστε να βρούμε την καταλληλότερη τιμή του. Οι επιδράσεις της επιλογής διαφόρων τιμών του k, συνοψίζονται περιληπτικά παρακάτω[11]:

Όσο αυξάνεται η τιμή του k, αυξάνεται η ακρίβεια του εκτιμητή αλλά ταυτόχρονα αυξάνεται η διασπορά του καθώς και το υπολογιστικό κόστος (περισσότερα πειράματα). Αντίθετα, όταν επιλέγουμε μικρότερες τιμές για το k, μειώνεται το υπολογιστικό κόστος και η διασπορά του εκτιμητή μας, αλλά

δεν έχουμε τόσο καλή ακρίβεια. Δηλαδή, κάνουμε είτε πολύ αισιόδοξες εκτιμήσεις είτε πολύ απαισιόδοξες εκτιμήσεις της παραμέτρου που μας ενδιαφέρει.

Πρακτικά, η επιλογή του k, εξαρτάται από το μέγεθος του συνόλου δεδομένων που έχουμε στην διάθεσή μας. Δηλαδή, για μεγάλα σύνολα δεδομένων αρκούν και μικρές τιμές του k, καθώς ακόμα και με μικρό αριθμό εκπαιδεύσεων, η ακρίβεια του εκτιμητή θα είναι ικανοποιητική. Αν όμως έχουμε μικρό αριθμό δειγμάτων, πιθανότατα πρέπει να χρησιμοποιήσουμε ακόμα και leave-one-out cross validation (όπου k είναι ίσο με το μέγεθος του δείγματος), ώστε να προκύψει ικανοποιητική ακρίβεια του εκτιμητή.

Για να γίνουν πιο κατανοητά τα παραπάνω, παραθέτουμε ένα παράδειγμα:

Έστω ότι έχουμε δύο σύνολα δεδομένων, ένα με μέγεθος 1200 δείγματα και ένα με μέγεθος 27 δειγμάτων. Για k=3, ο εκτιμητής μας στην πρώτη περίπτωση θα εκπαιδευτεί 3 φορές με train set μεγέθους 800 δειγμάτων κάθε φορά, εκπαίδευση η οποία μπορεί να οδηγήσει σε ικανοποιητική ακρίβεια. Στην δεύτερη περίπτωση, ο εκτιμητής μας θα εκπαιδευτεί 3 φορές με train set μεγέθους 18 δειγμάτων κάθε φορά που πιθανότατα να μην οδηγήσει σε ικανοποιητική ακρίβεια για τον εκτιμητή.

Αντίθετα, εφαρμόζοντας leave-one-out cross validation, στην πρώτη περίπτωση θα είχαμε 1200 εκπαιδεύσεις του εκτιμητή με train set μεγέθους 1199 δειγμάτων που προφανώς θα οδηγούσε σε ικανοποιητική ακρίβεια του τελευταίου(αλλά και σε πολύ μεγάλο υπολογιστικό κόστος και διασπορά), ενώ στην δεύτερη περίπτωση θα είχαμε 27 εκπαιδεύσεις του εκτιμητή με train set μεγέθους 26 δείγματος που πιθανότατα θα οδηγούσαν σε ικανοποιητική ακρίβεια (τουλάχιστον πιο ικανοποιητική απ’ ότι για k=3).

Μια συνήθης επιλογή για την τιμή του k είναι k=10.

Στα πειράματα που πραγματοποιήσαμε χρησιμοποιήσαμε μια παραλλαγή της μεθόδου αυτής. Η διαφορά είναι ότι κάθε φορά χωρίζαμε το train και test set επιλέγοντας κάθε φορά το 10% των δεδομένων σαν test set και το υπόλοιπο 90% σαν train set. Ο χωρισμός των δεδομένων σε train και test set δεν γίνεται με τυχαίο τρόπο όσον αφορά την κλάση από την οποία προέρχονται τα δείγματα. Κάθε φορά, συμμετέχουν δεδομένα και από τις δύο κλάσεις σε συγκεκριμένη αναλογία. Επιλέγουμε δηλαδή συγκεκριμένο αριθμό δειγμάτων από την θετική κλάση για να συμμετάσχουν στο train και test set. Έτσι τα δύο σύνολα που προκύπτουν από την διαδικασία του cross validation έχουν συγκεκριμένο αριθμό δειγμάτων από κάθε κλάση και ο αριθμός καθορίζεται από την αναλογία δειγμάτων θετικής και αρνητικής κλάσης στον αρχικό πληθυσμό. Για το λόγο αυτό, ποτέ δεν θα προκύψει ένα test set για παράδειγμα που θα έχει δείγματα μόνο από την θετική ή μόνο από την αρνητική κλάση.

Τα δείγματα της κάθε κλάσης έχουν ίσες πιθανότητες να επιλεγούν για κάθε σύνολο, οπότε υπό αυτήν την έννοια, δημιουργούμε με τυχαίο τρόπο κάθε φορά που εφαρμόζουμε την διαδικασία του cross validation, δύο σύνολα με σταθερές αναλογίες δειγμάτων από κάθε κλάση. Σε κάθε εφαρμογή της cross validation, αρχικά κατηγοριοποιείται το train set, test set και το independent test set.

Στη συνέχεια, περνάμε στην φάση της επιλογής γονιδίων από την οποία προκύπτουν από τον ταξινομητή που κάνει την επιλογή χαρακτηριστικών, τα νέα σύνολα που αποτελούνται φυσικά από τα ίδια δείγματα κάθε ένα από τα οποία έχει όμως μικρότερο αριθμό γονιδίων. Τα νέα σύνολα ανατροφοδοτούνται στον ταξινομητή που υλοποιεί την κατηγοριοποίηση ώστε να υπολογίσουμε την νέα ακρίβειά με τον μειωμένο αριθμό γονιδίων.

Η διαδικασία επαναλαμβάνεται μέχρι την τελική εξάλειψη των γονιδίων. Σε κάθε μέθοδο εφαρμόζουμε συνολικά δέκα φορές την 10-fold cross validation, πραγματοποιούμε δηλαδή 100 πειράματα.

Κεφάλαιο 2: Περιγραφή του προβλήματος

2.1. Αλγόριθμοι επιλογής γονιδίων

2.1.1.Wrapper μέθοδοι

Σε αυτήν την παράγραφο, δίνουμε μία γενική εικόνα των ταξινομητών οι οποίοι χρησιμοποιούνται για να υλοποιήσουν την επιλογή χαρακτηριστικών στις διάφορες υπό δοκιμή wrapper μεθόδους. Από τους ταξινομητές αυτούς προκύπτουν τα βάρη των χαρακτηριστικών (γονιδίων) με βάση τα οποία τα τελευταία εξαλείφονται ή όχι και δημιουργούν τα νέα train, test και independent test set που στη συνέχεια θα κατηγοριοποιηθούν με τον δεύτερο ταξινομητή ώστε να αξιολογήσουμε την ακρίβεια της μεθόδου με τον μειωμένο αριθμό γονιδίων σε κάθε επανάληψη της μεθόδου.

2.1.1.1.Support Vector Machines (SVM)

Ο SVM [1], ψάχνει το υπερεπίπεδο το οποίο ξεχωρίζει καλύτερα τις δύο κλάσεις που μας ενδιαφέρουν, την θετική(+1) και την αρνητική(-1). Το πρόβλημα αυτό, μπορεί ισοδύναμα να μοντελοποιηθεί και ως εξής:

$$\begin{aligned} &minimize \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{j=1}^{n}\xi_j^2 \\ &subject\,to\, y_j\left(\left(\boldsymbol{w}\cdot\boldsymbol{x}_j\right)+b\right)\geq 1-\xi_j,\ \xi_j\geq 0,\ i=1,\ldots n \end{aligned} \tag{1}$$

Για την λύση αυτού του προβλήματος, παρατηρούμε την παρακάτω έκφραση για το διάνυσμα κατεύθυνσης w:

$$\boldsymbol{w} = \sum_{j=1}^{n}\lambda_j y_j \boldsymbol{x}_j \tag{2}$$

Όπου,

$$1 \leq \lambda_j \leq C \tag{3}$$

Το οποίο είναι στην πραγματικότητα μία επέκταση των δειγμάτων εκπαίδευσης εκείνων με μηδενική λ_j, π.χ. των support vectors. Τα λ_js αντιστοιχούν στους πολλαπλασιαστές Lagrange, y_j αντιστοιχούν στην

ταμπέλα της κλάσης του δείγματος $\boldsymbol{x}_j$. Ένα πολλά υποσχόμενο πλεονέκτημα των SVMs είναι η δυνατότητα που έχουν να ενσωματώνουν πολυώνυμα και πυρήνες RBF και επομένως μπορούν να χρησιμοποιηθούν και σε μια μη-γραμμική μορφή.

2.1.1.2.Least Square Support Vector Machines (LSSVM)

Οι LSSVM μοντελοποιούν το πρόβλημα κατηγοριοποίησης των SVM σε μία έκδοση ελαχίστων τετραγώνων ως εξής:

$$\begin{aligned} &minimize \frac{1}{2}\|\boldsymbol{w}\|^2 + \gamma \frac{1}{2}\sum_{j=1}^{n} e_j^2 \\ &subject\,to\, y_j\left(\left(\boldsymbol{w}\cdot\boldsymbol{x}_j\right)+b\right) \geq 1-e_j,\ j\geq 0,\ i=1,\ldots n \end{aligned} \tag{4}$$

Για την λύση αυτού του προβλήματος, παρατηρούμε την παρακάτω έκφραση για το διάνυσμα κατεύθυνσης w:

$$\boldsymbol{w} = \sum_{j=1}^{n} \lambda_j y_j \boldsymbol{x}_j \tag{5}$$

Το πλεονέκτημα του LSSVM είναι ότι τελικά λύνει ένα γραμμικό πρόβλημα σε αντίθεση με τον τυπικό SVM ο οποίος λύνει ένα τετραγωνικό πρόβλημα. Αξίζει να αναφέρουμε ότι ο LSSVM έχει τις ίδιες δυνατότητες ενσωμάτωσης με τον τυπικό SVM.

2.1.1.3.The Gradient Descent Linear Neuron (LNW-GD)

Εφαρμόζοντας την μέθοδο Gradient Descent, μπορούμε να εκπαιδεύσουμε ένα γραμμικό νευρωνικό [3] χρησιμοποιώντας το παρακάτω σύνολο εξισώσεων και θέτοντας τον κατάλληλο ρυθμό μάθησης μ:

$$w_i(t+1) = w(t) + \mu e f'(u)\, g_i \tag{6}$$

$$f'(u_j) = y_j(1-y_j) \tag{7}$$

$$y_j = \frac{1}{1+e^{u_j}} = f\left(u_j\right) \tag{8}$$

$$u_j = \sum_{i=1}^{m} w_i x_{ij} \tag{9}$$

Αποδεικνύεται ότι μπορούμε να εφαρμόσουμε μια τέτοια μέθοδο στην επιλογή γονιδίων. Για περισσότερες λεπτομέρειες, ο αναγνώστης μπορεί να μελετήσει την αναφορά [4]. Μια τέτοια προσέγγιση μπορεί να επεκταθεί σε multilayer perceptron και εφαρμόζεται επίσης σε μη-γραμμικά προβλήματα.

2.1.1.4.The Ridge Regression Classifier

Ο ταξινομητής Ridge Regression, μοντελοποιεί το πρόβλημα της εύρεσης κλάσης ως εξής:

$$\begin{aligned} &minimize\, a\|\boldsymbol{w}\|^2 + \sum_{j=1}^{n} \xi_j^2 \\ &subject\, to\left(y_j - \boldsymbol{w}\cdot\boldsymbol{x}_j = \xi_j\right),\ j \geq 0,\ i = 1,\ldots n \end{aligned} \quad . \tag{10}$$

Για την λύση αυτού του προβλήματος, παρατηρούμε την παρακάτω έκφραση για το διάνυσμα κατεύθυνσης w:

$$\boldsymbol{w} = \frac{1}{2a}\sum_{j=1}^{n} \lambda_j x_j \tag{11}$$

2.1.1.5.Fisher's Linear Discriminant

Η προσέγγιση του Fisher (Fisher 1936) βασίζεται στην προβολή πολυδιάστατων δεδομένων πάνω σε μια γραμμή, με τέτοια τρόπο ώστε οι κλάσεις πάνω σε αυτή τη γραμμή να είναι καλά διαχωρισμένες. Το διάνυσμα κατεύθυνσης μιας τέτοιας γραμμής δίνεται από τον εξής τύπο:

$$\boldsymbol{w} = \frac{1}{2} k\left(\Sigma_1 + \Sigma_2\right)^{-1}\left(\mu_1 + \mu_2\right) \tag{12}$$

όπου

$$k = \frac{\sigma_{y1} + \sigma_{y1}}{\mu_{y1} - \mu_{y1}} \tag{13}$$

Από αυτό επηρεάζεται ουσιαστικά το $\|\boldsymbol{w}\|$ και σ_{y1}, σ_{y2}, μ_{y1} και μ_{y2} αντιστοιχούν στην τυπικές αποκλίσεις και τα μέσα των προβολών της αρνητικής και θετικής κλάσης, τα οποία δίνονται από τους τύπους:

$$\begin{aligned} \mu_{yi} &= \boldsymbol{w} \cdot \boldsymbol{\mu}_i \quad i = 1,2 \\ \sigma_{yi} &= \boldsymbol{w}' \Sigma_i \boldsymbol{w} \;\; i = 1,2 \end{aligned} \tag{14}$$

Όπου Σ_1, Σ_2 και μ_1, μ_2 αντιστοιχούν στους πίνακες συνδιακύμανσης και στα διανύσματα μέσων τιμών για την θετική και αρνητική κλάση
αντίστοιχα.

2.1.2. Συνδυάζοντας τις wrapper με τις μεθόδους φίλτρου

Οι συνδυαστικές μέθοδοι επιλογής χαρακτηριστικών εντάσσονται στην κατηγορία των wrapper μεθόδων εφ' όσον τα βάρη των γονιδίων δεν μένουν σταθερά αλλά υιοθετούν δυναμικά τις τιμές τους κατά την διάρκεια της διαδικασίας επιλογής. Επιπλέον, η διαδικασία εκπαίδευσής τους εμπλουτίζεται με ένα κριτήριο φιλτραρίσματος, που μας οδηγεί σε ένα υβρίδιο μεταξύ wrapper και filter προσέγγισης. Η πρώτη μέθοδος, σχετίζεται με την σωστή υιοθέτηση γραμμικών νευρωνικών βαρών, ενώ η δεύτερη σχετίζεται με την σωστή χρήση των βαρών που προκύπτουν από έναν SVM ταξινομητή.

2.1.2.1.Η προσέγγιση των γραμμικών νευρωνικών βαρών(LNW)

Αυτή η μεθοδολογία σύντηξης μιας wrapper και μιας filter προσέγγισης, επιτυγχάνεται μέσω της χρήσης ενός κατάλληλα εκπαιδευμένου νευρωνικού δικτύου ως εξής:

$$w_i(t+1) = w_i(t) + \mu \sum_{j=1}^{n} sign\left(e_j f'\left(u_j\right)\right) sign\left(g_{ij}\right) f_2\left(g_i\right) \tag{15}$$

όπου

$$f_2(g_i) = \frac{\sum_{j=1}^{n} \left| g_{ij} - c(g_i) \right|}{\sigma_+(g_i) + \sigma_-(g_i)} \tag{16}$$

και

$$c(g_i) = \frac{\left(\mu_+(g_i) + \mu_-(g_i)\right)}{2} \tag{17}$$

Μια εναλλακτική προσέγγιση η οποία δεν απαιτεί την σωστή ρύθμιση του ρυθμού μάθησης μ, δίνεται παρακάτω:

$$w_i(t+1) = w(t) + |d - y| \cdot sign\left(e \cdot f'(u)\right) \cdot sign(g_i) \cdot f_2(g_i) \tag{18}$$

Σημειώνουμε ότι μέσω του όρου $f_2(g_i)$, υλοποιούμε μια παραλλαγή του λόγου Fisher, η οποία με την διαδικασία εκπαίδευσης ενός γραμμικού νευρωνικού δικτύου μας οδηγεί στην σύντηξη μιας wrapper και μιας filter προσέγγισης. Σημειώνουμε ακόμη ότι αυτή η διαδικασία εκπαίδευσης θα μπορούσε να ενσωματωθεί καλά και με multilayer perceptron που προσομοιώνουν μη-γραμμικές συναρτήσεις.

2.1.2.2.Η προσέγγιση του Fisher's metric με Support Vectors (FSVs)

Με παρόμοιο τρόπο όπως στην παραπάνω προσέγγιση, μία παραλλαγή του λόγου Fisher, χρησιμοποιείται σωστά και καταλήγουμε στο διάνυσμα βαρών της εξίσωσης (2) ενός SVM vw ως εξής:

Έστω SVs είναι το σύνολο των support vectors και S το σύνολο των δεικτών που ορίζεται ως εξής:

$$S = \left\{ k : \boldsymbol{x}_k \in SVs \right\} \tag{19}$$

Τότε βασιζόμενη στην (2), προκύπτει ένα νέο διάνυσμα w':

$$w_i' = \sum_{j \in S} sign(\lambda_j) \cdot y_j \cdot sign(x_{ij}) \cdot \frac{|\mu_{s+}(g_i) - \mu_{s-}(g_i)|}{\sigma_{s+}(g_i) + \sigma_{s-}(g_i)} \quad (20)$$

Ακόμα, δείχνουμε ότι χρησιμοποιώντας την δυνατότητα του πυρήνα του SVM, μπορούμε να προμηθευτούμε αρκετά σύνολα από support vectors και επομένως αρκετά διανύσματα βαρών w'.

2.2. Διαστήματα εμπιστοσύνης

Στη συγκεκριμένη εργασία, η συσχέτιση της επιλεγμένης υπογραφής γονιδίων με το κλινικό αποτέλεσμα ελέγχεται μέσω της ταξινόμησης των δεδομένων. Ο τελικός ταξινομητής που λειτουργεί στα δεδομένα με το επιλεγμένο σύνολο γονιδίων μπορεί να παρουσιαστεί σαν ένας εκτιμητής του κλινικού αποτελέσματος που στην περίπτωσή μας για κάθε δείγμα που εξετάζεται, παίρνει ψηφιακή μορφή (0 για την καλή πρόγνωση και 1 για την κακή).

Το αποτέλεσμα του εκτιμητή σε νέα δεδομένα εκτιμάται μέσω του cross validation. Όμως λόγω του ότι το αποτέλεσμα διαφέρει από ένα σύνολο δεδομένων σε άλλο, χρειάζεται η παρουσίαση της δυνατότητας εκτίμησης μέσα σε ένα διάστημα εμπιστοσύνης μέσα στο οποίο κυμαίνεται η ακρίβεια του εκτιμητή σε τυχαία άγνωστα δεδομένα.

Η ακρίβεια του εκτιμητή καθώς και το διάστημα εμπιστοσύνης της μπορούν να μοντελοποιηθούν στην δική μας εργασία ως εξής:

Έστω A είναι ένας πίνακας που αποτελείται από S στήλες, όπου S είναι ο αριθμός των δειγμάτων(ασθενών) οι οποίοι μπορεί να συμμετέχουν ή όχι στο test set. Ο πίνακας αυτός έχει επίσης R γραμμές, όπου η κάθε γραμμή αντιστοιχεί σε ένα από τα 100 τρεξίματα της μεθόδου. Ορίζουμε το κάθε στοιχείο S_{ij} ως εξής:

- S_{ij} =1, αν το δείγμα συμμετέχει στο test set και έχει κατηγοριοποιηθεί σωστά
- S_{ij} =0, αν το δείγμα συμμετέχει στο test set και δεν έχει κατηγοριοποιηθεί σωστά
- S_{ij} =NULL, αλλιώς.

Αν θέλουμε λοιπόν υπολογίσουμε για το τρέξιμο i την μέση ακρίβεια της εκτίμησης θα έχουμε:

$$A_i = \frac{1}{R}\sum_{j=1}^{R} S_{ij}$$

Επιπλέον, μπορούμε όπως θα δούμε παρακάτω, να υπολογίσουμε το διάστημα εμπιστοσύνης μέσα στο οποίο κινείται η ακρίβεια αυτή, θεωρώντας ότι ακολουθεί διωνυμική κατανομή σαν άθροισμα δοκιμών Bernouli (το κάθε δείγμα μπορεί να θεωρηθεί σαν μια δοκιμή Bernouli).

Κάθε ένα από αυτά τα διαστήματα εμπιστοσύνης μπορεί να μας δώσει στοιχεία για την σταθερότητα της μεθόδου στο train set, αφού για όλους τους ασθενείς του τρεξίματος i έχουμε το ίδιο train set. Επομένως, μικρά τέτοια διαστήματα εμπιστοσύνης μας δείχνουν ότι στα συγκεκριμένα τρεξίματα η μέθοδος παρουσιάζει σταθερότητα στο train set. Αν υπολογίσουμε το μέσο διάστημα εμπιστοσύνης για όλα τα τρεξίματα, μπορούμε να βγάλουμε ένα γενικότερο συμπέρασμα για την σταθερότητα της μεθόδου σε σχέση με το train set.

Με την ίδια λογική, μπορούμε να υπολογίσουμε το διάστημα εμπιστοσύνης ανά ασθενή j και να βγάλουμε συμπεράσματα για την σταθερότητα της μεθόδου σχετικά με το test set.

2.2.1.Ορισμός του διαστήματος εμπιστοσύνης

Τα όρια εμπιστοσύνης για μία παράμετρο χρησιμοποιούνται όταν δεν μπορούμε να θεωρήσουμε την παράμετρο αυτή σαν μια σημειακή εκτίμηση. Αντί αυτού μπορούμε να πούμε με μία σχετική βεβαιότητα ότι η παράμετρος αυτή βρίσκεται μέσα σε συγκεκριμένα όρια ανοχής[6]. Αυτό εισάγει την έννοια της εκτίμησης διαστήματος για την παράμετρο αυτήν, η οποία είναι ένα διάστημα (θ1,θ2), τα ακραία σημεία του οποίου είναι συναρτήσεις $\theta1=g_1(X)$ και $\theta2=g_2(X)$ του διανύσματος παρατήρησης X.

Το αντίστοιχο τυχαίο διάνυσμα (θ1,θ2) είναι ένα διάστημα εμπιστοσύνης της X. Λέμε ότι το (θ1,θ2) είναι ένα διάστημα εμπιστοσύνης της X, αν:

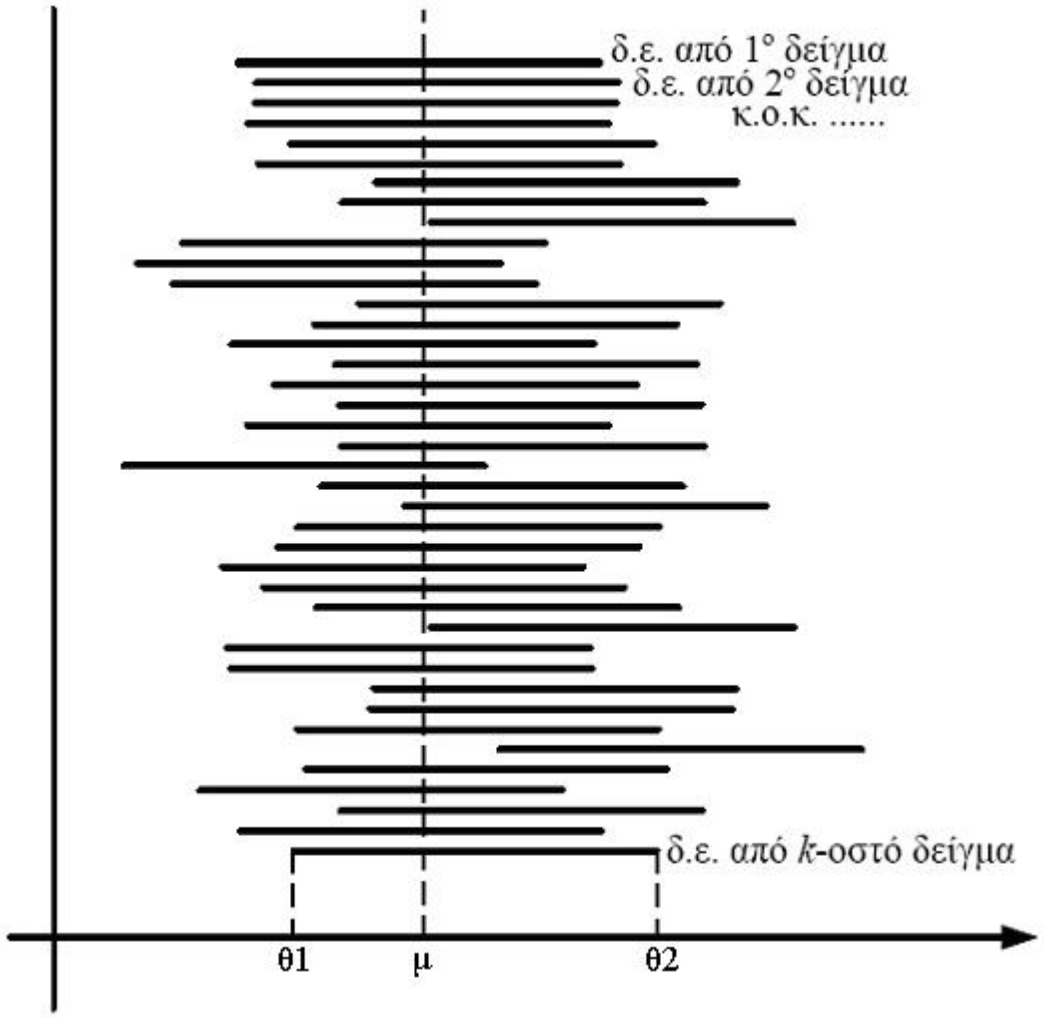
$P\{\theta1<\theta<\theta2\}= \gamma$

Η σταθερά γ είναι ο συντελεστής εμπιστοσύνης της εκτίμησης και η διαφορά α= 1-γ είναι το επίπεδο εμπιστοσύνης. Έτσι η γ είναι ένα υποκειμενικό μέτρο της εμπιστοσύνης μας, ότι η άγνωστη θ βρίσκεται στο διάστημα (θ1,θ2)[6].

Αν το γ είναι κοντά στο 1 μπορούμε να περιμένουμε σχεδόν με βεβαιότητα ότι ισχύει η εκτίμησή μας, ότι δηλαδή η άγνωστη παράμετρος βρίσκεται στην πραγματικότητα μέσα στο διάστημα (θ1,θ2). Η εκτίμησή μας είναι σωστή στο 100γ% των περιπτώσεων. Ο αντικειμενικός σκοπός της εκτίμησης διαστήματος εμπιστοσύνης είναι να καθορίσει τις συναρτήσεις $g_1(X)$ και $g_2(X)$ έτσι ώστε να ελαχιστοποιείται το μήκος θ2-θ1 του διαστήματος (θ1,θ2) για συγκεκριμένες τιμές του γ ή του α.

Έστω ότι έχουμε ένα πληθυσμό και θέλουμε να υπολογίσουμε μια παράμετρο αυτού θ. Με δειγματοληψία προκύπτει μια τιμή της παραμέτρου αυτής και ένα διάστημα εμπιστοσύνης με συντελεστή γ. Αυτό πρακτικά μας δείχνει ότι αν κάναμε δειγματοληψία πολλές φορές και υπολογίζαμε κάθε φορά το διάστημα εμπιστοσύνης της παραμέτρου θ(το οποίο θα ήταν διαφορετικό για κάθε δείγμα), το γ% των διαστημάτων εμπιστοσύνης θα περιείχε την πραγματική

τιμή της παραμέτρου. Ένα τέτοιο παράδειγμα, φαίνεται και στο παρακάτω σχήμα όπου έχουμε υπολογίσει k διαστήματα εμπιστοσύνης για την παράμετρο μ ενός πληθυσμού:

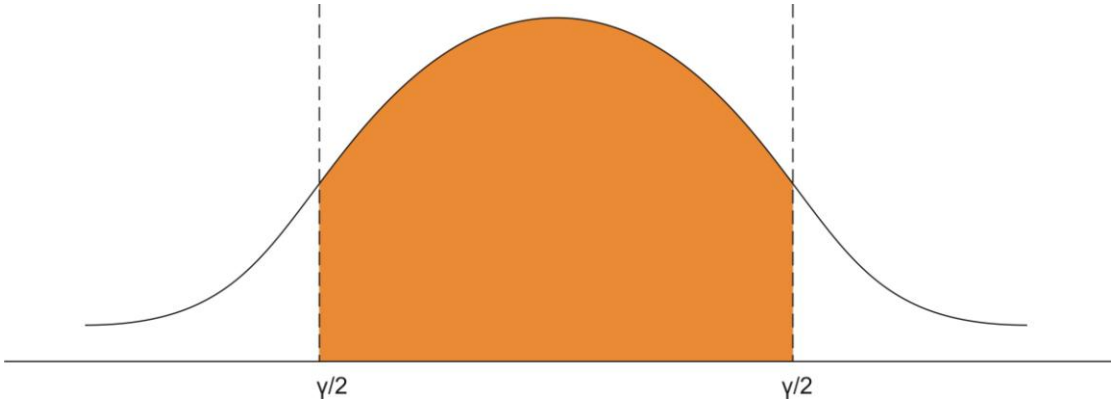


Εικόνα 4 :Διαστήματα εμπιστοσύνης για την παράμετρο μ ενός πληθυσμού

Βλέπουμε ότι στα περισσότερα διαστήματα εμπιστοσύνης περιέχεται η πραγματική τιμή της παραμέτρου μ, για την ακρίβεια περιέχεται στο γ% αυτών.

Πώς όμως ο συντελεστής γ επηρεάζει τα άκρα του διαστήματος εμπιστοσύνης; Μπορούμε να δούμε ότι όσο αυξάνουμε τον συντελεστή γ, αυξάνουμε δηλαδή τις πιθανότητες να βρίσκεται μέσα στο διάστημα εμπιστοσύνης η πραγματική τιμή της εκτιμώμενης παραμέτρου, αυξάνεται και

το μήκος του διαστήματος εμπιστοσύνης. Προφανώς για ένα υποθετικό διάστημα εμπιστοσύνης άπειρου μήκους είμαστε απολύτως σίγουροι ότι η πραγματική τιμή της παραμέτρου είναι μέσα σε αυτό (δηλαδή γ=1) .Οι πιθανότητες μειώνονται όσο μικραίνει το διάστημα εμπιστοσύνης, γεγονός που φαίνεται ξεκάθαρα στο παρακάτω σχήμα [6]:



Εικόνα 5: Σχέση μεταξύ συντελεστή εμπιστοσύνης και μήκους διαστήματος εμπιστοσύνης

Όπως μπορούμε να δούμε, για μεγαλύτερο γ, πρέπει να συμπεριλάβουμε μεγαλύτερο ποσοστό της συνάρτησης κατανομής της παραμέτρου και επομένως το διάστημα εμπιστοσύνης θα είναι μεγαλύτερο. Το αντίθετο συμβαίνει όσο μικραίνει ο συντελεστής γ.

Σαν παράδειγμα, μπορούμε να υπολογίσουμε το διάστημα εμπιστοσύνης για τον μέσο κανονικής κατανομής όταν είναι γνωστό το σ^2.

Έστω $X_1,X_2,...,X_n$ από $N(\mu,\sigma^2)$ με σ^2 γνωστό. Ζητάμε να βρούμε ένα διάστημα μέσα στο οποίο βρίσκεται το μ με πιθανότητα 1−α.

Επειδή ο δειγματικός μέσος $\bar{X}$ είναι μία αμερόληπτη εκτιμήτρια του μ θα αναζητήσουμε ένα διάστημα της μορφής [$\bar{X}$ −d, $\bar{X}$ +d]. Σύμφωνα με τα παραπάνω το d θα πρέπει να είναι τέτοιο ώστε να ισχύει:

$$P\left(\mu \in \left[\bar{X}-d, \bar{X}+d\right]\right) = P\left(\bar{X}-d \le \mu \le \bar{X}+d\right) = 1-a\,.$$

Είναι γνωστό ότι ο δειγματικός μέσος $\bar{X}$ προερχόμενος από κανονικό δείγμα είναι κανονικός (κάθε γραμμική συνάρτηση ανεξάρτητων τυχαίων μεταβλητών από την κανονική κατανομή ακολουθεί κανονική κατανομή). Και επειδή, ως γνωστό, $E\left(\bar{X}\right)=\mu, Var\left(\bar{X}\right)=\frac{\sigma}{n}$ ισχύει ότι $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ ή ισοδύναμα, $Z=\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$ ~N(0,1).Επομένως, το d θα πρέπει να είναι τέτοιο ώστε, $P\left(\bar{X}-d\le\mu\le\bar{X}+d\right)=1-a \Leftrightarrow P\left(-d\le\bar{X}-\mu\le d\right)=1-a$.Επομένως, αν η Φ^{-1} είναι η αντίστροφη συνάρτηση της Φ (η Φ ως γνήσια αύξουσα συνάρτηση είναι 1-1 και άρα αντιστρέφεται) θα ισχύει ότι:

$$P\left(\frac{-d}{\sqrt{\sigma^2/n}}\le\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\le\frac{d}{\sqrt{\sigma^2/n}}\right)=1-a \Leftrightarrow P\left(\frac{-d}{\sqrt{\sigma^2/n}}\le Z\le\frac{d}{\sqrt{\sigma^2/n}}\right)=1-a$$

$$\Leftrightarrow \Phi\left(\frac{d}{\sqrt{\sigma^2/n}}\right)-\Phi\left(-\frac{d}{\sqrt{\sigma^2/n}}\right)=1-a \Leftrightarrow \Phi\left(\frac{d}{\sqrt{\sigma^2/n}}\right)-1+\Phi\left(\frac{d}{\sqrt{\sigma^2/n}}\right)=1-a$$

$$\Leftrightarrow \Phi\left(\frac{d}{\sqrt{\sigma^2/n}}\right)=1-a/2 .$$

Επομένως ένα διάστημα εμπιστοσύνης για το μ συντελεστού 1−α θα είναι το $\left[\bar{X}-d,\bar{X}+d\right]=\left[\bar{X}-\sqrt{\frac{\sigma^2}{n}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)},\bar{X}+\sqrt{\frac{\sigma^2}{n}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}\right]$ όπου, Φ είναι η συνάρτηση κανονικής κατανομής. Με βάση τα όσα θα επεξηγηθούν παρακάτω για το Z_α, προκύπτει το διάστημα εμπιστοσύνης[7]:

$$\left[\bar{X}-\frac{\sigma}{\sqrt{n}}Z_{a/2},\bar{X}+\frac{\sigma}{\sqrt{n}}Z_{a/2}\right]$$

2.2.2.Θεωρητική εύρεση διαστημάτων εμπιστοσύνης

Παρακάτω περιγράφουμε μια γενική μεθοδολογία κατασκευής ενός διαστήματος εμπιστοσύνης συντελεστού γ=1-α για μια παράμετρο θ οποιουδήποτε πληθυσμού[7]:

Έστω $X_1,X_2,...,X_n$ ένα τυχαίο διάνυσμα από την F(x,θ).

- Αρχικά, βρίσκουμε μία στατιστική συνάρτηση[1] $T(X_1,X_2,...,X_n)$ της οποίας η κατανομή να εξαρτάται από το θ. Συνήθως ως T εκλέγουμε μία εκτιμήτρια[2] του θ.
- Στη συνέχεια, κατασκευάζουμε συνάρτηση Y =h(T, g(θ)) η κατανομή της οποίας να μην εξαρτάται από το θ.
- Υπολογίζουμε δύο σταθερές θ1, θ2 έτσι ώστε να ισχύει P(θ1 ≤ Y ≤ θ2) = 1−α.
- Εφόσον έχουν βρεθεί τα θ1, θ2, λύνουμε τη σχέση θ1 ≤ Y = h(T, g(θ)) ≤ θ2 ως προς g(θ). Έτσι, προκύπτει μία ανισότητα της μορφής:

  $L = L(X_1,X_2,...,X_n) \le g(\theta) \le U(X_1,X_2,...,X_n) = U$

Το παραπάνω ενδεχόμενο θα έχει και αυτό πιθανότητα 1−α και επομένως το διάστημα (L, U) θα είναι ένα διάστημα εμπιστοσύνης για το g(θ) συντελεστού 1−α.Τα θ1,θ2 συνήθως επιλέγονται έτσι ώστε P(Y > θ2) = P(Y < θ1) = α/2. Δηλαδή το θ2 είναι το άνω α/2-σημείο της κατανομής της Y, ενώ το θ1 είναι το άνω 1− α/2-σημείο της ίδιας κατανομής.

---

[1] Υπενθυμίζουμε ότι στατιστική (ή δειγματική) συνάρτηση θα λέγεται κάθε συνάρτηση T(X) = T(X1,X2, ...,Xn) των τυχαίων μεταβλητών του δείγματος X1,X2,...,Xn που δεν εξαρτάται από τις προς εκτίμηση παραμέτρους.

Προφανώς, κάθε στατιστική συνάρτηση είναι και αυτή μία τυχαία μεταβλητή. Για παράδειγμα, γνωστές στατιστικές συναρτήσεις είναι:

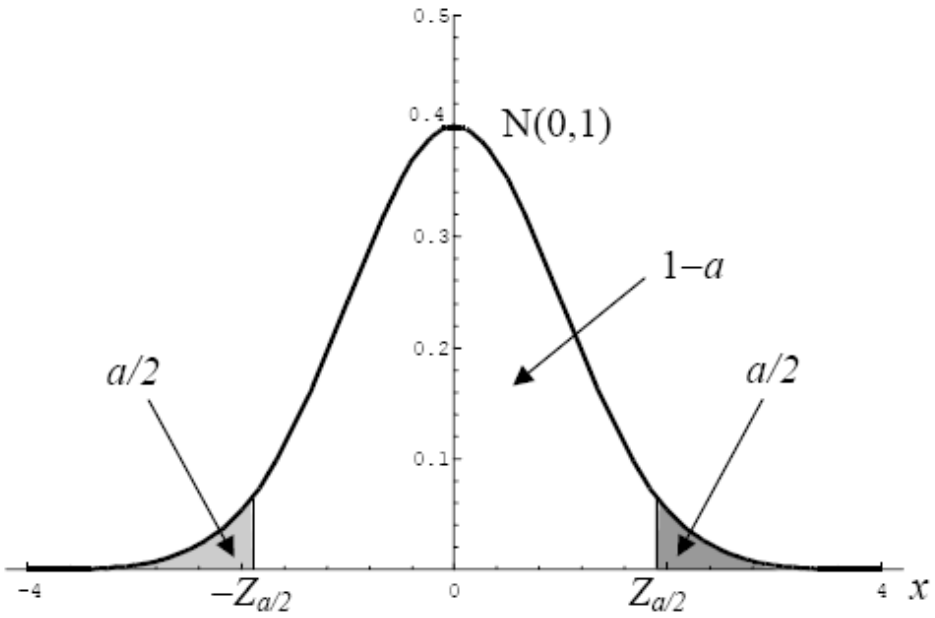
$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Xi$ (δειγματικός μέσος)

$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$ (δειγματικές ροπές τάξεως r)

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ (δειγματική διασπορά)

$R = \max\{X_1, X_2, ...X_n\} - \min\{X_1, X_2, ...X_n\} = X_{(n)} - X_{(1)}$ (δειγματικό εύρος)

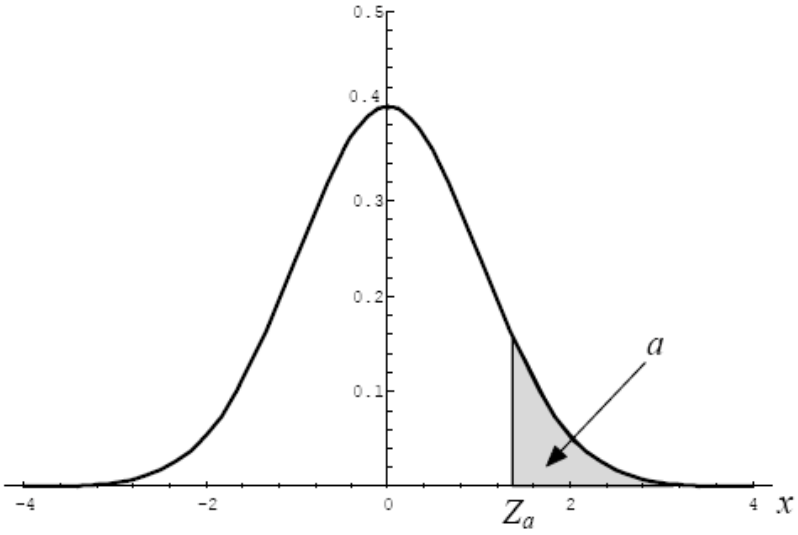
[2] Υπενθυμίζουμε ακόμα ότι εκτιμήτρια συνάρτηση μίας παραμέτρου θ θα καλείται μία στατιστική συνάρτηση T(X1,X2,...,Xn) η οποία χρησιμοποιείται για την εκτίμηση της θ.



Εικόνα 6:Γραφική αναπαράσταση του άνω α/2-σημείου και του άνω 1-α/2-σημείου της κανονικής κατανομής με μέση τιμή 0 και διασπορά 1[7].

Το άνω α-σημείο της κανονικής κατανομής συμβολίζεται συνήθως με $\mathbf{Z_\alpha = \Phi^{-1}(1-\alpha)}$. Για να δούμε σχηματικά ποιο είναι το άνω α-σημείο μιας κατανομής παίρνουμε το γράφημα της συνάρτησης πυκνότητας πιθανότητας αυτής της κατανομής. Έστω ότι έχουμε τυπική κανονική κατανομή.

Το άνω α-σημείο Za θα βρίσκεται στον άξονα των x έτσι ώστε το εμβαδόν κάτω από τη συνάρτηση πυκνότητας πιθανότητας από το Za έως το άπειρο να είναι ίσο με α:



Εικόνα 7: Γραφική αναπαράσταση του άνω α-σημείου της κανονικής κατανομής που συμβολίζεται με Za [7].

Είτε από το παραπάνω σχήμα, είτε από τη γνωστή σχέση Φ(x)=1−Φ(−x) αποδεικνύεται εύκολα ότι Z1-a = −Za. Πράγματι, αν X ~ N(0,1),τότε:

$$P\left(X > -Z_\alpha\right) = 1 - \Phi\left(-Z_\alpha\right) = \Phi\left(Z_\alpha\right) = \Phi\left(\Phi^{-1}\left(1-\alpha\right)\right) = 1-\alpha \ ,$$

και επομένως το $-Z_\alpha$ είναι το 1−α-σημείο της τυπικής κανονικής.

2.2.3. Εμπειρική κατανομή και διαστήματα εμπιστοσύνης

‘Εστω ότι λαμβάνουμε ένα πραγματικό τυχαίο δείγμα $X_1,X_2,\ldots X_n$ από ένα πληθυσμό με κατανομή F, και επιθυμούμε, με βάση το δείγμα αυτό να εξάγουμε κάποια συμπεράσματα σχετικά με μια παράμετρο θ της κατανομής F. Αν προσπαθήσουμε να εφαρμόσουμε την παραπάνω μεθοδολογία κατασκευής διαστημάτων εμπιστοσύνης για την παράμετρο αυτήν, προκύπτουν κάποια βασικά ερωτήματα όπως[8]:

- Ποια στατιστική συνάρτηση T θα χρησιμοποιήσουμε;
- Ποια είναι τα χαρακτηριστικά της T(όπως η κατανομή της, η διασπορά της, κτλ);

- Πως μπορούμε τελικά να κατασκευάσουμε ένα διάστημα εμπιστοσύνης για την θ χρησιμοποιώντας την T;

Οι απαντήσεις στα παραπάνω ερωτήματα βασίζονται στις υποθέσεις που κάνουμε για το εκάστοτε μοντέλο. Μπορούμε να διακρίνουμε δύο μεγάλες κατηγορίες μοντέλων, τα παραμετρικά και τα μη-παραμετρικά μοντέλα. Στα παραμετρικά μοντέλα, η κατανομή F,θεωρείται γνωστή, εκτός από κάποιες παραμέτρους της (προφανώς άγνωστη είναι και η παράμετρος θ), ενώ στα μη παραμετρικά μοντέλα, η κατανομή F θεωρείται εντελώς άγνωστη.

Σε πολλά παραμετρικά μοντέλα τα παραπάνω ερωτήματα μπορούν να απαντηθούν με ευκολία. Για παράδειγμα, αν $X_1, X_2, ..., X_n \sim N(\mu,\sigma^2)$ και θ = μ, τότε το θ εκτιμάται από την στατιστική συνάρτηση T = $\overline{X}$, η οποία ακολουθεί κατανομή $N(\mu,\sigma^2/n)$. Ακόμη και αν είναι άγνωστη η διασπορά της στατιστικής συνάρτησης T, μπορεί να εκτιμηθεί από το S^2/n, όπου S^2 είναι η δειγματική διασπορά. Επίσης, είναι εύκολο να κατασκευάσουμε διάστημα εμπιστοσύνης για το θ είτε το σ είναι γνωστό είτε είναι άγνωστο. Σε αρκετά όμως παραμετρικά μοντέλα, δεν μπορούμε εύκολα να προσδιορίσουμε μία κατάλληλη στατιστική συνάρτηση T και στη συνέχεια να βρούμε ή να εκτιμήσουμε τα χαρακτηριστικά της.

Μια λύση στο παραπάνω πρόβλημα για την εύρεση της κατάλληλης στατιστικής συνάρτησης T, προκειμένου να εκτιμήσουμε την παράμετρο θ της κατανομής F χωρίς να κάνουμε καμία υπόθεση για την μορφή της F, δίνει η εμπειρική κατανομή.

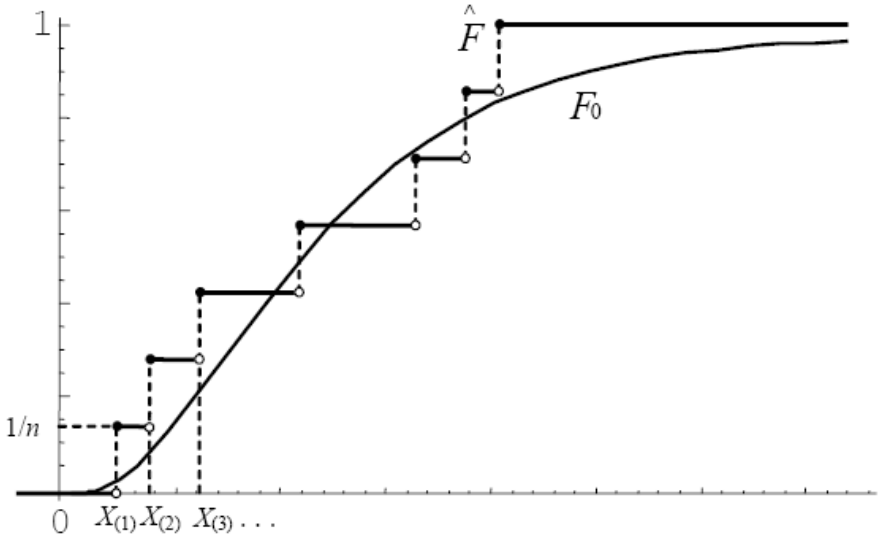
Το πρώτο βήμα είναι να περιγράψουμε την εξάρτηση της παραμέτρου θ από την F. Θα γράφουμε ότι $\theta = \theta_F$ για να υποδηλώσουμε την εξάρτηση αυτή. Για παράδειγμα, αν θ είναι η μέση τιμή, ή η διασπορά, ή το άνω α-σημείο της κατανομής F, τότε

$$\theta_F = \int_{-\infty}^{\infty} x dF(x), \text{ ή } \theta_F = \int_{-\infty}^{\infty} x^2 dF(x) - \left(\int_{-\infty}^{\infty} x dF(x) \right)^2, \text{ ή } \theta_F = F^{-1}(1-a)$$

αντίστοιχα. Εναλλακτικά, μπορούμε να περιγράψουμε την εξάρτηση της θ από την F, εκφράζοντας την θ με τη βοήθεια μιας τυχαίας μεταβλητής X η οποία ακολουθεί την κατανομή F. Για παράδειγμα, αν θ είναι και πάλι η μέση τιμή, ή η διασπορά, ή το άνω α-σημείο της κατανομής F, τότε η εξάρτηση αυτή από το δείγμα τιμών X_i, περιγράφεται γράφοντας $\theta_F = E(X)$, ή $\theta_F = V(X)$, ή θ_F: $Pr (X > \theta_F) = \alpha$ αντίστοιχα, όπου X ακολουθεί την κατανομή F.

Μπορούμε να δείξουμε ότι η εμπειρική συνάρτηση κατανομής $\hat{F}$ είναι μια συνεπής, αμερόληπτη εκτιμήτρια της συνάρτησης κατανομής F. Η εμπειρική συνάρτηση κατανομής (δηλαδή το αθροιστικό ιστόγραμμα) που προκύπτει από το τυχαίο διάνυσμα $X_1, X_2, \ldots X_n$ είναι η συνάρτηση:

$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) = \frac{\#\{X_i \le x\}}{n}, x \in R$, όπου $I(X_i \le x) = 1$ ή 0, ανάλογα με το αν $X_i \le x$ ή όχι. Επομένως, έχοντας στην διάθεσή μας τα δείγματα X_i, μπορούμε να κατασκευάσουμε την εμπειρική αθροιστική συνάρτηση κατανομής $\hat{F}$ με βάση τον παραπάνω κανόνα. Η συνάρτηση αυτή μας δείχνει ουσιαστικά για κάθε τιμή του x, τον αριθμό των δειγμάτων X_i που έχουν την τιμή αυτή ή μικρότερη αυτής και στο παρακάτω σχήμα δίνεται μια ενδεικτική μορφή της για ένα δείγμα μεγέθους 1000 τιμών:



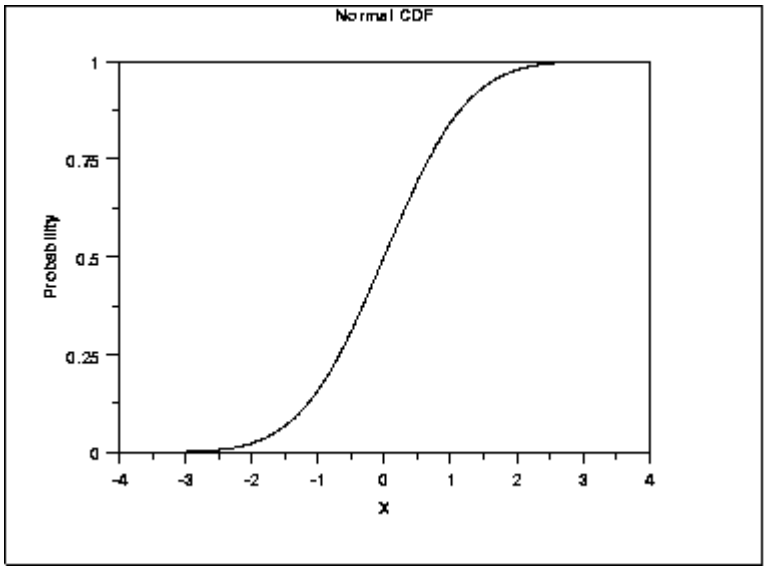
Εικόνα 8: Ενδεικτική εμπειρική συνάρτηση $\hat{F}$ (αθροιστικό ιστόγραμμα) σε σχέση με την αντίστοιχη αθροιστική συνάρτηση κατανομής F_0

Από τον νόμο των μεγάλων αριθμών γνωρίζουμε ότι για μεγάλο μέγεθος δείγματος (δηλαδή για $n \to \infty$), η πραγματική τιμή μιας παραμέτρου τείνει προς την αναμενόμενη θεωρητική. Με δεδομένο το ότι ο όρος $\frac{1}{n}\sum_{i=1}^{n} I\left(X_i \le x\right)$ είναι ένας μέσος όρος, μπορούμε να περιμένουμε ότι για μεγάλες τιμές του n θα προσεγγίζει τον αναμενόμενο θεωρητικό μέσο όρο $E\left(I\left(X_i \le x\right)\right)$. Δηλαδή:

$$\hat{F}\left(x\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \le x\right) \to n \to \infty \to E\left(I\left(X_i \le x\right)\right) = \Pr\left(I\left(X_i \le x\right) = 1\right) = \Pr\left(X_i \le x\right) = F\left(x\right)$$

με πιθανότητα 1 για κάθε x. Τα παραπάνω αποδεικνύονται και από το θεώρημα Glivenko-Cantelli το οποίο εκφράζει την ασυμπτωτική συμπεριφορά της εμπειρικής κατανομής όσο αυξάνεται ο αριθμός των παρατηρήσεων [12].
Επιπλέον, μπορούμε να εξηγήσουμε με έναν περισσότερο διαισθητικό τρόπο την παραπάνω σύγκλιση. Γνωρίζουμε ότι η αθροιστική συνάρτηση κατανομής εκφράζει την πιθανότητα τα δείγματα X_i να έχουν τιμή μικρότερη ή ίση από την τιμή x. Όσο το x αυξάνεται, αυξάνεται και αριθμός των δειγμάτων X_i τα οποία έχουν τιμή μικρότερη ή ίση με το x, επομένως αυξάνεται και η πιθανότητα πραγματοποίησης του ενδεχομένου αυτού, μέχρι κάποια τιμή του x, για την οποία η πιθανότητα είναι ίση με 1, δηλαδή, όλα τα δείγματα έχουν μικρότερη ή ίση τιμή με το x.

Παρακάτω παρατίθεται μια ενδεικτική μορφή αθροιστικής συνάρτησης κατανομής, η οποία προσεγγίζεται από την αθροιστική εμπειρική κατανομή της προηγούμενης εικόνας:



Εικόνα 9: Ενδεικτική αθροιστική συνάρτηση κατανομής

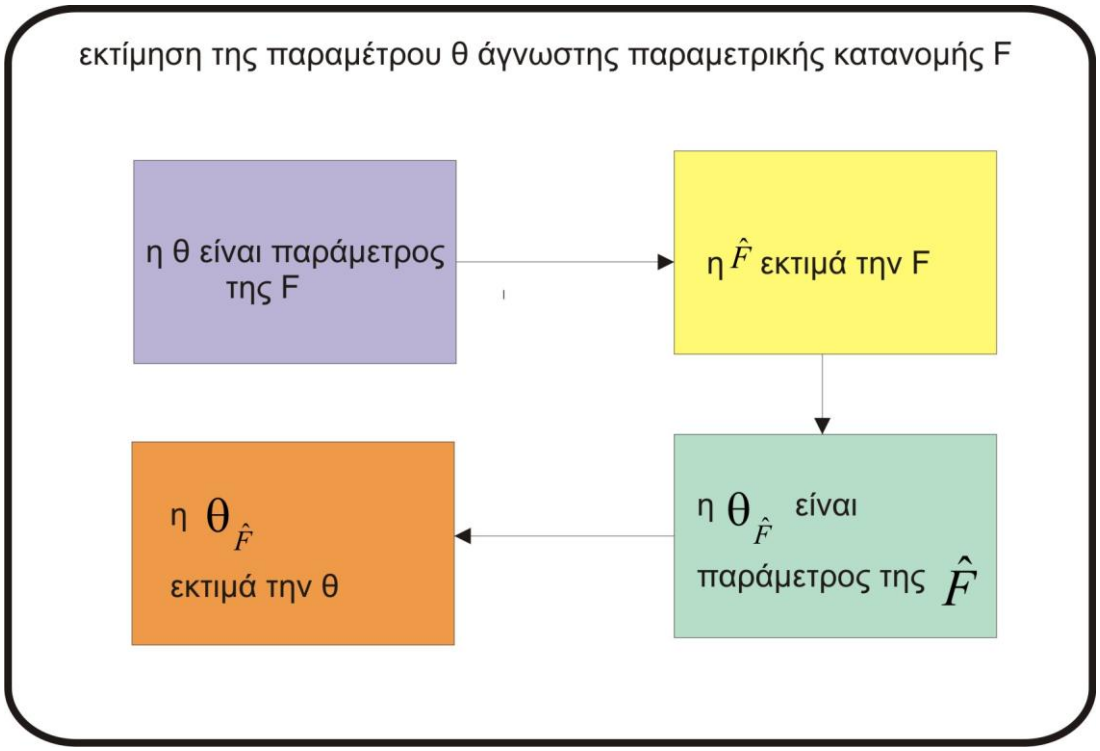
Αυτό που μπορούμε να παρατηρήσουμε σε σχέση με την συνάρτηση εμπειρικής κατανομής είναι ότι στην τελευταία, έχουμε κάποιο περιορισμένο αριθμό δειγμάτων X_i και επομένως μπορούμε να βρούμε ένα x, από το οποίο όλα τα X_i που έχουμε στην διάθεσή μας έχουν μικρότερη ή ίση με αυτό τιμή, χωρίς όμως αυτό να σημαίνει ότι το x αυτό έχει την ίδια τιμή με το αντίστοιχο της κατανομής. Όσο περισσότερα δείγματα έχουμε όμως στην διάθεσή μας, τόσο αυξάνονται οι πιθανότητες οι τιμές των x μεταξύ της εμπειρικής κατανομής και της κατανομής να συμπίπτουν. Αν είχαμε στην διάθεσή μας όλο τον πληθυσμό και όχι ένα δείγμα του, η εμπειρική κατανομή θα συνέπιπτε με την κατανομή του.

Η παραπάνω εκτίμηση της F είναι μη-παραμετρική διότι δεν βασίζεται σε καμία υπόθεση για την μορφή της F. Η κατανομή που έχει σαν συνάρτηση κατανομής την $\hat{F}$ που προέρχεται από δείγμα $x_1, x_2, ...x_n$ κατανέμει πιθανότητα 1/n σε κάθε ένα από τα δειγματικά σημεία $x_1, x_2, ...x_n$. Αυτό πρακτικά σημαίνει ότι το ενδεχόμενο να επιλέξουμε ένα από αυτά είναι ισοπίθανο με το ενδεχόμενο επιλογής κάποιου άλλου. Δηλαδή, θεωρώντας ότι επιλέγουμε ένα δείγμα X^* από την εμπειρική κατανομή,

ισχύει: $\Pr\left(X^* = x_i\right) = \hat{F}\left(x_i\right) - \hat{F}\left(x_{i-1}\right) = \frac{1}{n}, i = 1,2,...n$, όπου x_i είναι η πραγματική τιμή του μετά την εκτέλεση του πειράματος.

Καταλήγουμε επομένως στο ότι η συνάρτηση κατανομής F, μπορεί να εκτιμηθεί χωρίς να κάνουμε καμία υπόθεση για την μορφή της, από την εμπειρική κατανομή της $\hat{F}$, που προκύπτει από το δείγμα $x_1, x_2, ...x_n$, όπως προαναφέραμε. Μπορούμε λοιπόν να χρησιμοποιήσουμε την εμπειρική κατανομή της F, για να εκτιμήσουμε και την παράμετρο θ. Θα χρησιμοποιήσουμε δηλαδή ως εκτίμηση της θ_F την $\hat{\theta}_F$. Για παράδειγμα, αν η θ_F είναι η μέση της κατανομής, τότε ως εκτιμήτρια μπορούμε να θεωρήσουμε την $\hat{\theta}_F = \int_{-\infty}^{\infty} x d\hat{F}(x) = E\left(X^*\right)$, όπου X^* είναι μια τυχαία μεταβλητή που ακολουθεί την κατανομή $\hat{F}$.

Με τον τρόπο αυτό μπορούμε να προτείνουμε μία εκτιμήτρια για οποιαδήποτε παράμετρο θ_F μιας (άγνωστης) κατανομής F χωρίς να κάνουμε καμία υπόθεση για την μορφή της F. Αν π.χ. η θ_F μπορεί να γραφεί στη μορφή θ_F = E(g(X)) με X να ακολουθεί την κατανομή της F, τότε λαμβάνουμε ως εκτίμηση της θ την $\theta_{\hat{F}} = E\left(g\left(X^*\right)\right), X^* \sim \hat{F}$.Λόγω του ότι $\hat{F} \to$ F (όταν $n \to \infty$), η παραπάνω θα είναι συνεπής εκτιμήτρια του θ. Τα βασικότερα σημεία του παραπάνω συλλογισμού συνοψίζονται στο διάγραμμα που ακολουθεί:



Εικόνα 10: Διάγραμμα εκτίμησης παραμέτρου θ άγνωστης κατανομής

Έστω τώρα μια κατάλληλη στατιστική συνάρτηση εκτίμησης T(X). Για να προσεγγίσουμε την τιμή της εκτιμήτριας από το δείγμα πρέπει να προσδιορίσουμε τα χαρακτηριστικά της, μέσω την κατανομής F_T. Επειδή όμως δεν έχουμε κάνει κάποια υπόθεση για την F, θα πρέπει με κάποιο τρόπο να εκτιμήσουμε και αυτά τα χαρακτηριστικά ή γενικότερα την κατανομή F_T της T από το δείγμα.

Η συνάρτηση κατανομής της T = T(X_1, X_2,.., X_n) εξαρτάται από την F (την κατανομή των X_i), η οποία όπως είδαμε μπορεί να εκτιμηθεί από την $\hat{F}$. Η βασική ιδέα της μεθόδου bootstrap είναι να εκτιμήσουμε την κατανομή F_T της T χρησιμοποιώντας αντί της (άγνωστης) F, την $\hat{F}$. Συγκεκριμένα, εκτιμούμε την κατανομή της $T = T(X_1, X_2, ...X_n)$, όπου X_i ακολουθεί την κατανομή της F, από την κατανομή της τυχαίας μεταβλητής $T^* = T(X_1^*, X_2^*, ...X_n^*)$, όπου κάθε X_i^* ακολουθεί την κατανομή $\hat{F}$.

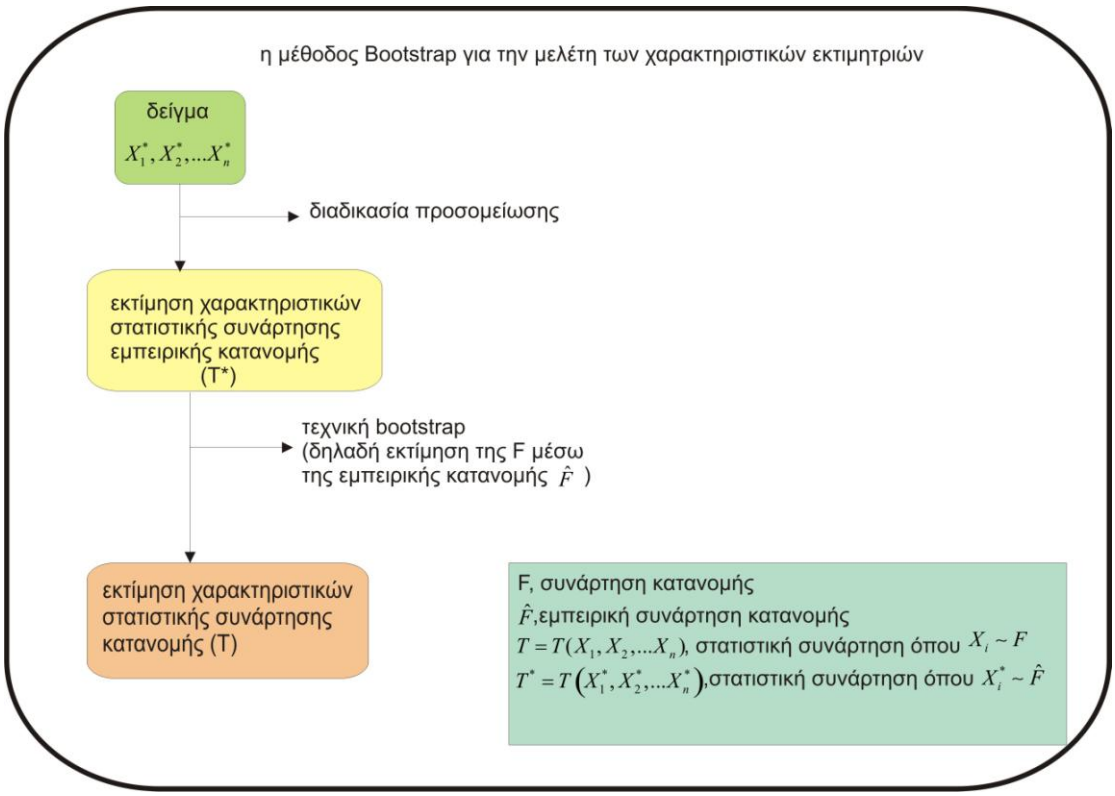
Επομένως, όλα τα ζητούμενα χαρακτηριστικά της T μπορούν να εκτιμηθούν από τα αντίστοιχα χαρακτηριστικά της T*. Με την λογική αυτή, η μέση τιμή για παράδειγμα μιας συνάρτησης της T, μπορεί να προκύψει από τον εξής τύπο:

$$\hat{E}\left(g\left(T\right)\right)=E\left(g\left(T^{*}\right)\right)=E\left(g\left(T\left(X_{1}^{*},X_{2}^{*},...X_{n}^{*}\right)\right)\right)=\frac{1}{n^{n}}\sum_{i_1}^{n}...\sum_{i_n}^{n}g\left(T\left(x_{i_1},x_{i_2},...x_{i_n}\right)\right).$$

Πρόκειται για ένα πολλαπλό άθροισμα το οποίο αποτελείται από n^n όρους. Επειδή ενέχει μεγάλη πολυπλοκότητα, έχει προταθεί μια εναλλακτική λύση η οποία βασίζεται στην προσομοίωση. Μπορούμε μέσω αυτής να υπολογίσουμε ή έστω να προσεγγίσουμε την μέση τιμή $E\left(g\left(T^{*}\right)\right)$ όπως περιγράφεται παρακάτω.

Αρχικά παράγουμε n τυχαίους αριθμούς $X_{1}^{*},X_{2}^{*},...X_{n}^{*}$ από την συνάρτηση κατανομής $\hat{F}$. Η παραγωγή ενός τυχαίου αριθμού από την $\hat{F}$ είναι εύκολη, γιατί ή $\hat{F}$ κατανέμει πιθανότητα 1/n σε κάθε ένα από τα σημεία $x_{1},x_{2},...x_{n}$.

Ουσιαστικά, επιλέγουμε τυχαία n αριθμούς $X_{1}^{*},X_{2}^{*},...X_{n}^{*}$ από το πραγματικό δείγμα $x_{1},x_{2},...x_{n}$ με επανάθεση. Από αυτούς τους n αριθμούς υπολογίζουμε την $T^{*}=T\left(X_{1}^{*},X_{2}^{*},...X_{n}^{*}\right)$. Στην συνέχεια, επαναλαμβάνουμε το ίδιο k φορές και υπολογίζουμε διαδοχικά τα $T_{1}^{*},T_{2}^{*},...T_{n}^{*}$. Μια (Monte Carlo) εκτίμηση λοιπόν της $E\left(g\left(T^{*}\right)\right)$, είναι η : $\hat{E}\left(g\left(T^{*}\right)\right)=\frac{1}{k}\sum_{i=1}^{k}g\left(T_{i}^{*}\right)$. Αυτό που συνήθως κάνουμε είναι να χρησιμοποιούμε την τεχνική της προσομοίωσης από την οποία προκύπτει μια εκτίμηση της εμπειρικής κατανομής η οποία με την σειρά της μπορεί μέσω της bootstrap μεθόδου να χρησιμοποιηθεί για να εκτιμήσει την άγνωστη κατανομή. Η εκτίμηση μέσω προσομοίωσης βελτιώνεται όσο αυξάνουμε τον αριθμό των επαναλήψεων, ενώ η μέθοδος bootstrap εξαρτάται αποκλειστικά από το μέγεθος του δείγματος και δεν επιδέχεται περαιτέρω βελτίωση. Η παραπάνω διαδικασία συνοψίζεται στην εικόνα που ακολουθεί:



Εικόνα 11: Διάγραμμα απεικόνισης της μεθόδου Bootstrap για την μελέτη χαρακτηριστικών εκτιμητριών

Έστω λοιπόν ότι έχουμε βρει την εκτιμήτρια T και θέλουμε να την χρησιμοποιήσουμε για να κατασκευάσουμε το διάστημα εμπιστοσύνης για την παράμετρο θ. Βασιζόμαστε και πάλι σε ένα τυχαίο δείγμα $X_1, X_2, ...X_n$ που ακολουθεί την άγνωστη κατανομή F και στην στατιστική συνάρτηση $T = T\left(X_1, X_2, ...X_n\right)$ που υπολογίσαμε προηγουμένως. Υποθέτουμε ότι η εκτιμήτρια που χρησιμοποιούμε δεν ακολουθεί κανονική κατανομή. Σε αυτή την περίπτωση μπορούμε και πάλι να αξιοποιήσουμε την ιδέα του bootstrap και να εκτιμήσουμε μέσω της μεθόδου αυτής και τα ποσοστημόρια της T.

Συγκεκριμένα, για να κατασκευάσουμε διάστημα εμπιστοσύνης για το θ θα πρέπει να εκτιμήσουμε τα ποσοστημόρια της τυχαίας μεταβλητής T−θ. Ειδικότερα, πρέπει να εκτιμήσουμε τα σημεία $c_{a/2}$, $c_{1-a/2}$ για τα οποία ισχύουν:

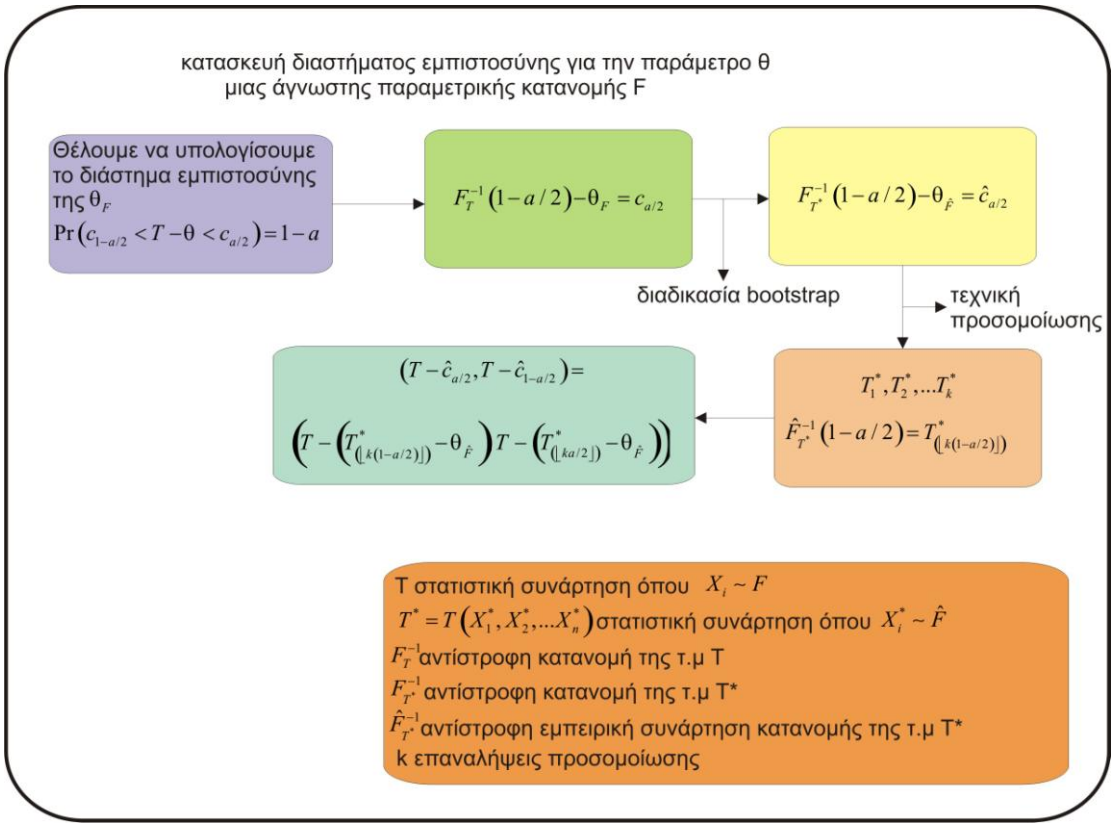
$\Pr\left(T-\theta > c_{a/2}\right) = a/2, \Pr\left(T-\theta > c_{1-a/2}\right) = 1-a/2$, διότι

τότε $\Pr\left(c_{1-a/2} \le T-\theta \le c_{a/2}\right) = 1-a$ και το διάστημα $\left(T-c_{a/2}, T-c_{1-a/2}\right)$, είναι το

διάστημα εμπιστοσύνης συντελεστού εμπιστοσύνης 1-α για την θ. Για το $c_{a/2}$ ισχύει: $\Pr\left(T-\theta_F > c_{a/2}\right) = a/2 \Leftrightarrow F_T\left(c_{a/2}+\theta_F\right) = 1-a/2 \Leftrightarrow c_{a/2} = F_T^{-1}\left(1-a/2\right)-\theta_F$ και επομένως η bootstrap εκτίμησή του (αντικαθιστούμε την F με την $\hat{F}$ και την T με την $\hat{T}$) θα είναι η: $\hat{c}_{a/2} = F_{T^*}^{-1}\left(1-a/2\right)-\theta_{\hat{F}}$, όπου $T^* = T\left(X_1^*, X_2^*, ...X_n^*\right)$ και $X_1^*, X_2^*, ...X_n^*$ είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την εμπειρική συνάρτηση κατανομής $\hat{F}$, που προέρχεται από το αρχικό δείγμα $x_1, x_2, ...x_n$. Επειδή όμως όπως είδαμε παραπάνω είναι πρακτικά αδύνατο να υπολογιστεί η $F_{T^*}^{-1}\left(1-a/2\right)$, χρησιμοποιούμε και πάλι προσομοίωση [8].

Αν παράγουμε $T_1^*, T_2^*, ...T_k^*$ τυχαίους αριθμούς τότε η συνάρτηση κατανομής F_{T^*} προσεγγίζεται από την εμπειρική συνάρτηση κατανομής $\hat{F}_{T^*}\left(x\right) = \frac{1}{k}\sum_{i=1}^{k} I\left(T_i^* \le x\right)$. Αν διατάξουμε τους τυχαίους αριθμούς $T_1^*, T_2^*, ...T_k^*$ από τον μικρότερο προς τον μεγαλύτερο, παρατηρούμε ότι στο σημείο $x = T_{(j)}^*$ η F_{T^*} θα προσεγγίζεται από $\hat{F}_{T^*}\left(T_{(j)}^*\right) = \frac{1}{k}\sum_{i=1}^{k} I\left(T_i^* \le T_{(j)}^*\right) = \frac{j}{k}$, απ' όπου προκύπτει ότι $\hat{F}_{T^*}^{-1}\left(j/k\right) = T_{(j)}^*$. Άρα αν $j/k = 1-a/2 \Leftrightarrow j = k\left(1-a/2\right)$, μπορούμε να πάρουμε $\hat{F}_{T^*}^{-1}\left(1-a/2\right) = T_{\left(\lfloor k(1-a/2)\rfloor\right)}^*$ [8].

Συνεπώς τελικά, $\hat{c}_{a/2} = T_{\left(\lfloor k(1-a/2)\rfloor\right)}^* - \theta_{\hat{F}}, \hat{c}_{1-a/2} = T_{\left(\lfloor ka/2\rfloor\right)}^* - \theta_{\hat{F}}$ και το διάστημα $\left(T-\hat{c}_{a/2}, T-\hat{c}_{1-a/2}\right) = \left(T-\left(T_{\left(\lfloor k(1-a/2)\rfloor\right)}^* - \theta_{\hat{F}}\right), T-\left(T_{\left(\lfloor ka/2\rfloor\right)}^* - \theta_{\hat{F}}\right)\right)$ είναι ένα διάστημα εμπιστοσύνης συντελεστού εμπιστοσύνης 1−α για το θ. Αν μάλιστα έχουμε θέσει $T = \theta_{\hat{F}}$ τότε το διάστημα εμπιστοσύνης είναι ίσο με $\left(2T - T_{\left(\lfloor k(1-a/2)\rfloor\right)}^*, 2T - T_{\left(\lfloor ka/2\rfloor\right)}^*\right)$. Το παραπάνω διάστημα εμπιστοσύνης καλείται βασικό bootstrap διάστημα εμπιστοσύνης συντελεστού εμπιστοσύνης 1−α για το θ [8]. Συνοπτικά όλα τα παραπάνω παρουσιάζονται στο διάγραμμα που ακολουθεί:



Εικόνα 12: Διάγραμμα κατασκευής διαστήματος εμπιστοσύνης για την παράμετρο θ άγνωστης παραμετρικής κατανομής

Στην απλή περίπτωση που το θ είναι ο μέσος μ της κατανομής F, τότε, για μεγάλο k (επαναλήψεις προσομοίωσης), το παραπάνω διάστημα εμπιστοσύνης θα είναι $\left(2\overline{X}-\overline{X}^*_{(\lfloor k(1-a/2)\rfloor)}, 2\overline{X}-\overline{X}^*_{(\lfloor ka/2\rfloor)}\right) \approx \left(2\overline{X}-F^{-1}_{\overline{X}^*}(1-a/2), 2\overline{X}-F^{-1}_{\overline{X}^*}(a/2)\right)$ και αν το μέγεθος n του αρχικού (πραγματικού) δείγματος είναι και αυτό αρκετά μεγάλο, τότε $\sqrt{n}\left(\overline{X}-\mu\right)/\sigma$ ~N(0,1) προσεγγιστικά, και το παραπάνω διάστημα εμπιστοσύνης γίνεται περίπου ίσο με $\left(2\overline{X}-\mu-\frac{\sigma}{\sqrt{n}}z_{a/2}, 2\overline{X}-\mu+\frac{\sigma}{\sqrt{n}}z_{a/2}\right) \approx \left(\overline{X}-\frac{S}{\sqrt{n}}z_{a/2}, \overline{X}-\frac{S}{\sqrt{n}}z_{a/2}\right)$,[8] (επειδή για $n\to\infty, \overline{X}\approx\mu, S^2\approx\sigma^2$), προσεγγίζει το γνωστό διάστημα εμπιστοσύνης για το μέσο της κατανομής.

2.2.4. 'Ελεγχοι καλής προσαρμογής

Κατά την προσπάθεια να κατασκευάσουμε ένα διάστημα εμπιστοσύνης για την παράμετρο θ μιας (άγνωστης) παραμετρικής κατανομής μπορεί να χρειαστεί να χρειαστεί να βρούμε πληροφορίες σχετικά με την μορφή της κατανομής αυτής από την οποία προέρχεται το τυχαίο μας δείγμα [9]. Συνεπώς είναι πολύ χρήσιμο να έχουμε τη δυνατότητα να ελέγχουμε αν κάποια δεδομένα προέρχονται από μία συγκεκριμένη κατανομή ή όχι, προκειμένου να κάνουμε τις σωστές υποθέσεις για το δείγμα μας πριν προχωρήσουμε στην χρήση έτοιμων τύπων για την κατασκευή του διαστήματος εμπιστοσύνης.

Οι έλεγχοι αυτής της μορφής καλούνται «έλεγχοι καλής προσαρμογής» των δεδομένων σε μια συγκεκριμένη κατανομή και έχουν προταθεί αρκετοί. Θα παρουσιάσουμε μερικούς ελέγχους καλής προσαρμογής οι οποίοι χρησιμοποιούν το ιστόγραμμα των δεδομένων και στην συνέχεια θα εξηγήσουμε πως τους χρησιμοποιήσαμε για να ελέγξουμε την κατανομή των δεδομένων μας και να κάνουμε τις σωστές υποθέσεις για τον υπολογισμό των διαστημάτων εμπιστοσύνης.

Ο έλεγχος χ^2 (χι-τετράγωνο) καλής προσαρμογής

Επιθυμούμε να ελέγξουμε αν κάποιες παρατηρήσεις ενός τυχαίου διανύσματος $X_1, X_2, \ldots, X_n$ προέρχονται από μια συγκεκριμένη κατανομή με συνάρτηση κατανομής F_0. Ο Pearson, ήδη από τις αρχές του προηγούμενου αιώνα (1900), πρότεινε για το σκοπό αυτό τη χρήση μιας στατιστικής συνάρτησης η οποία, υπό την υπόθεση H_0: Xi ~ F_0, ακολουθεί (προσεγγιστικά) κατανομή χ^2 (με κάποιους βαθμούς ελευθερίας) ενώ όταν δεν ισχύει η υπόθεση, λαμβάνει «μεγάλες» τιμές[9].

Πριν δούμε ποια είναι η μορφή αυτής της στατιστικής συνάρτησης στο συγκεκριμένο πρόβλημα, αξίζει να θυμηθούμε ένα σημαντικό θεωρητικό αποτέλεσμα το οποίο αφορά την πολυωνυμική κατανομή και αποτελεί την βάση του χ^2 ελέγχου καλής προσαρμογής.

Πρόταση: Αν το τυχαίο διάνυσμα $N=(N_1,N_2,...N_k)$ ακολουθεί πολυωνυμική κατανομή με παραμέτρους n και $p_1,p_2,...p_k$ με ($\sum_{i=1}^{k} p_i=1$), τότε η στατιστική συνάρτηση $T=\sum_{i=1}^{k}\frac{(N_i-np_i)^2}{np_i}$, ακολουθεί ασυμπτωτικά ($n\to\infty$) κατανομή χ^2_{k-1} [9].

Η πολυωνυμική κατανομή είναι η από κοινού κατανομή του πλήθους των επιτυχιών 1ου-είδους, 2ου-είδους,...,k-είδους σε μία ακολουθία n ανεξάρτητων και ισόνομων δοκιμών με k δυνατά είδη επιτυχιών (πιθανότητα επιυχίας i-είδους = p_i)

Έστω τώρα $X_1,X_2,\ldots,X_n$ ένα τυχαίο δείγμα και έστω ότι επιθυμούμε να ελέγξουμε την υπόθεση H_0: Xi ~ F_0. Προκειμένου να χρησιμοποιήσουμε το αποτέλεσμα της παραπάνω πρότασης εργαζόμαστε ως εξής: διαμερίζουμε το πεδίο τιμών των X_i σε k σύνολα A_1, A_2, …, A_k (συνήθως έτσι ώστε στο κάθε σύνολο να αναμένονται τουλάχιστον 5 παρατηρήσεις). Στη συνέχεια θεωρούμε τις τυχαίες μεταβλητές: N_i = πλήθος των $X_1,X_2,\ldots,X_n$ που ανήκουν στο σύνολο A_i, i = 1,2,…,n.

Είναι προφανές ότι όταν ισχύει η υπόθεση H_0: Xi ~ F_0 τότε το τυχαίο διάνυσμα $N=(N_1,N_2,...N_k)$ ακολουθεί πολυωνυμική κατανομή με παραμέτρους n και $p_1,p_2,...p_k$ όπου $p_i=P(X_1\in A_i/H_0:X_i\sim F_0), i=1,2,...,n$

Επομένως, υπό την H_0, η στατιστική συνάρτηση $T(X)=\sum_{i=1}^{k}\frac{(N_i-np_i)^2}{np_i}$

ακολουθεί προσεγγιστικά κατανομή χ^2 με k-1 βαθμούς ελευθερίας ενώ υπό την υπόθεση H_1: X_i ~ G ≠ F_0 θα λαμβάνει «μεγάλες» τιμές. Το τελευταίο συμβαίνει διότι, το np_i είναι το αναμενόμενο πλήθος παρατηρήσεων στο A_i υπό την H_0 $\left(E(N_i)=np_i=nP(X_1\in A_i/H_0)\right)$ και επομένως όταν δεν ισχύει η H_0 κάθε Ni (παρατηρούμενη συχνότητα) θα διαφέρει αρκετά από το npi (αναμενόμενη συχνότητα υπό την H_0).

Άρα, με βάση την παραπάνω στατιστική συνάρτηση μπορούμε να κατασκευάσουμε έναν έλεγχο για την υπόθεση H_0: $X_i \sim F_0$. Συγκεκριμένα θα απορρίπτουμε την H_0 (σε επίπεδο σημαντικότητας α περίπου) όταν, με βάση τις παρατηρήσεις $x_1, x_2, \ldots, x_n$, $T(x) > c = \chi^2_{k-1}(a)$: άνω α-σημείο της χ^2_{k-1}.

Παραπάνω προφανώς θεωρήσαμε ότι τα p_i είναι γνωστά (καθορίζονται πλήρως από την κατανομή F_0). Υπάρχουν όμως περιπτώσεις όπου τα p_i δεν είναι απολύτως γνωστά, αλλά εξαρτώνται από κάποιες άγνωστες παραμέτρους, δηλαδή $p_i = p_i(\theta)$ με $\theta = (\theta_1,\theta_2,\ldots,\theta_r)$ άγνωστο. Η περίπτωση αυτή εμφανίζεται π.χ. κατά τον έλεγχο καλής προσαρμογής δεδομένων σε μία γνωστή κατανομή (π.χ. κανονική) με άγνωστες όμως παραμέτρους (π.χ. μ, σ, δηλ. $p_i = p_i(\mu,\sigma)$). Στην περίπτωση αυτή χρησιμοποιούμε την τροποποιημένη στατιστική συνάρτηση: $T'(X) = \sum_{i=1}^{k} \frac{\left(N_i - np_i\left(\hat{\theta}\right)\right)^2}{np_i\left(\hat{\theta}\right)}$, όπου $\hat{\theta}$ είναι η εκτίμηση της θ, από τα δεδομένα. Τώρα, υπό την H_0, αποδεικνύεται ότι η Τ΄ ακολουθεί ασυμπτωτικά χι-τετράγωνο κατανομή με $k - r - 1$ βαθμούς ελευθερίας, όπου r είναι το πλήθος των παραμέτρων που χρειάστηκε να εκτιμηθούν από τα δεδομένα. Επομένως τώρα, απορρίπτουμε την H_0 σε επίπεδο σημαντικότητας α (περίπου) όταν $T'(x) > \chi^2_{k-1-r}(a)$.

2.2.5.Διαστήματα εμπιστοσύνης για παραμέτρους γνωστών κατανομών

Έχοντας ήδη περιγράψει τον τρόπο υπολογισμού του μέσου μ κανονικής κατανομής με γνωστό σ^2, ακολουθούν τα διαστήματα εμπιστοσύνης μερικών ακόμα παραμέτρων γνωστών κατανομών.

2.2.5.1.Κανονική κατανομή

α. Διάστημα εμπιστοσύνης για το μέσο μ κανονικής κατανομής όταν σ^2 άγνωστο.

Έστω $X_1,X_2,...,X_n$ από $N(\mu,\sigma^2)$ με σ^2 άγνωστο. Ζητάμε να βρούμε ένα διάστημα μέσα στο οποίο βρίσκεται το μ με πιθανότητα 1−α. Εφαρμόζοντας την γενική μεθοδολογία για την εύρεση του ζητούμενου διαστήματος εμπιστοσύνης καταλήγουμε στον τύπο:

$$[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(a/2), \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(a/2)]$$

όπου, S^2 είναι η δειγματική διασπορά που αποτελεί εκτιμήτρια του σ^2.

Όπου η ποσότητα t αναφέρεται στην κατανομή του Student και συμβολίζεται με t_{n-1} (κατανομή t με n−1 βαθμούς ελευθερίας).

Τα άνω α-σημεία της κατανομής t_n είναι πινακοποιημένα για διάφορες τιμές των α και n. Για n>30 μπορούμε προσεγγιστικά να πάρουμε ότι $t_n(a) \approx Z_a$ και το παραπάνω διάστημα είναι σχεδόν ίσο με το[7]:

$$[\bar{X} - \frac{S}{\sqrt{n}} Z_{a/2}, \bar{X} + \frac{S}{\sqrt{n}} Z_{a/2}]$$

β. Διάστημα εμπιστοσύνης για τη διασπορά κανονικής κατανομής όταν μ γνωστό.

Έστω X1,X2,...,Xn από $N(\mu,\sigma^2)$ με μ γνωστό. Ζητάμε να βρούμε ένα διάστημα μέσα στο οποίο βρίσκεται το σ^2 με πιθανότητα 1−α. Εφαρμόζοντας την γενική μεθοδολογία για την εύρεση του ζητούμενου διαστήματος εμπιστοσύνης καταλήγουμε στον τύπο:

$$\left[\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\chi_n^2(a/2)}, \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\chi_n^2(1-a/2)} \right]$$

Όπου η ποσότητα χ_n^2 αναφέρεται στην κατανομή του χι-τετράγωνο με n βαθμούς ελευθερίας. Τα άνω α-σημεία της κατανομής χ_n^2 είναι πινακοποιημένα για διάφορες τιμές των α και n. Για n>100 μπορούμε προσεγγιστικά να πάρουμε ότι $\chi_n^2(\alpha)=n+\sqrt{2n}\,Z_\alpha$.

γ. Διάστημα εμπιστοσύνης για τη διασπορά κανονικής κατανομής όταν μ άγνωστο.

Έστω $X_1,X_2,...,X_n$ από $N(\mu,\sigma^2)$ με μ άγνωστο. Ζητάμε να βρούμε ένα διάστημα μέσα στο οποίο βρίσκεται το σ^2 με πιθανότητα 1−α. Εφαρμόζοντας την γενική μεθοδολογία για την εύρεση του ζητούμενου διαστήματος εμπιστοσύνης καταλήγουμε στον τύπο[7]:

$$\left[\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\chi^2_{n-1}\left(a/2\right)},\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\chi^2_{n-1}\left(1-a/2\right)}\right]$$

όπου $\bar{X}$,είναι ο δειγματικός μέσος.

δ. Διάστημα εμπιστοσύνης για τη διαφορά των μέσων δύο ανεξάρτητων κανονικών πληθυσμών

Έστω $X_1,X_2,\ldots X_n$ και $Y_1, Y_2,\ldots Y_n$ δύο ανεξάρτητα δείγματα από $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$ αντίστοιχα. Ζητάμε διάστημα εμπιστοσύνης συντελεστού 1−α για τη διαφορά των μέσων $\mu_1-\mu_2$. Διαστήματα αυτής της μορφής χρησιμοποιούνται συνήθως για τη σύγκριση των δύο μέσων.

Θα εξετάσουμε αρχικά την περίπτωση που οι διασπορές είναι γνωστές. Καταλήγουμε στον τύπο[7]:

$$\left[\left(\bar{X}-\bar{Y}\right)-\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}Z_{a/2},\left(\bar{X}-\bar{Y}\right)+\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}Z_{a/2}\right]$$

Είναι το διάστημα εμπιστοσύνης συντελεστού 1−α για τη διαφορά $\mu_1-\mu_2$ όταν τα σ_1, σ_2 είναι άγνωστα αλλά ίσα.

Κάτω άκρο: $$\left(\bar{X}-\bar{Y}\right)-\sqrt{\frac{\left(n_1+n_2\right)\left(\left(n_1-1\right)S_1^2+\left(n_2-1\right)S_2^2\right)}{n_1n_2\left(n_1+n_2-2\right)}}t_{n_1+n_2-2}\left(a/2\right)$$

Άνω άκρο: $$\left(\bar{X}-\bar{Y}\right)+\sqrt{\frac{\left(n_1+n_2\right)\left(\left(n_1-1\right)S_1^2+\left(n_2-1\right)S_2^2\right)}{n_1n_2\left(n_1+n_2-2\right)}}t_{n_1+n_2-2}\left(a/2\right)$$

ε. Διάστημα εμπιστοσύνης για το λόγο των διασπορών δύο ανεξάρτητων κανονικών πληθυσμών

Έστω $X_1,X_2,\ldots X_n$ και $Y_1, Y_2,\ldots Y_n$ δύο ανεξάρτητα δείγματα από $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$ αντίστοιχα. Ζητάμε διάστημα εμπιστοσύνης συντελεστού 1−α για το πηλίκο σ_2^2/σ_1^2. Διαστήματα αυτής της μορφής χρησιμοποιούνται συνήθως για τη σύγκριση των δύο διασπορών.

Θα εξετάσουμε αρχικά την περίπτωση που οι μέσες τιμές μ_1 και μ_2 είναι γνωστές:

$$\left[\frac{n_1\sum_{i=1}^{n_2}\left(Y_i-\mu_2\right)^2}{n_2\sum_{i=1}^{n_1}\left(X_i-\mu_1\right)^2}F_{n_1,n_2\left(1-\frac{a}{2}\right)},\frac{n_1\sum_{i=1}^{n_2}\left(Y_i-\mu_2\right)^2}{n_2\sum_{i=1}^{n_1}\left(X_i-\mu_1\right)^2}F_{n_1,n_2\left(\frac{a}{2}\right)}\right]$$

Η παραπάνω κατανομή ονομάζεται κατανομή Snedecor ή κατανομή F με n1 και n2 βαθμούς ελευθερίας η οποία γράφεται ως το πηλίκο δύο ανεξάρτητων χι-τετράγωνο κατανομών διαιρεμένων δια τους βαθμούς ελευθερίας τους. Η κατανομή Snedecor έχει μελετηθεί και έχουν πινακοποιηθεί τα άνω α-σημεία της για διάφορες τιμές του α και των βαθμών ελευθερίας n1 και n2. Ακόμα, μπορεί να αποδειχτεί ότι ισχύει για την κατανομή αυτή:

$$F_{n_1,n_2}(a)=1/F_{n_2,n_1}(1-a) \text{ και } F_{1,n}(a)=\left(t_n(a)\right)^2$$

Στην περίπτωση τώρα που τα μ_1, μ_2 είναι άγνωστα, ακολουθώντας τα ίδια βήματα με παραπάνω προκύπτει το διάστημα εμπιστοσύνης:

$$\left[\frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1\left(1-\frac{a}{2}\right)},\frac{S_2^2}{S_1^2}F_{n_1-1,n_2-1\left(\frac{a}{2}\right)}\right]$$

2.2.5.2.Διωνυμική κατανομή

α. Κατασκευή διαστήματος εμπιστοσύνης για ποσοστό p ενός πληθυσμού.

- The Wald method

Έστω ότι θέλουμε να κατασκευάσουμε διάστημα εμπιστοσύνης για το ποσοστό p ενός πληθυσμού που έχει κάποιο χαρακτηριστικό. Αν πάρουμε ένα τυχαίο διάνυσμα $X_1,X_2,\ldots X_n$ από αυτόν τον πληθυσμό και θέσουμε $X_i=1$ αν το i-άτομο του δείγματος έχει το προς εξέταση χαρακτηριστικό και $X_i=0$ διαφορετικά τότε, ως γνωστό, οι παρατηρήσεις Xi θα ακολουθούν διωνυμική κατανομή B(v=1,p). Ειδικότερα, $P(X_i=x)=p^x(1-p)^{1-x}, x=0,1$

Για μεγάλο n (>30) ισχύει: $\bar{X}-\sqrt{\frac{p(1-p)}{n}}Z_{a/2}\le p\le \bar{X}+\sqrt{\frac{p(1-p)}{n}}Z_{a/2}$

Η παραπάνω παραδοχή γίνεται πάντα αποδεκτή στην πράξη για αρκετά μεγάλα δείγματα (n≥100) ενώ για μέτρια δείγματα (30<n<100) μπορούμε αν θέλουμε να ακολουθήσουμε μία πιο συντηρητική διαδικασία και να πάρουμε διάστημα εμπιστοσύνης με επίπεδο σημαντικότητας 1−α αντί συντελεστού 1−α (δηλαδή η πιθανότητα το p να ανήκει στο διάστημα εμπιστοσύνης να είναι τουλάχιστον 1−α αντί να είναι ίση με 1−α)

Αυτό γίνεται εύκολα λαμβάνοντας ως διάστημα εμπιστοσύνης το μεγαλύτερο διάστημα: $\left[\bar{X} - \frac{Z_{a/2}}{\sqrt{4n}}, \bar{X} + \frac{Z_{a/2}}{\sqrt{4n}} \right]$

Η μέθοδος αυτή είναι πολύ απλή όσον αφορά τους υπολογισμούς της. Δυστυχώς όμως, παράγει πολύ στενά διαστήματα όταν έχουμε μικρό μέγεθος δείγματος. Η μέση κάλυψή της είναι κατά προσέγγιση 60% για confidence level 95%.Αυτό δημιουργεί μία εσφαλμένη αίσθηση ακρίβειας σε αυτούς που την χρησιμοποιούν και θα ήθελαν η ονομαστική κάλυψη να συμπίπτει και στην πράξη με την πραγματική.

- The exact or Clopper-Pearson method

$$\left[1 + \frac{n-x+1}{xF_{2x,2(n-x+1),1-a/2}} \right]^{-1} < p < \left[1 + \frac{n-x}{(x+1)F_{2(x+1),2(n-x),a/2}} \right]^{-1}$$

Η μέθοδος αυτή παρέχει πιο αξιόπιστα διαστήματα εμπιστοσύνης όταν έχουμε μικρό μέγεθος δείγματος. Στην πραγματικότητα όμως, παράγει πολύ συντηρητικά διαστήματα με κάλυψη 99% για confidence level 95%.Αυτό είναι πολύ κακό για μικρό μέγεθος δείγματος(π.χ για n < 15). Αυτό σημαίνει ότι τα διαστήματα που προκύπτουν από την μέθοδο αυτή είναι πολύ ευρεία ενώ αυτά που προκύπτουν από την εφαρμογή της μεθόδου Walt είναι πολύ στενά.

Η μέθοδος αυτή χρησιμοποιεί την αθροιστική συνάρτηση της διωνυμικής κατανομής. Το διάστημα εμπιστοσύνης μπορεί να γραφεί και ως εξής:

$$\{\theta; P[Bin(n;\theta) \le X] \ge \alpha/2\} \bigcap \{\theta; P[Bin(n;\theta) \ge X] \ge a/2\}$$

Όπου X είναι ο αριθμός των επιτυχιών που έχουμε παρατηρήσει στο δείγμα και Bin(n;θ) είναι μία τυχαία μεταβλητή που ακολουθεί διωνυμική κατανομή με n δοκιμές και πιθανότητα επιτυχίας θ.

Το διάστημα εμπιστοσύνης που προκύπτει από την μέθοδο αυτή είναι πολύ ακριβές γιατί ο υπολογισμός του βασίζεται στην διωνυμική κατανομή και όχι στην προσέγγιση που γίνεται στις άλλες μεθόδους ότι για μεγάλα n γίνεται κανονική.

- The Wilson score method

$$\left(\hat{p} + \frac{z_{a/2}^2}{2n} \pm z_{a/2} \sqrt{\left[\hat{p}(1-\hat{p}) + z_{a/2}^2 / 4n \right] / n} \right) / \left(1 + z_{a/2}^2 / n\right)$$

Η μέθοδος αυτή είναι μία βελτίωση της Wald method ,δεν είναι πολύ συντηρητική, και παρέχει κάλυψη 95% για confidence level 95%.Έχει καλές προοπτικές ακόμη και για μικρό μέγεθος δείγματος. Ωστόσο, παρουσιάζει μεγάλη υπολογιστική πολυπλοκότητα, και έχει σοβαρά προβλήματα κάλυψης για συγκεκριμένες περιπτώσεις.

β. Διάστημα εμπιστοσύνης για τη διαφορά αναλογιών δύο ανεξάρτητων πληθυσμών

Σε αυτή την παράγραφο θα αναζητήσουμε διάστημα εμπιστοσύνης συντελεστού 1−α για τη διαφορά δύο ποσοστών p1−p2 από ανεξάρτητους πληθυσμούς. Διαστήματα αυτής της μορφής χρησιμοποιούνται συνήθως για τη σύγκριση δύο ποσοστών.

Έστω λοιπόν $X_1, X_2, \ldots X_n$ και $Y_1, Y_2 \ldots Y_n$ δύο ανεξάρτητα τυχαία δείγματα από B(1, p1) και B(1, p2) αντίστοιχα. Γνωρίζουμε ότι για μεγάλα n_1 και n_2, (από Κ.Ο.Θ.) ισχύει για τα δειγματικά ποσοστά ότι:

$$\bar{X} \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \text{ και } \bar{Y} \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

καταλήγουμε στο διάστημα:

$$\left[\bar{X}-\bar{Y}-\sqrt{\frac{\bar{X}(1-\bar{X})}{n_1}+\frac{\bar{Y}(1-\bar{Y})}{n_2}}Z_{a/2}, \bar{X}-\bar{Y}+\sqrt{\frac{\bar{X}(1-\bar{X})}{n_1}+\frac{\bar{Y}(1-\bar{Y})}{n_2}}Z_{a/2}\right]$$

2.2.5.3. Εκθετική κατανομή

Έστω ότι έχουμε μία τυχαία μεταβλητή (X) που ακολουθεί εκθετική κατανομή με συνάρτηση πυκνότητας $f(x,\lambda)=(1/\lambda)e^{-x/\lambda}U(x)$, και θέλουμε να βρούμε το διάστημα εμπιστοσύνης γ (γ=2α-1) της παραμέτρου λ.

Όπως γνωρίζουμε για την εκθετική κατανομή, και η μέση τιμή και η τυπική απόκλιση είναι λ. Άρα για μεγάλο n, ο δειγματικός μέσος $\bar{x}$ της τυχαίας μεταβλητής ακολουθεί κανονική κατανομή $N(\lambda,\lambda\sqrt{n})$. Αυτό μας δίνει τελικά το διάστημα $\bar{x}(1\pm z/\sqrt{n})$, όπου z_a προκύπτει από τους πίνακες της κανονικής κατανομής.

2.2.5.4.Κατανομή poisson

Έστω ότι έχουμε μία τυχαία μεταβλητή (X) που ακολουθεί κατανομή poisson με παράμετρο λ:

$P\{x=k\}= e^{-\lambda}(\lambda^{k}/k!)$, όπου k=1,2,….

Όπως γνωρίζουμε για την εκθετική κατανομή, η μέση τιμή της είναι λ και η διασπορά της είναι λ. Άρα για μεγάλο n, ο δειγματικός μέσος $\bar{x}$ της τυχαίας μεταβλητής ακολουθεί κανονική κατανομή $N(\lambda,\lambda\sqrt{n})$. Αυτό μας δίνει τελικά ότι το διάστημα εμπιστοσύνης γ της παραμέτρου λ είναι το κατακόρυφο ευθύγραμμο τμήμα (λ_1,λ_2) όπου τα λ_1 και λ_2 είναι οι ρίζες της δευτεροβάθμιας εξίσωσης: $(\lambda - \bar{x})^2= (z_u^2/n)\lambda$

Ακόμα, μια θεμελιώδης ιδιότητα της κατανομής Poisson είναι η εξής.

Αν:

το N είναι αρκετά «μεγάλο» (ας πούμε αν N >100)
p είναι αρκετά «μικρό» (π.χ. p <1/25)
το γινόμενό τους Np να είναι «της τάξεως του 1»

Τότε: ΔΙΩΝ(N,p) $\approx$ Poisson(λ), με λ = N p

Κεφάλαιο 3: Πειραματική διαδικασία και συνεισφορά

Στην παράγραφο αυτή δίνουμε μερικά στοιχεία για τα πειράματα που κάναμε και για τα δύο σύνολα δεδομένων.

Και οι δύο σειρές περαμάτων υλοποιούνται με την παραλλαγή της 10-fold cross validation που αναφέρθηκε παραπάνω. Το 90% των δεδομένων αποτελεί κάθε φορά το σύνολο που εκπαιδεύει τον ταξινομητή ενώ το 10% αυτών είναι το σύνολο που κατηγοριοποιείται για να υπολογίσουμε την ακρίβεια του.

Η διαδικασία επαναλαμβάνεται $10(\,folds)\times 10(runs\,/\,fold)$ φορές για κάθε μέθοδο. Η πληροφορία που κρατείται είναι για 100 γονίδια και κάτω, αριθμός γονιδίων μέσα στον οποίο είναι η επανάληψη στην οποία η εκάστοτε μέθοδος παρουσιάζει την καλύτερη ακρίβεια.

3.1.Πειράματα δεδομένων καρκίνου του μαστού

Τα δεδομένα που σχετίζονται με τον καρκίνο του μαστού, περιέχουν 24.481 γονίδια για κάθε ασθενή και 78 ασθενείς, κάθε ένας από τους οποίους χαρακτηρίζεται από την κλάση στην οποία ανήκει. Από τους 78 ασθενείς, 44 ανήκουν στην αρνητική κλάση που σημαίνει ότι παραμένουν υγιείς για μια χρονική περίοδο τουλάχιστον πέντε ετών, ενώ οι υπόλοιποι 34 ανήκουν στην θετική κλάση που σημαίνει ότι έκαναν μετάσταση μέσα σε πέντε χρόνια.

Από τους 78 αυτούς ασθενείς, οι 7(περίπου 10%), αποτελούν κάθε φορά το test set του cross validation και οι υπόλοιποι 71(περίπου 90%) χρησιμοποιούνται για να εκπαιδεύσουν τον ταξινομητή. Από τους 7 ασθενείς του test set, 4 ανήκουν στην αρνητική και 3 στην θετική κλάση.

Το ανεξάρτητο σύνολο, αποτελείται από 19 ασθενείς από τους οποίους 7 ανήκουν στην αρνητική και 12 στην θετική κλάση.

Όλοι οι αλγόριθμοι εκπαιδεύονται χρησιμοποιώντας το σύνολο των 71 ασθενών, ακολουθείται η διαδικασία εξάλειψης των γονιδίων που περιγράψαμε παραπάνω. Σε κάθε στάδιο εξάλειψης, ο ταξινομητής εκπαιδεύεται χρησιμοποιώντας τα εναπομείναντα χαρακτηριστικά ενώ η ακρίβεια του, ελέγχεται με βάση το test set μέσω της διαδικασίας cross validation, χρησιμοποιώντας τα ίδια γονίδια. Τελικά, το ελάχιστο σύνολο γονιδίων που πετυχαίνει την καλύτερη ακρίβεια και κατηγοριοποιεί σωστά το ίδιο το training set, είναι το τελικό σύνολο γονιδίων δεικτών.

Όλες οι μέθοδοι που εφαρμόσαμε, καθώς και οι παράμετροι που χρησιμοποιήσαμε για το σύνολο δεδομένων που σχετίζεται με τον καρκίνο του μαστού, φαίνονται στον παρακάτω πίνακα:

Όνομα μεθόδου	**Ταξινομητής ανάθεσης βαρών**	**Ταξινομητής μέτρησης ακρίβειας**	**Παράμετροι για την επιλογή χαρακτη-ριστικών**	**Παράμετροι κατηγοριο-ποίησης**
RFE-SVM	SVM	SVM	C = 100	C = 100
RFE-LNW-GD	LNW-GD	SVM	$\mu = 10^{-2}$, Epochs = 500	C = 100
RFE-LSSVM	LSSVM	LSSVM	γ = 0.1	γ = 0.1
RFE-RR	RR	SVM	$a = 10^{-1}$	LK†, C =1
RFE-FLD	FLD	SVM		LK†, C = 100
RFE-LNW1	LNW	SVM	$\mu = 10^{-2}$ *	LK†, C = 100
RFE-LNW2	LNW	SVM	$\mu = 10^{-4}$ Epochs = 500	LK†, C = 100
RFE-FSVs-7DK	RFE-FSVs	SVM	7 degree kernel C = 100	LK‡, C = 100
Filter	-	SVM	-	LK†, C = 100

∗ Χρησιμοποιούμε 3000 epochs όσο ο αριθμό των γονιδίων που μένουν είναι μεγαλύτερος από 100,αλλιώς χρησιμοποιούμε 200 epochs και μεταβλητό βήμα μάθησης με βάση την εξισωση (18).

†Linear Kernel, ‡ 7 degree polynomial kernel

Πίνακας 2: Μέθοδοι και παράμετροι για τα δεδομένα καρκίνου του μαστού.

Σε όλες τις μεθόδους έχουμε επιλέξει τις τιμές των παραμέτρων έτσι ώστε να πετυχαίνουμε την καλύτερη απόδοση κατά την διαδικασία κατηγοριοποίησης.

3.2.Πειράματα δεδομένων λευχαιμίας

Τα δεδομένα που σχετίζονται με την λευχαιμία, περιέχουν 7.129 γονίδια για κάθε ασθενή και 38 ασθενείς, κάθε ένας από τους οποίους χαρακτηρίζεται από την κλάση στην οποία ανήκει. Από τους 38 ασθενείς, 27 ανήκουν στην αρνητική κλάση που σημαίνει ότι παραμένουν υγιείς για μια χρονική περίοδο τουλάχιστον πέντε ετών, ενώ οι υπόλοιποι 11 ανήκουν στην θετική κλάση που σημαίνει ότι έκαναν μετάσταση μέσα σε πέντε χρόνια.

Από τους 38 αυτούς ασθενείς, οι 4 (περίπου 10%), αποτελούν κάθε φορά το test set για το cross validation και οι υπόλοιποι 34 (περίπου 90%) χρησιμοποιούνται για να εκπαιδεύσουν τον ταξινομητή. Από τους 4 ασθενείς του test set, 3 ανήκουν στην αρνητική και 1 στην θετική κλάση.

Το ανεξάρτητο σύνολο, αποτελείται από 34 ασθενείς από τους οποίους 20 ανήκουν στην αρνητική και 14 στην θετική κλάση.

Όλοι οι αλγόριθμοι εκπαιδεύονται χρησιμοποιώντας το σύνολο των 34 ασθενών, ακολουθείται η διαδικασία εξάλειψης των γονιδίων που περιγράψαμε παραπάνω. Σε κάθε στάδιο εξάλειψης, ο ταξινομητής εκπαιδεύεται με χρησιμοποιώντας τα εναπομείναντα χαρακτηριστικά ενώ η ακρίβεια του, ελέγχεται με βάση το test set ,χρησιμοποιώντας τα ίδια γονίδια.

Τελικά, το ελάχιστο σύνολο γονιδίων που πετυχαίνει την καλύτερη ακρίβεια και κατηγοριοποιεί εξίσου καλά το ίδιο το training set,είναι το τελικό σύνολο γονιδίων δεικτών.

Όλες οι μέθοδοι που εφαρμόσαμε, καθώς και οι παράμετροι που χρησιμοποιήσαμε για το σύνολο δεδομένων που σχετίζεται με τον καρκίνο του μαστού, φαίνονται στον παρακάτω πίνακα:

Όνομα μεθόδου	**Ταξινομητής ανάθεσης βαρών**	**Ταξινομητής μέτρησης ακρίβειας**	**Παράμετροι για την επιλογή χαρακτη-ριστικών**	**Παράμετροι κατηγοριο-ποίησης**
RFE-SVM	SVM	SVM	C = 100	C = 100
RFE-FSVs-4DK	RFE-FSVs	SVM	4 degree kernel C = 100	LK‡, C = 100
Filter	-	SVM	-	LK†, C = 100

Linear Kernel, ‡ 4 degree polynomial kernel

Πίνακας 3: Μέθοδοι και παράμετροι για τα δεδομένα λευχαιμίας.

3.3. Συνεισφορά εργασίας

Αφού περιγράψαμε αναλυτικά την πειραματική διαδικασία μπορούμε τώρα να αναφέρουμε κάποια στοιχεία σχετικά με την διαδικασία και την επεξεργασία των δεδομένων μας, καθώς και τις υποθέσεις που έγιναν ώστε να προκύψουν τα τελευταία. Για το σύνολο δεδομένων που σχετίζεται με τον καρκίνο του μαστού, όπως προαναφέραμε υλοποιήσαμε 9 μεθόδους και βρήκαμε για κάθε μία των αριθμό των γονιδίων στον οποίο η μέθοδος πέτυχε το καλύτερο αποτέλεσμα στην κατηγοριοποίηση του ανεξάρτητου test set, ενώ ταυτόχρονα κατηγοριοποιούσε τέλεια το ίδιο το training set.

Για τον αριθμό αυτό, υπολογίσαμε για κάθε μέθοδο έναν τελικό πίνακα οι στήλες του οποίου ήταν οι ασθενείς που συμμετείχαν στο πείραμα συνολικά (δηλαδή οι ασθενείς του cross validation και του ανεξάρτητου test set), ενώ οι

γραμμές του ήταν τα 100 τρεξίματα που πραγματοποιήθηκαν για κάθε μέθοδο.

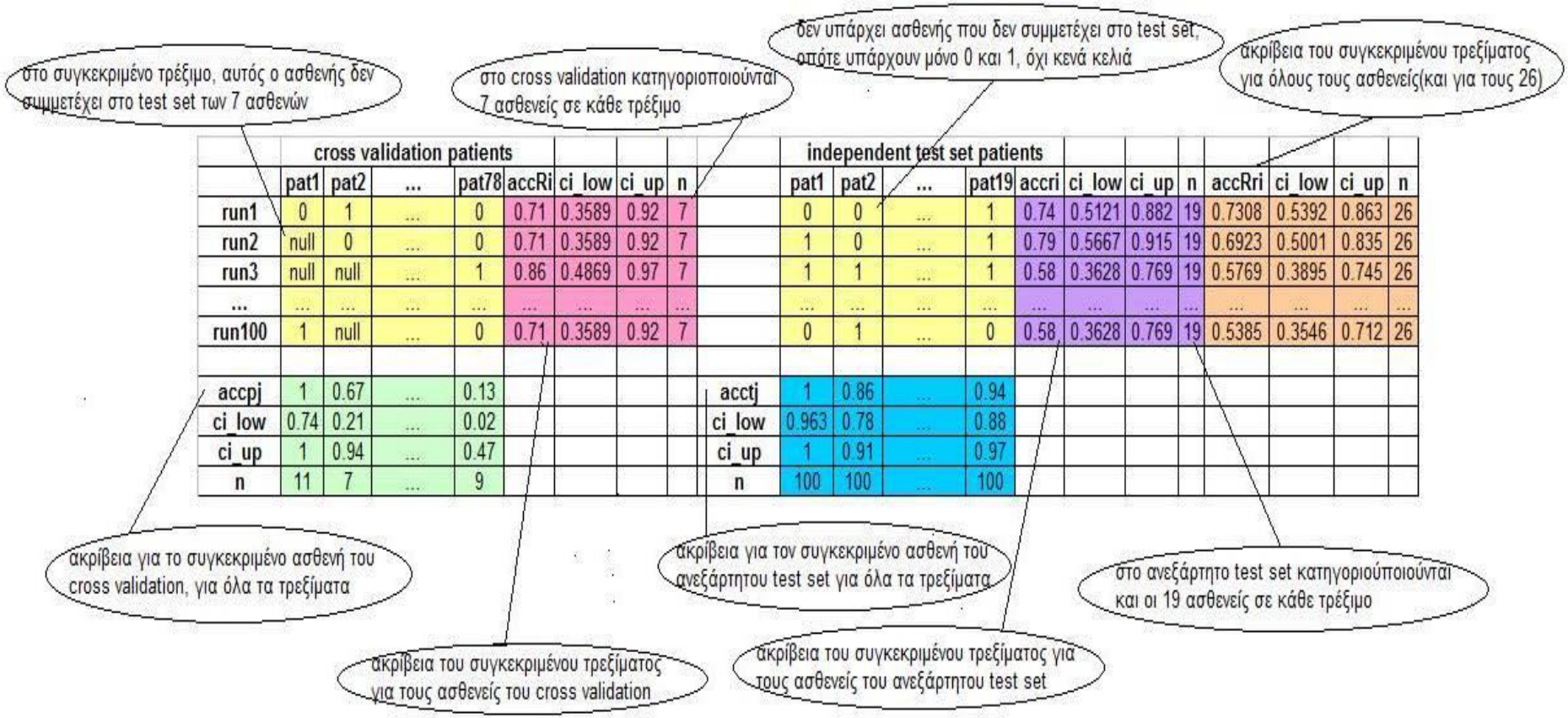
Για κάθε ασθενή του cross validation σημειώναμε 1,0 ή κενό στο αντίστοιχο τρέξιμο με την εξής λογική:

- 1,αν ο ασθενής είχε συμμετάσχει στο test set των 7 ασθενών για το συγκεκριμένο τρέξιμο και είχε κατηγοριοποιηθεί σωστά,
- 0, αν ο ασθενής είχε συμμετάσχει στο test set των 7 ασθενών για το συγκεκριμένο τρέξιμο και δεν είχε κατηγοριοποιηθεί σωστά,
- Και κενό αν ο ασθενής δεν είχε συμμετάσχει στο test set των 7 ασθενών για το συγκεκριμένο τρέξιμο

Για κάθε ασθενή του ανεξάρτητου test set σημειώναμε 1 ή 0 (εφ' όσον όλοι οι ασθενείς του ανεξάρτητου test set συμμετείχαν στο test set των 19 ασθενών για κάθε τρέξιμο δεν σημειώνουμε κενό) στο αντίστοιχο τρέξιμο με την εξής λογική:

- 1,αν ο ασθενής είχε κατηγοριοποιηθεί σωστά,
- 0, αν ο ασθενής δεν είχε κατηγοριοποιηθεί σωστά

Μια ενδεικτική μορφή ενός τέτοιου τελικού πίνακα για ένα συγκεκριμένο αριθμό γονιδίων (ο οποίος υπολογίστηκε όπως εξηγήσαμε παραπάνω) είναι η εξής:



	cross validation patients									independent test set patients												
	pat1	pat2	...	pat78	accRi	ci_low	ci_up	n		pat1	pat2	...	pat19	accri	ci_low	ci_up	n	accRri	ci_low	ci_up	n
run1	0	1	...	0	0.71	0.3589	0.92	7		0	0	...	1	0.74	0.5121	0.882	19	0.7308	0.5392	0.863	26
run2	null	0	...	0	0.71	0.3589	0.92	7		1	0	...	1	0.79	0.5667	0.915	19	0.6923	0.5001	0.835	26
run3	null	null	...	1	0.86	0.4869	0.97	7		1	1	...	1	0.58	0.3628	0.769	19	0.5769	0.3895	0.745	26
...	...	...	...	...	...	...	...	...		...	...	...	...	...	...	...	...	...	...	...	...
run100	1	null	...	0	0.71	0.3589	0.92	7		0	1	...	0	0.58	0.3628	0.769	19	0.5385	0.3546	0.712	26
accpj	1	0.67	...	0.13					acctj	1	0.86	...	0.94								
ci_low	0.74	0.21	...	0.02					ci_low	0.963	0.78	...	0.88								
ci_up	1	0.94	...	0.47					ci_up	1	0.91	...	0.97								
n	11	7	...	9					n	100	100	...	100								

Εικόνα 13: Ενδεικτικός τελικός πίνακας για τα δεδομένα καρκίνου του μαστού

Με το κίτρινο χρώμα φαίνονται τα μηδενικά, οι άσσοι και τα κενά κελιά που συμπληρώνονται με τον τρόπο που εξηγήσαμε παραπάνω, για τους ασθενείς του cross validation αλλά και του ανεξάρτητου test set.

Με το ροζ χρώμα σημειώνουμε την ακρίβεια για όλους τους ασθενείς του cross validation για το κάθε τρέξιμο της μεθόδου. Δηλαδή, το accRi είναι η ακρίβεια του i τρεξίματος σε όλους τους ασθενείς του cross validation.Επειδή έχουμε 100 τρεξίματα για κάθε μέθοδο, έχουμε 100 τιμές για το accRi. Ακόμη, υπολογίζουμε για κάθε accRi το διάστημα εμπιστοσύνης του, με συντελεστή εμπιστοσύνης 95%. Κάθε accRi είναι η αναλογία των άσσων προς το σύνολο άσσων και μηδενικών. Δηλαδή, βρίσκουμε κάθε φορά το ποσοστό των δειγμάτων που κατηγοριοποιήθηκαν σωστά, σε σχέση με το σύνολο των δειγμάτων που κατηγοριοποιήθηκαν τα οποία είναι πάντα 7 όπως έχουμε πει παραπάνω. Οι 7 ασθενείς που συμμετέχουν κάθε φορά στο test set, μπορούν να πάρουν την τιμή 1 ή 0 ανάλογα με το αν έχουν ή όχι κατηγοριοποιηθεί σωστά. Εμείς ψάχνουμε αυτήν την αναλογία μηδενικών και άσσων. Ψάχνουμε επομένως μια αναλογία σε μια σειρά πειραμάτων τύχης Bernoulli,και αυτό μας οδηγεί στο να συμπεράνουμε ότι για να υπολογίσουμε τα διαστήματα εμπιστοσύνης αυτών των μεγεθών(accRi), πρέπει να χρησιμοποιήσουμε τους τύπους της διωνυμικής κατανομής. Συγκεκριμένα τα διαστήματα εμπιστοσύνης των accRi υπολογίζονται από τον τύπο: $\left(\hat{p} + \frac{z_{a/2}^2}{2n} \pm z_{a/2} \sqrt{\left[\hat{p}(1-\hat{p}) + z_{a/2}^2 / 4n \right] / n} \right) / \left(1 + z_{a/2}^2 / n\right)$, όπου $\hat{p}$ είναι το αντίστοιχο accRi, η ποσότητα $z_{a/2}$ είναι για την κανονική κατανομή και για συντελεστή εμπιστοσύνης 95% ίση με 1,96 και n είναι το σύνολο των 7 ασθενών του test set.

Στο ίδιο συμπέρασμα οδηγούμαστε και στην προσπάθειά μας να υπολογίσουμε διαστήματα εμπιστοσύνης για την ακρίβεια ενός ασθενή του cross validation σε όλα τα τρεξίματα(accpj) που φαίνεται στο πράσινο χρώμα. Επειδή έχουμε 78 ασθενείς για το cross validation, έχουμε και 78 τιμές για το accpj.Χρησιμοποιούμε και πάλι τον ίδιο τύπο για τον υπολογισμό των

διαστημάτων εμπιστοσύνης και κρατάμε κάθε φορά το πλήθος των εμφανίσεων του συγκεκριμένου ασθενή στα test set των 100 τρεξιμάτων.

Σημειωμένα με το μωβ χρώμα, φαίνονται τα κελιά που περιέχουν την ακρίβεια για κάθε τρέξιμο της μεθόδου σε όλους του ασθενείς του test set(accri), οι οποίοι συμμετέχουν όλοι κάθε φορά στο test set. Γι' αυτό είναι πάντα 19 ο αριθμός των ασθενών που συμμετέχουν στο test set για κάθε τρέξιμο. Τα διαστήματα εμπιστοσύνης υπολογίζονται με τον παραπάνω τρόπο.

Με γαλάζιο χρώμα βλέπουμε την ακρίβεια για κάθε έναν από τους 19 ασθενείς του ανεξάρτητου test set(acctj), και τα διαστήματα εμπιστοσύνης της που υπολογίζονται με τον τρόπο που περιγράψαμε. Έχουμε 19 τιμές για το acctj όπως είναι αναμενόμενο. Ακόμα, βλέπουμε ότι η συχνότητα εμφάνισης των ασθενών στα test set των 100 τρεξιμάτων είναι πάντα 100, αφού συμμετέχουν όλοι στο test set του κάθε τρεξίματος.

Τέλος, βλέπουμε με πορτοκαλί χρώμα, την ακρίβεια του κάθε τρεξίματος για όλους τους ασθενείς και των δύο test set που δοκιμάζονται σε κάθε τρέξιμο(accRri). Βλέπουμε ότι κάθε φορά συμμετέχουν 26 ασθενείς (7 από το test set του cross validation και 19 από το ανεξάρτητο test set) και υπολογίζουμε με τον τρόπο που περιγράψαμε παραπάνω τα διαστήματα εμπιστοσύνης.

Το ίδιο κάναμε και για τα δεδομένα που σχετίζονται με τη λευχαιμία. Φτιάξαμε τέτοιους πίνακες για τις τρεις μεθόδους που εφαρμόσαμε σε αυτά τα δεδομένα και υπολογίσαμε τα ίδια μέτρα απόδοσης και τα διαστήματα εμπιστοσύνης αυτών με τον ίδιο τρόπο που θα περιγραφεί αναλυτικότερα παρακάτω. Ένας τελικός πίνακας για τα δεδομένα που σχετίζονται με την λευχαιμία φαίνεται παρακάτω:

στο cross validation συμμετέχουν 38 ασθενείς

4 ασθενείς κάθε φορά είναι στο test set

στο ανεξάρτητο test set συμμετέχουν 34 ασθενείς

	cross validation patients									independent test set patients											
	pat1	pat2	...	pat38	accRi	ci_low	ci_up	n		pat1	pat2	...	pat34	accri	ci_low	ci_up	n	accRri	ci_low	ci_up	n
run1	0	1	...	0	1	0.5101	1	4		0	0	...	1	0.79	0.632	0.897	34	0.8158	0.6658	0.908	38
run2	null	0	...	0	0.75	0.3006	0.95	4		1	0	...	1	0.82	0.6649	0.917	34	0.8158	0.6658	0.908	38
run3	null	null	...	1	1	0.5101	1	4		1	1	...	1	0.85	0.6987	0.936	34	0.8421	0.6958	0.926	38
...	...	...	...	...	...	...	...	...		...	...	...	...	...	...	...	...	...	...	...	...
run100	1	null	...	0	0.75	0.3006	0.95	4		0	1	...	0	0.79	0.632	0.897	34	0.8158	0.6658	0.908	38
accpj	1	1	...	1					acctj	1	0.92	...	1								
ci_low	0.74	0.65	...	0.72					ci_low	0.963	0.85	...	0.96								
ci_up	1	1	...	1					ci_up	1	0.96	...	1								
n	11	7	...	10					n	100	100	...	100								

Εικόνα 14: Ενδεικτικός τελικός πίνακας για τα δεδομένα της λευχαιμίας

Αφού λοιπόν δημιουργήσαμε πίνακες της παραπάνω μορφής για όλους τους αλγόριθμους, θελήσαμε να μελετήσουμε κάθε μια από τις μεταβλητές που προαναφέρθηκαν χωριστά σαν να αποτελούν δείγματα ενός ανεξάρτητου πληθυσμού και να βρούμε τις μέσες τιμές των πληθυσμών, καθώς και τα διαστήματα εμπιστοσύνης.

Για να το κάνουμε αυτό, έπρεπε να βρούμε την κατανομή αυτού του πληθυσμού. Κατ' αρχάς δεν ήταν διωνυμική κατανομή γιατί δεν είχαμε τώρα μια σειρά από πειράματα κατανομής Bernoulli. Υποθέσαμε στην συνέχεια ότι ήταν κανονική η μορφή της πράγμα που θελήσαμε να επιβεβαιώσουμε χρησιμοποιώντας την εμπειρική κατανομή. Αν και τα δεδομένα δεν ακολουθούσαν ακριβώς κανονική κατανομή, χρησιμοποιώντας την εμπειρική και γνωρίζοντας ότι σαν μέσες τιμές ενός πληθυσμού ακολουθούν περίπου κανονική κατανομή, κάναμε τελικά αυτήν την υπόθεση και υπολογίσαμε τα διαστήματα εμπιστοσύνης αφαιρώντας και προσθέτοντας από την μέση τιμή την ποσότητα που ορίζεται από την κανονική κατανομή για συντελεστή εμπιστοσύνης 95% πολλαπλασιασμένη με την τυπική απόκλιση.

Στο επόμενο κεφάλαιο παρουσιάζονται αρχικά οι μεταβλητές που αναφέραμε στην αρχή αυτού του κεφαλαίου καθώς και τα διαστήματα εμπιστοσύνης που υπολογίσαμε από τους τύπους της διωνυμικής κατανομής για όλες τις μεθόδους.

Στην συνέχεια, για τις τρεις μεθόδους που έχουν εφαρμοστεί και στα δύο σύνολα δεδομένων παρουσιάζονται διαγράμματα που μας δείχνουν την μέση τιμή κάθε μέτρου απόδοσης καθώς και το διάστημα εμπιστοσύνης αυτής(που υπολογίζεται με τύπους κανονικής κατανομής), για τα 100 τελευταίες επαναλήψεις των μεθόδων, δηλαδή για τα 100 τελευταία γονίδια που επιβιώνουν κατά την εφαρμογή των μεθόδων επιλογής χαρακτηριστικών.

Παρουσιάζονται επίσης συγκεντρωτικοί πίνακες με τα μέτρα απόδοσης για όλες τις μεθόδους και για τα δύο σύνολα δεδομένων ώστε να γίνεται πιο εύκολα η σύγκρισή τους.

Στη συνέχεια, αναλύουμε τα αποτελέσματα των πειραμάτων που πραγματοποιήσαμε και για τα δύο σύνολα δεδομένων. Βγάζουμε συμπεράσματα για τα μέτρα απόδοσης των μεθόδων, για τα διαστήματα εμπιστοσύνης τους, καθώς και για την επικάλυψη των γονιδίων, σταματώντας σε διαφορετικούς κάθε φορά αριθμούς γονιδίων τις επαναληπτικές μεθόδους.

Συγκρίνουμε τις wrapper με τις μεθόδους φίλτρου, καθώς και τις συνδυαστικές μεθόδους με κάθε μια από τις παραπάνω κατηγορίες και ελέγχουμε αν τα συμπεράσματα στα οποία οδηγηθήκαμε από την εφαρμογή των μεθόδων στα δεδομένα καρκίνου του μαστού επιβεβαιώνονται και από την εφαρμογή τους στα δεδομένα της λευχαιμίας.

Τέλος, στο κεφάλαιο 5, συνοψίζουμε τα αποτελέσματα της εργασίας μας και αναφέρουμε την μελλοντική προοπτική της δουλειάς που έγινε σε αυτήν την εργασία προτείνοντας κάποιες από τις ιδέες που θα μπορούσαν να την αναπτύξουν και να την επεκτείνουν.

Κεφάλαιο 4: Παρουσίαση των αποτελεσμάτων

4.1.Αποτελέσματα δεδομένων καρκίνου του μαστού

Για κάθε μέθοδο από τις 9 που υλοποιήσαμε υπολογίσαμε τέσσερα μεγέθη:

- Την ακρίβεια της μεθόδου ανά τρέξιμο για τους 19 ασθενείς του ανεξάρτητου test set(accri)
- Την ακρίβεια της μεθόδου ανά τρέξιμο για όλους ασθενείς του ανεξάρτητου test set και του test set(accRri)
- Την ακρίβεια της μεθόδου ανά ασθενή για τους ασθενείς του test set που εμφανίστηκαν πάνω από 10 φορές συνολικά στα test set των 7 ασθενών, στα 100 τρεξίματα της μεθόδου(accpj)
- την ακρίβεια της μεθόδου ανά ασθενή για όλους τους ασθενείς του ανεξάρτητου test set(acctj)

Για τα παραπάνω αυτά τα μεγέθη, υπολογίσαμε τα όρια εμπιστοσύνης, θεωρώντας ότι ακολουθούν διωνυμική κατανομή ως αθροίσματα δοκιμών Bernouli, με βάση τον τύπο:

$$\left(\hat{p} + \frac{z_{a/2}^2}{2n} \pm z_{a/2} \sqrt{\left[\hat{p}\left(1-\hat{p}\right) + z_{a/2}^2 / 4n \right] / n} \right) / \left(1 + z_{a/2}^2 / n\right)$$

όπου, $\hat{p}$ είναι η μέση τιμή του κάθε μέτρου απόδοσης όπως αυτή προκύπτει από τις πειραματικές μετρήσεις, Z_α είναι η ποσότητα που προκύπτει από το άνω α-σημείο της κανονικής κατανομής και n είναι το πλήθος των στοιχείων.

Για κάθε μέθοδο, έχουμε υπολογίσει τον βέλτιστο αριθμό γονιδίων ο οποίος είναι αυτός στον οποίο πετυχαίνουμε απόλυτη ακρίβεια στο training set και την καλύτερη ακρίβεια στο ανεξάρτητο test set των 19 ασθενών. Όλα τα αποτελέσματα παρουσιάζονται για αυτόν τον αριθμό ο οποίο είναι διαφορετικός για κάθε αλγόριθμο:

ΟΝΟΜΑ ΜΕΘΟΔΟΥ	ΒΕΛΤΙΣΤΟΣ ΑΡΙΘΜΟΣ ΓΟΝΙΔΙΩΝ
RFE-FSVs-7DK	73
GSM	61
SVM	32
LSSVM	45
FLD	28
LNW-GD	22
LNW1	44
LNW2	64
RR	7

Πίνακας 4: Μέθοδοι που υλοποιήθηκαν και δοκιμάστηκαν στο σύνολο δεδομένων του καρκίνου του μαστού και ο αριθμός γονιδίων στον οποίο πετυχαίνουν την καλύτερη ακρίβεια.

Παρακάτω, παρουσιάζονται για τους 9 αλγορίθμους, 5 διαγράμματα τα οποία είναι:

Α. Ακρίβεια για όλους τους ασθενείς του ανεξάρτητου test set ανά τρέξιμο (accri)

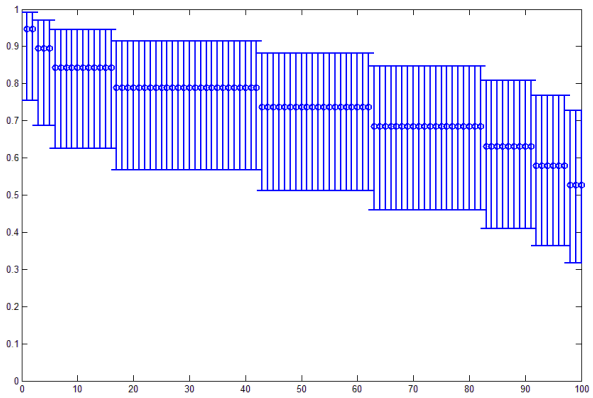
Β. Ακρίβεια για όλους τους ασθενείς του test set που προκύπτει από το cross validation και τους ασθενείς του ανεξάρτητου test set συνολικά, ανά τρέξιμο (accRri)

Γ. Ακρίβεια για όλα τα τρεξίματα, ανά ασθενή του cross validation, μόνο για όσους εμφανίζονται περισσότερες από 10 φορές σε test set(accpj με συχνότητα >10)

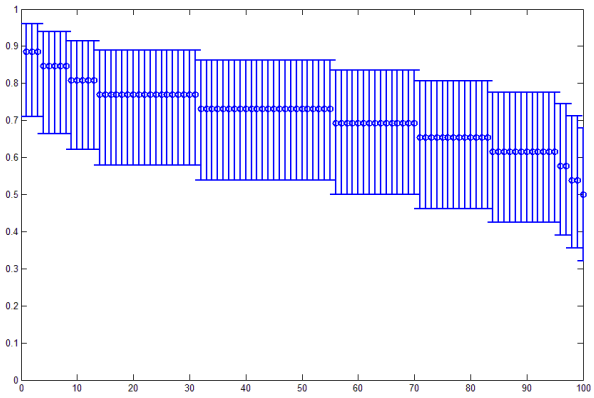
Δ. Ακρίβεια για όλα τα τρεξίματα, ανά ασθενή του ανεξάρτητου test set(acctj)

Ε. Ακρίβεια για όλα τα τρεξίματα ανά ασθενή, για όλους τους ασθενείς του ανεξάρτητου test set και όσους ασθενείς του cross validation εμφανίζονται περισσότερες από 10 φορές σε test set(acctj,accpj)

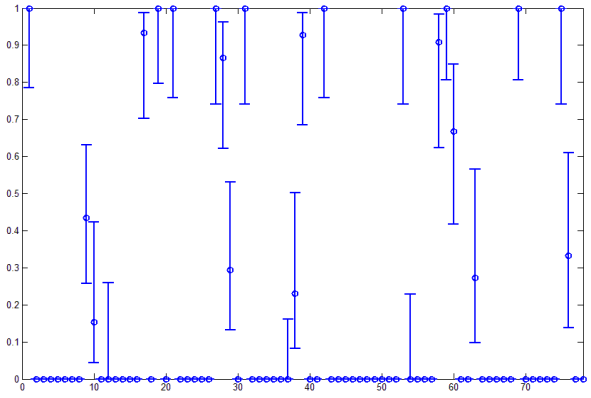
1.ΜΕΘΟΔΟΣ FSVs-7DK ΣΤΑ 73 GENES:



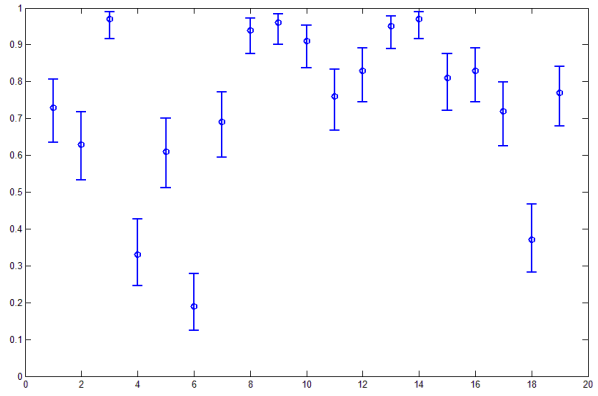
Διάγραμμα 1:1Α



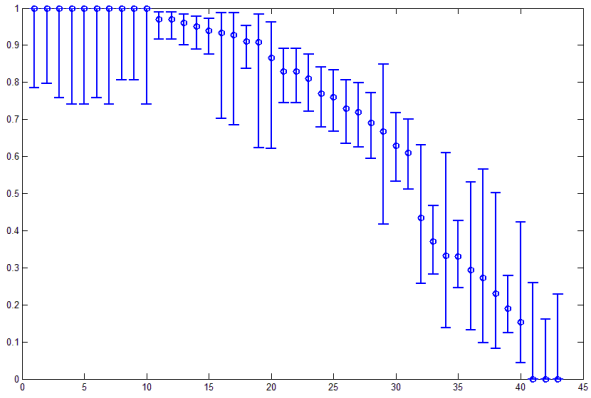
Διάγραμμα 2:1Β



Διάγραμμα 3:1Γ



Διάγραμμα 4:1Δ



Διάγραμμα 5:1Ε

Παρατηρούμε ότι η ακρίβεια για όλους τους ασθενείς του test set που προκύπτει από το cross validation και τους ασθενείς του ανεξάρτητου test set συνολικά, ανά τρέξιμο (διάγραμμα 1Β), έχει πιο μικρά διαστήματα εμπιστοσύνης από την ακρίβεια για τους ασθενείς του ανεξάρτητου test set ανά τρέξιμο (διάγραμμα 1Α).

Αυτό συμβαίνει γιατί, στο διάγραμμα 1Α έχουμε μικρότερο N_i στον υπολογισμό του διαστήματος εμπιστοσύνης καθώς πάντα έχουμε 19 δείγματα που κατηγοριοποιούνται, σε αντίθεση με το διάγραμμα 1Β που εξετάζει την κατηγοριοποίηση 19 δειγμάτων από το ανεξάρτητο test set και 7 δειγμάτων από το cross validation, σύνολο δηλαδή 26 δειγμάτων.

Επιπλέον, υπολογίζουμε την τυπική απόκλιση της κάθε ακρίβειας και βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 1Α (τυπική απόκλιση μέσης ακρίβειας 0.089) από ότι στο 1Β (τυπική απόκλιση μέσης ακρίβειας 0.077) .Παρατηρούμε ακόμα ότι στο διάγραμμα 1Β υπάρχουν 13 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

Το διάγραμμα 1Α (μέση ακρίβεια 0.735) έχει γενικά μεγαλύτερες τιμές ακρίβειας από το 1Β (μέση ακρίβεια 0.713), γιατί η συγκεκριμένη μέθοδος παρουσιάζει μεγαλύτερη ακρίβεια στην κατηγοριοποίηση του ανεξάρτητου test set από ότι στην κατηγοριοποίηση του cross validation, με αποτέλεσμα η απόδοση του cross validation να μειώνει την απόδοση της μεθόδου στο σύνολο των ασθενών (διάγραμμα 1Β) σε σχέση με αυτήν στο ανεξάρτητο test set (διάγραμμα 1Α).

Όσον αφορά τα διαγράμματα που απεικονίζουν την ακρίβεια για όλα τα τρεξίματα, ανά ασθενή του cross validation, μόνο για όσους εμφανίζονται περισσότερες από 10 φορές σε test set (διάγραμμα 1Γ) και την ακρίβεια για όλα τα τρεξίματα, ανά ασθενή του ανεξάρτητου test set (διάγραμμα 1Δ), παρατηρούμε ότι το διάγραμμα 1Γ έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 1Δ.

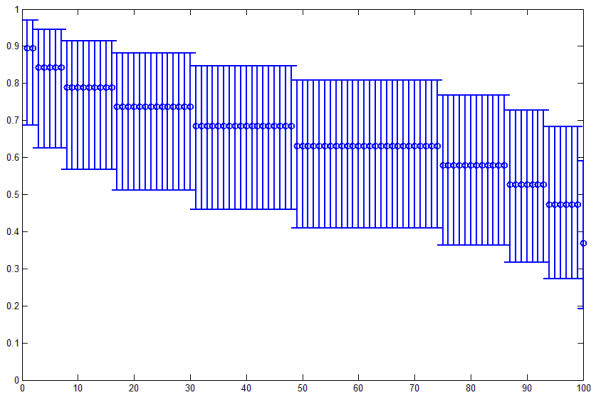
Αυτό συμβαίνει γιατί, στο διάγραμμα 1Γ έχουμε μικρότερο N_i στον υπολογισμό του διαστήματος εμπιστοσύνης καθώς το N_i είναι η συχνότητα με την οποία εμφανίζεται ένας ασθενής του cross validation στα test sets σε 100 τρεξίματα, σε αντίθεση με το διάγραμμα 1Δ στο οποίο κάθε ασθενής κατηγοριοποιείται όλες τις φορές στα 100 τρεξίματα της μεθόδου, επομένως η συχνότητα εμφάνισής του είναι πάντα 100. Το μέγεθος N_i επηρεάζει καθοριστικά το μέγεθος του διαστήματος εμπιστοσύνης ακόμα και αν στον υπολογισμό του τελευταίου για το διάγραμμα 1Γ συμμετέχουν μόνο οι ασθενείς με N_i μεγαλύτερο του 10.

Επιπλέον, υπολογίζουμε την τυπική απόκλιση της κάθε ακρίβειας και βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 1Γ (τυπική απόκλιση μέσης ακρίβειας 0.387) από ότι στο 1Δ (τυπική απόκλιση μέσης ακρίβειας 0.221).

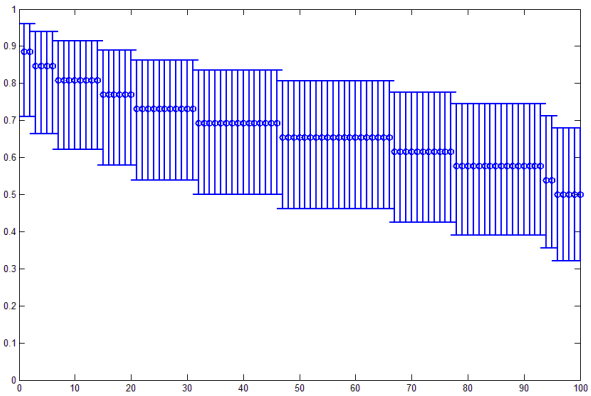
Το διάγραμμα 1Δ (μέση ακρίβεια 0.735) έχει γενικά μεγαλύτερες τιμές ακρίβειας από το 1Γ (μέση ακρίβεια 0.668), γιατί η συγκεκριμένη μέθοδος παρουσιάζει μεγαλύτερη ακρίβεια στην κατηγοριοποίηση του ανεξάρτητου test set (διάγραμμα 1Δ) από ότι στην κατηγοριοποίηση του cross validation (διάγραμμα 1Γ).

Τέλος, παρατηρούμε ότι στο διάγραμμα 1Ε υπάρχουν 23 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

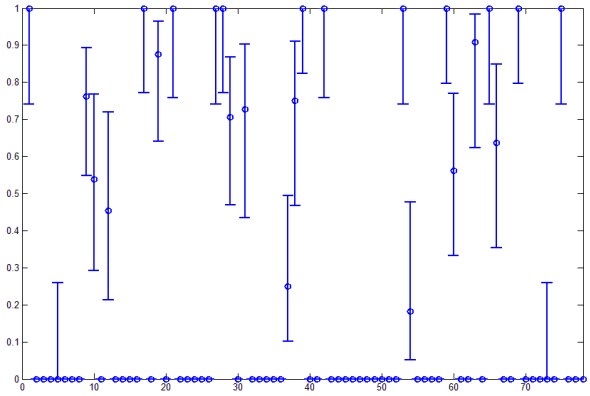
2. ΜΕΘΟΔΟΣ GSM ΣΤΑ 61 GENES:



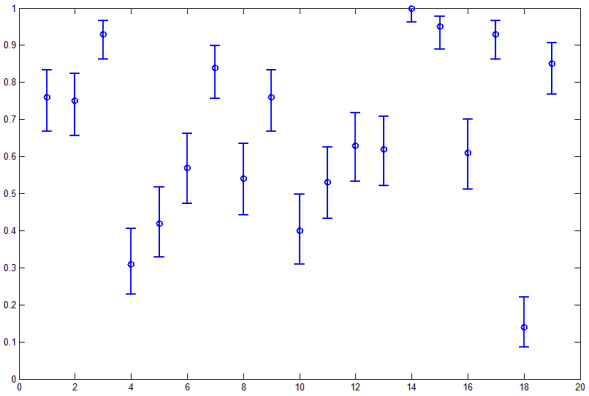
Διάγραμμα 6:2Α



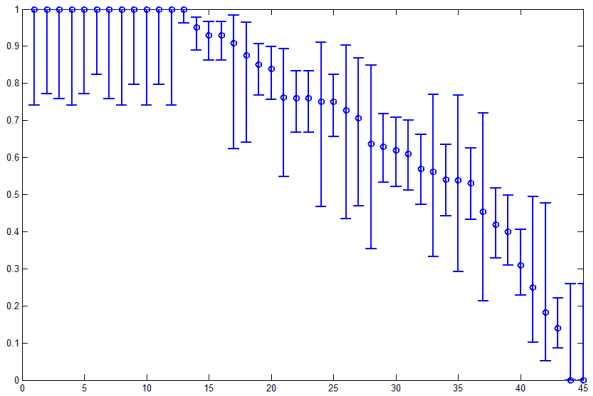
Διάγραμμα 7:2Β



Διάγραμμα 8:2Γ



Διάγραμμα 9:2Δ



Διάγραμμα 10:2Ε

Η συμπεριφορά της μεθόδου φίλτρου GSM, διαφοροποιείται από αυτήν της RFE-FSVs-7DK, όσον αφορά το διάγραμμα 2Β και 2Α, καθώς το διάγραμμα 2Α (μέση ακρίβεια 0.660) έχει γενικά μικρότερες τιμές ακρίβειας από το 2Β (μέση ακρίβεια 0.673), γιατί η συγκεκριμένη μέθοδος παρουσιάζει μικρότερη ακρίβεια στην κατηγοριοποίηση του ανεξάρτητου test set από ότι στην κατηγοριοποίηση του cross validation, με αποτέλεσμα η απόδοση του cross validation να αυξάνει την απόδοση της μεθόδου στο σύνολο των ασθενών (διάγραμμα 2Β) σε σχέση με αυτήν στο ανεξάρτητο test set (διάγραμμα 2Α).

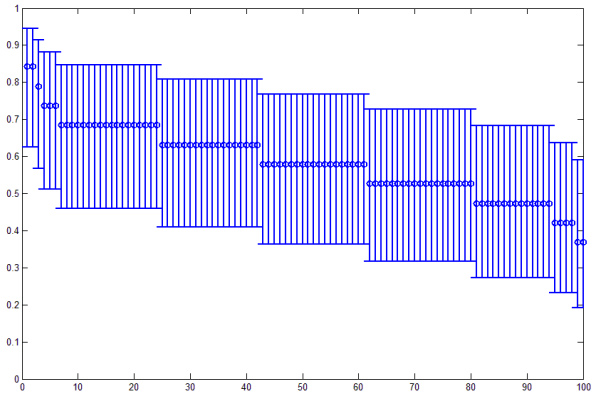
Μια ακόμα διαφοροποίηση παρατηρούμε ακόμα στα διαγράμματα 2Γ και 2Δ, όπου το διάγραμμα 2Δ (μέση ακρίβεια 0.660) έχει γενικά μικρότερες τιμές ακρίβειας από το 2Γ (μέση ακρίβεια 0.744), γιατί η συγκεκριμένη μέθοδος παρουσιάζει μικρότερη ακρίβεια στην κατηγοριοποίηση του ανεξάρτητου test set (διάγραμμα 2Δ) από ότι στην κατηγοριοποίηση του cross validation (διάγραμμα 2Γ).

Στα υπόλοιπα διαγράμματα παρατηρούμε παρόμοια συμπεριφορά των δύο μεθόδων καθώς:

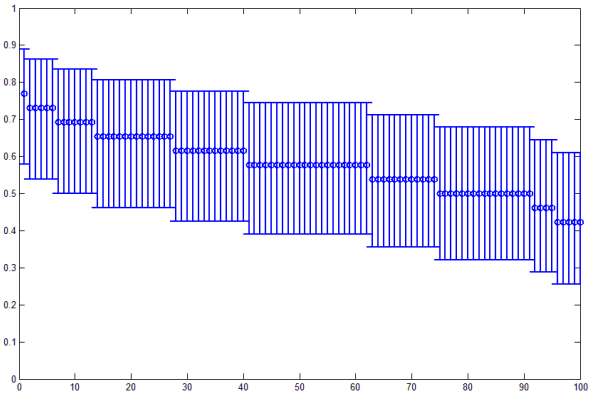
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 2Α (τυπική απόκλιση μέσης ακρίβειας 0.102) από ότι στο 2Β (τυπική απόκλιση μέσης ακρίβειας 0.091)
- το διάγραμμα 2Α έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 2Β
- το διάγραμμα 2Γ έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 2Δ
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 2Γ (τυπική απόκλιση μέσης ακρίβειας 0.320) από ότι στο 2Δ (τυπική απόκλιση μέσης ακρίβειας 0.230)

Παρατηρούμε ακόμα ότι στο διάγραμμα 2Β υπάρχουν 14 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.Τέλος, παρατηρούμε ότι στο διάγραμμα 2Ε υπάρχουν 20 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

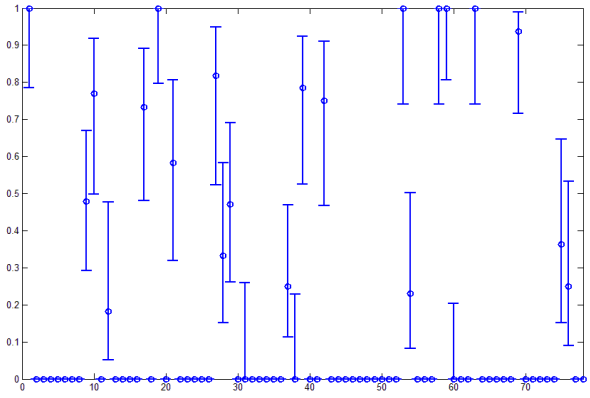
3. ΜΕΘΟΔΟΣ SVM ΣΤΑ 32 GENES:



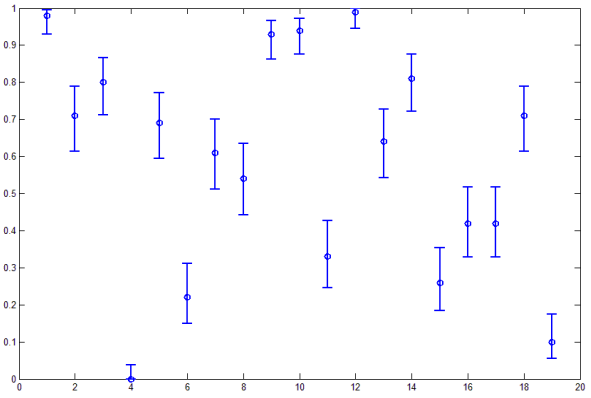
Διάγραμμα 11:3Α



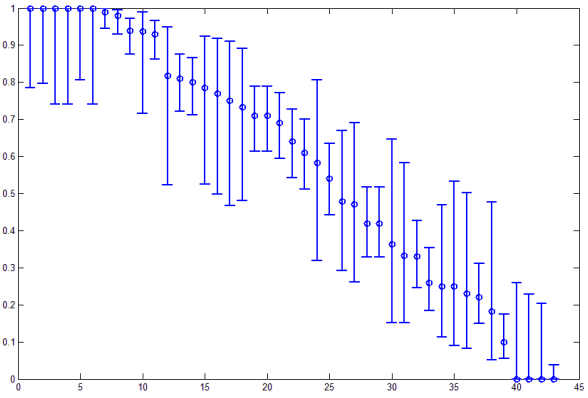
Διάγραμμα 12:3Β



Διάγραμμα 13:3Γ



Διάγραμμα 14:3Δ



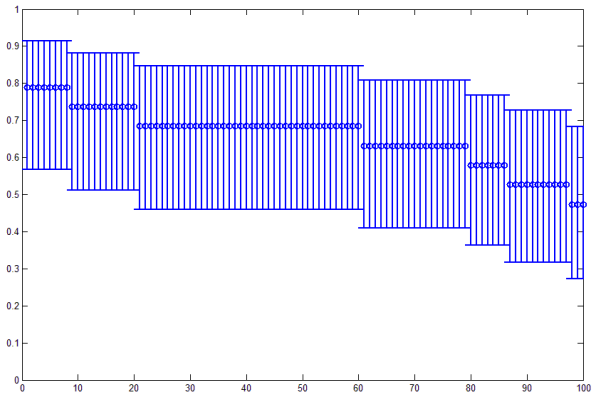
Διάγραμμα 15:3Ε

Η συμπεριφορά της wrapper μεθόδου RFE-SVM, είναι παρόμοια με αυτήν της μεθόδου RFE-FSVs-7DK καθώς:

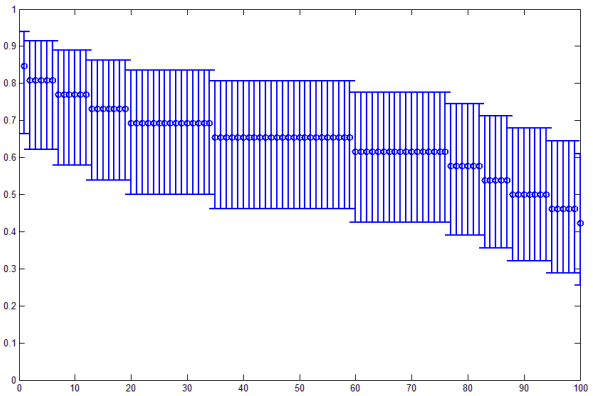
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 3Α (τυπική απόκλιση μέσης ακρίβειας 0.095) από ότι στο 3Β (τυπική απόκλιση μέσης ακρίβειας 0.080)
- το διάγραμμα 3Α έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 3Β
- το διάγραμμα 3Γ έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 3Δ
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 3Γ (τυπική απόκλιση μέσης ακρίβειας 0.354) από ότι στο 3Δ (τυπική απόκλιση μέσης ακρίβειας 0.294)
- το διάγραμμα 3Α (μέση ακρίβεια 0.584) έχει οριακά μεγαλύτερες τιμές ακρίβειας από το 3Β (μέση ακρίβεια 0.580)
- Το διάγραμμα 3Δ (μέση ακρίβεια 0.584) έχει οριακά μεγαλύτερες τιμές ακρίβειας από το 3Γ (μέση ακρίβεια 0.580)

Παρατηρούμε ακόμα ότι στο διάγραμμα 3Β δεν υπάρχει κανένα δείγμα με ακρίβεια μεγαλύτερη ή ίση του 0.8. Τέλος, παρατηρούμε ότι στο διάγραμμα 3Ε υπάρχουν 13 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

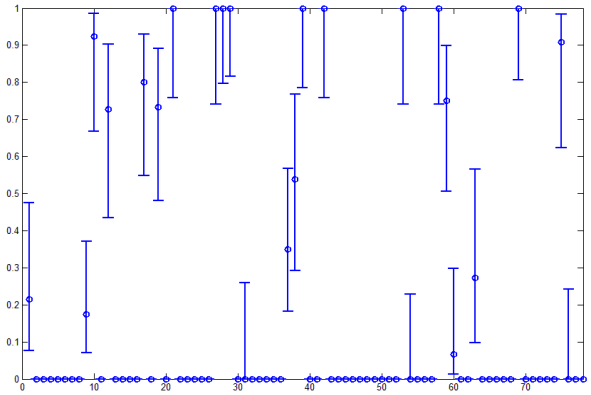
4.ΜΕΘΟΔΟΣ LSSVM ΣΤΑ 45 GENES:



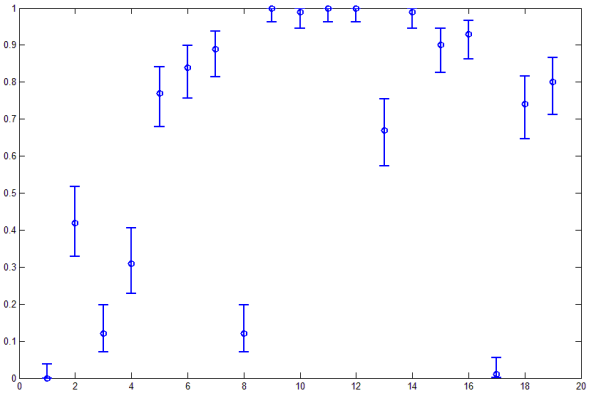
Διάγραμμα 16:4Α



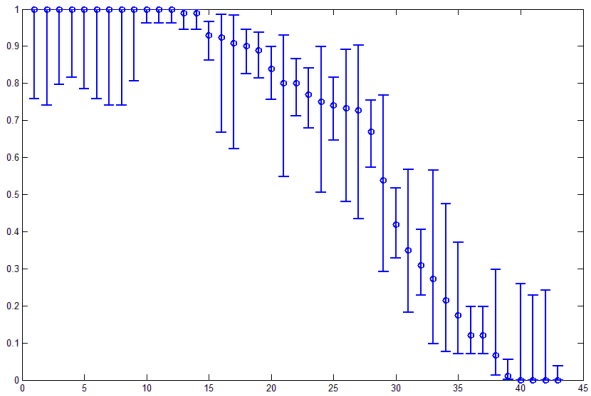
Διάγραμμα 17:4Β



Διάγραμμα 18:4Γ



Διάγραμμα 19:4Δ



Διάγραμμα 20:4Ε

Η συμπεριφορά της wrapper μεθόδου RFE-LSSVM, διαφοροποιείται από αυτήν της RFE-FSVs-7DK μεθόδου και της RFE-SVM, όσον αφορά την τυπική απόκλιση της κάθε ακρίβειας και βλέπουμε ότι έχει μικρότερες τιμές στο διάγραμμα 4Α (τυπική απόκλιση μέσης ακρίβειας) από ότι στο 4Β (τυπική απόκλιση μέσης ακρίβειας).

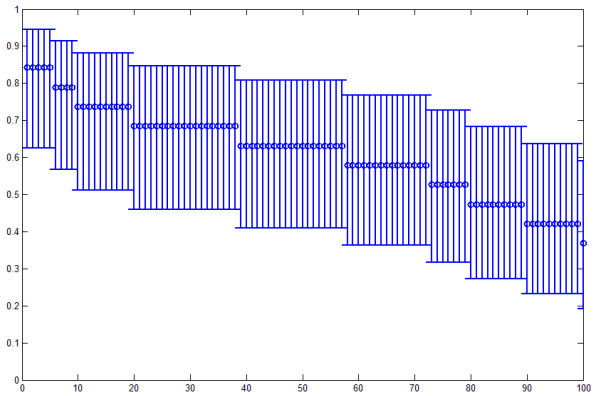
Διαφοροποίηση υπάρχει επιπλέον, στην τυπική απόκλιση της κάθε ακρίβειας όπου βλέπουμε ότι έχει ακριβώς την ίδια τιμή στο διάγραμμα 4Γ (τυπική απόκλιση μέσης ακρίβειας 0.358) με αυτήν του διαγράμματος 4Δ.

Στα υπόλοιπα διαγράμματα παρατηρούμε παρόμοια συμπεριφορά των τρίων μεθόδων καθώς:

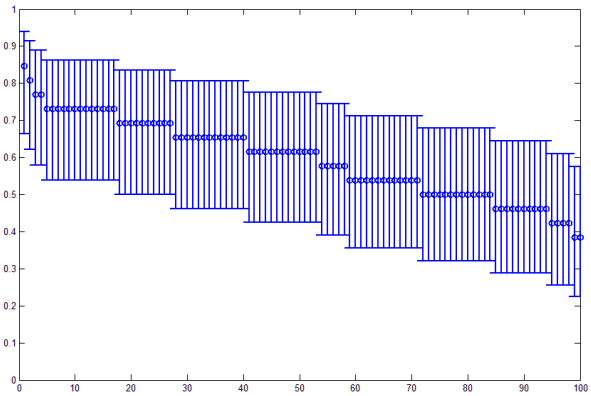
- το διάγραμμα 4Α έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 4Β
- το διάγραμμα 4Γ έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 4Δ
- το διάγραμμα 4Α (μέση ακρίβεια 0.658) έχει μεγαλύτερες τιμές ακρίβειας από το 4Β (μέση ακρίβεια 0.642)
- Το διάγραμμα 4Δ (μέση ακρίβεια 0.658) έχει μεγαλύτερες τιμές ακρίβειας από το 4Γ (μέση ακρίβεια 0.644)

Παρατηρούμε ακόμα ότι στο διάγραμμα 4Β υπάρχουν 6 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.Τέλος, παρατηρούμε ότι στο διάγραμμα 4Ε υπάρχουν 22 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

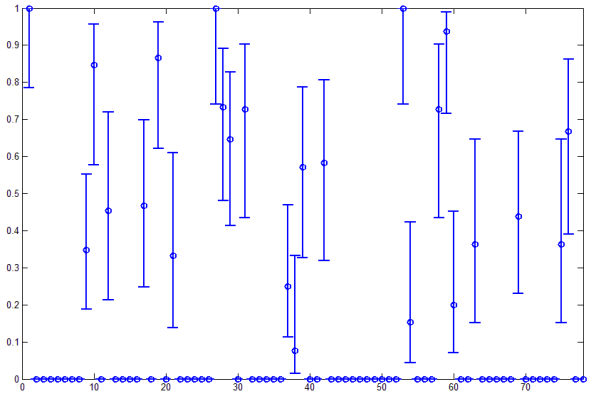
5.ΜΕΘΟΔΟΣ FLD ΣΤΑ 28 GENES:



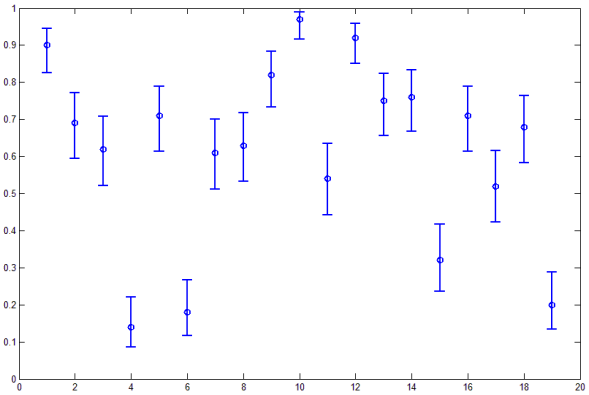
Διάγραμμα 21:5Α



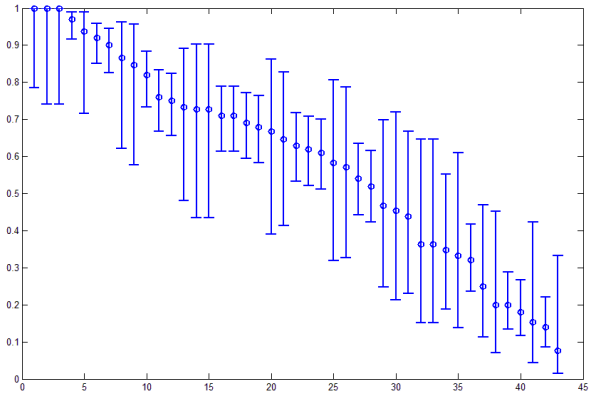
Διάγραμμα 22:5Β



Διάγραμμα 23:5Γ



Διάγραμμα 24:5Δ



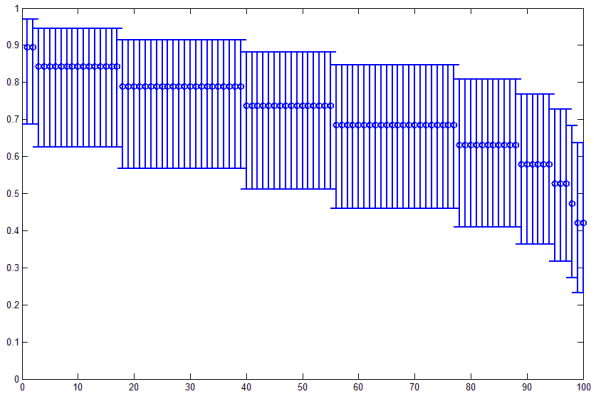
Διάγραμμα 25:5Ε

Η συμπεριφορά της wrapper μεθόδου RFE-FLD, είναι παρόμοια με αυτήν των μεθόδων RFE-FSVs-7DK και RFE-SVM, καθώς:

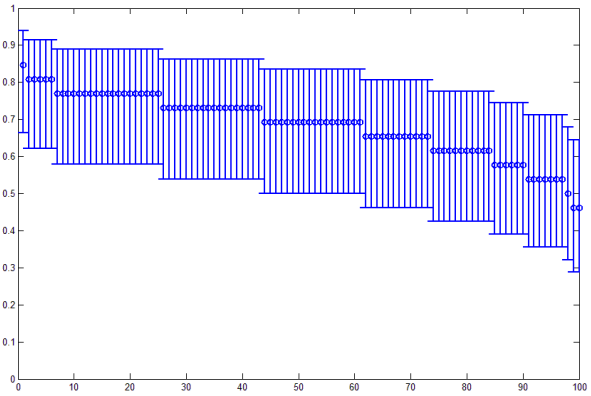
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 5Α (τυπική απόκλιση μέσης ακρίβειας 0.116) από ότι στο 5Β (τυπική απόκλιση μέσης ακρίβειας 0.105)
- το διάγραμμα 5Α έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 5Β
- το διάγραμμα 5Γ έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 5Δ
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 5Γ (τυπική απόκλιση μέσης ακρίβειας 0.276) από ότι στο 5Δ (τυπική απόκλιση μέσης ακρίβειας 0.240)
- το διάγραμμα 5Α (μέση ακρίβεια 0.614) έχει μεγαλύτερες τιμές ακρίβειας από το 5Β (μέση ακρίβεια 0.596)
- Το διάγραμμα 5Δ (μέση ακρίβεια 0.614) έχει μεγαλύτερες τιμές ακρίβειας από το 5Γ (μέση ακρίβεια 0.573)

Παρατηρούμε ακόμα ότι στο διάγραμμα 5Β υπάρχουν 2 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.Τέλος, παρατηρούμε ότι στο διάγραμμα 5Ε υπάρχουν 10 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

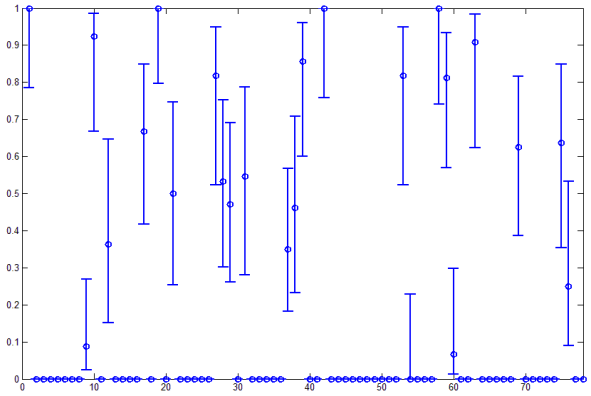
6.ΜΕΘΟΔΟΣ LNW_GD ΣΤΑ 22 GENES:



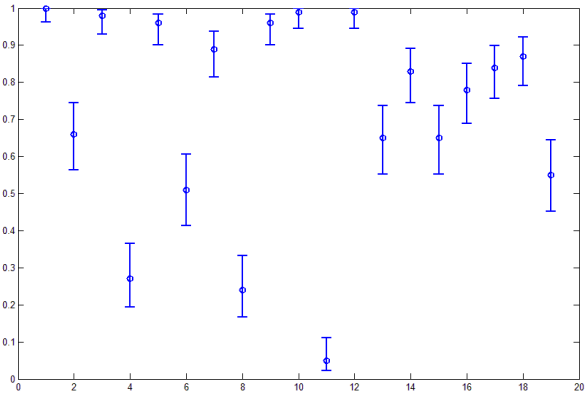
Διάγραμμα 26:6Α



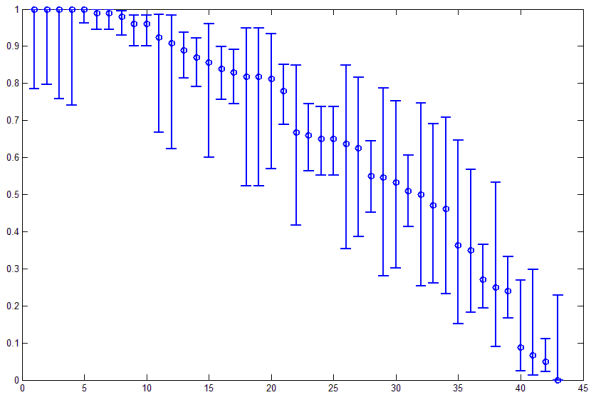
Διάγραμμα 27:6Β



Διάγραμμα 28:6Γ



Διάγραμμα 29:6Δ



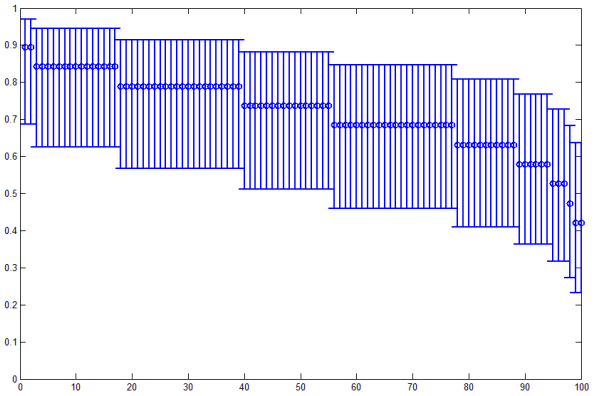
Διάγραμμα 30:6Ε

Η συμπεριφορά της wrapper μεθόδου RFE-LNW-GD, είναι παρόμοια με αυτήν των μεθόδων RFE-FSVs-7DK, RFE-FLD και RFE-SVM, καθώς:

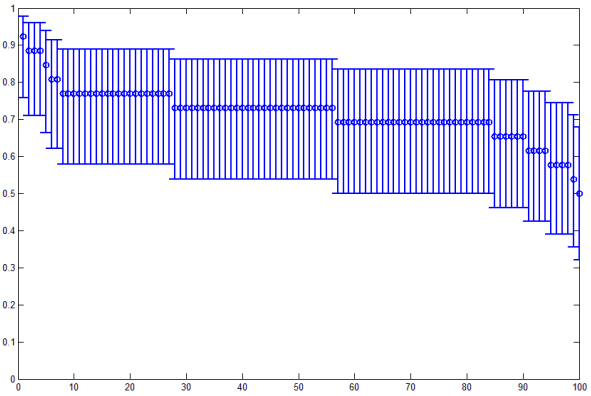
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 6Α (τυπική απόκλιση μέσης ακρίβειας 0.099) από ότι στο 6Β (τυπική απόκλιση μέσης ακρίβειας 0.083)
- το διάγραμμα 6Α έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 6Β
- το διάγραμμα 6Γ έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 6Δ
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 6Γ (τυπική απόκλιση μέσης ακρίβειας 0.306) από ότι στο 6Δ (τυπική απόκλιση μέσης ακρίβειας 0.277)
- το διάγραμμα 6Α (μέση ακρίβεια 0.719) έχει μεγαλύτερες τιμές ακρίβειας από το 6Β (μέση ακρίβεια 0.684)
- Το διάγραμμα 6Δ (μέση ακρίβεια 0.719) έχει μεγαλύτερες τιμές ακρίβειας από το 6Γ (μέση ακρίβεια 0.612)

Παρατηρούμε ακόμα ότι στο διάγραμμα 6Β υπάρχουν 6 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8. Τέλος, παρατηρούμε ότι στο διάγραμμα 6Ε υπάρχουν 43 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

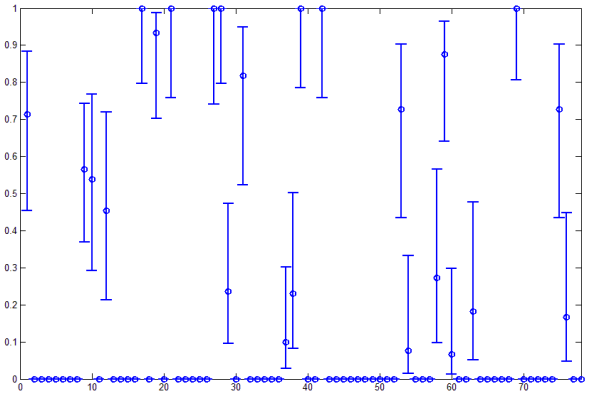
7.ΜΕΘΟΔΟΣ LNW1 ΣΤΑ 44 GENES:



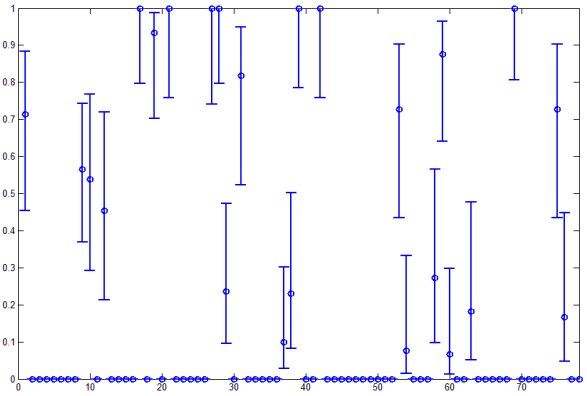
Διάγραμμα 31:7Α



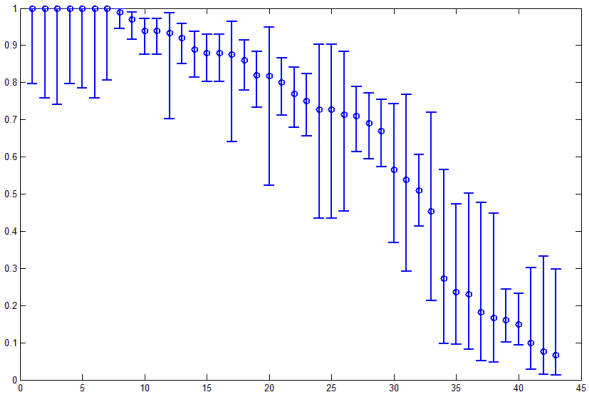
Διάγραμμα 32:7Β



Διάγραμμα 33:7Γ



Διάγραμμα 34:7Δ



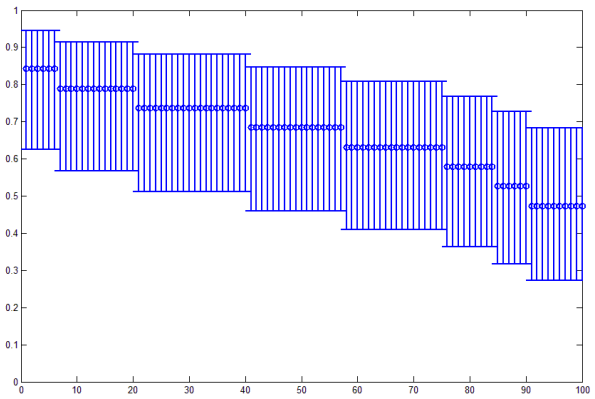
Διάγραμμα 35:7Ε

Η συμπεριφορά της συνδυαστικής μεθόδου RFE-LNW1, είναι παρόμοια με αυτήν των μεθόδων RFE-FSVs-7DK, RFE-FLD, RFE-LNW-GD και RFE-SVM, καθώς:

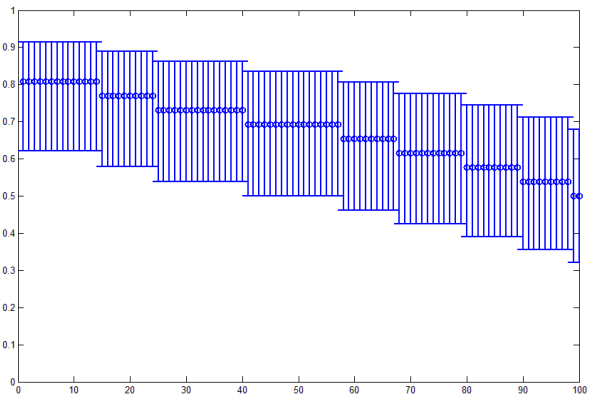
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 7Α (τυπική απόκλιση μέσης ακρίβειας 0.073) από ότι στο 7Β (τυπική απόκλιση μέσης ακρίβειας 0.068)
- το διάγραμμα 7Α έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 7Β
- το διάγραμμα 7Γ έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 7Δ
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 7Γ (τυπική απόκλιση μέσης ακρίβειας 0.352) από ότι στο 7Δ (τυπική απόκλιση μέσης ακρίβειας 0.236)
- το διάγραμμα 7Α (μέση ακρίβεια 0.753) έχει μεγαλύτερες τιμές ακρίβειας από το 7Β (μέση ακρίβεια 0.717)
- Το διάγραμμα 7Δ (μέση ακρίβεια 0.753) έχει μεγαλύτερες τιμές ακρίβειας από το 7Γ (μέση ακρίβεια 0.612)

Παρατηρούμε ακόμα ότι στο διάγραμμα 7Β υπάρχουν 7 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.Τέλος, παρατηρούμε ότι στο διάγραμμα 7Ε υπάρχουν 21 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

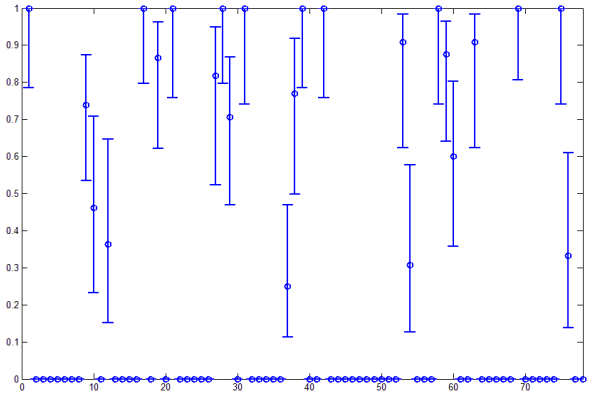
8.ΜΕΘΟΔΟΣ LNW2 ΣΤΑ 64 GENES:



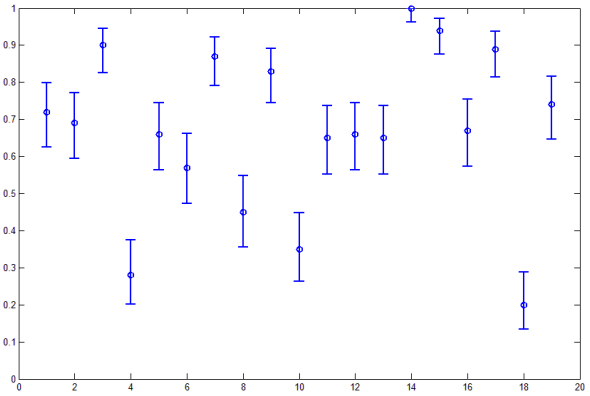
Διάγραμμα 36:8Α



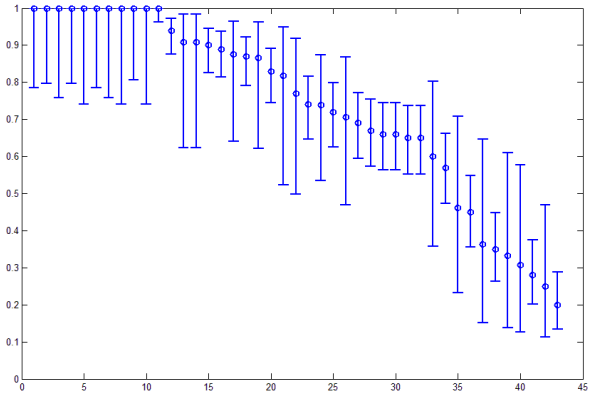
Διάγραμμα 37:8Β



Διάγραμμα 38:8Γ



Διάγραμμα 39:8Δ



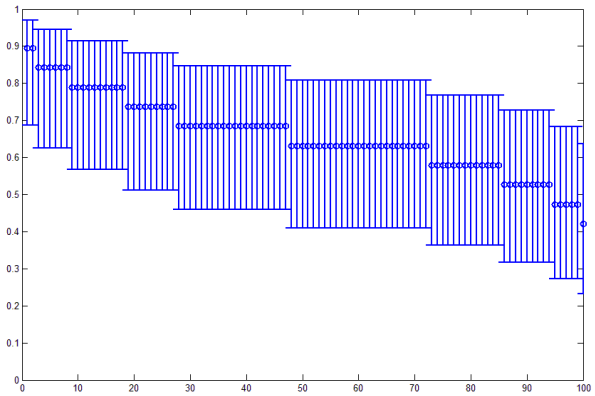
Διάγραμμα 40:8Ε

Η συμπεριφορά της συνδυαστικής μεθόδου RFE-LNW2, είναι παρόμοια με αυτήν της μεθόδου φίλτρου GSM καθώς:

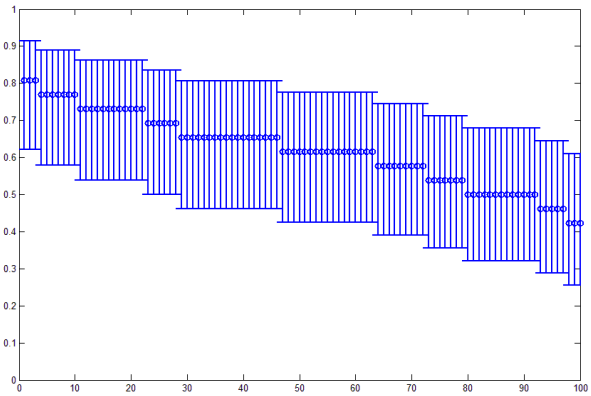
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 8Α (τυπική απόκλιση μέσης ακρίβειας 0.104) από ότι στο 8Β (τυπική απόκλιση μέσης ακρίβειας 0.087)
- το διάγραμμα 8Α έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 8Β
- το διάγραμμα 8Γ έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 8Δ
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 8Γ (τυπική απόκλιση μέσης ακρίβειας 0.254) από ότι στο 8Δ (τυπική απόκλιση μέσης ακρίβειας 0.217)
- το διάγραμμα 8Α (μέση ακρίβεια 0.669) έχει μικρότερες τιμές ακρίβειας από το 8Β (μέση ακρίβεια 0.680) σε αντίθεση με όλες τις υπόλοιπες μεθόδους
- Το διάγραμμα 8Δ (μέση ακρίβεια 0.669) έχει μικρότερες τιμές ακρίβειας από το 8Γ (μέση ακρίβεια 0.788) σε αντίθεση με όλες τις υπόλοιπες μεθόδους

Παρατηρούμε ακόμα ότι στο διάγραμμα 8Β υπάρχουν 14 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8. Τέλος, παρατηρούμε ότι στο διάγραμμα 8Ε υπάρχουν 21 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

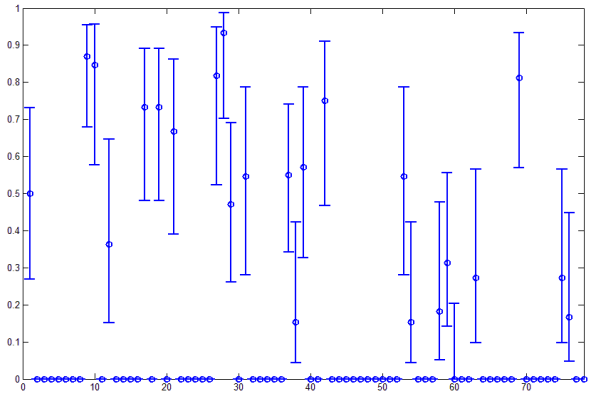
9.ΜΕΘΟΔΟΣ RR ΣΤΑ 7 GENES:



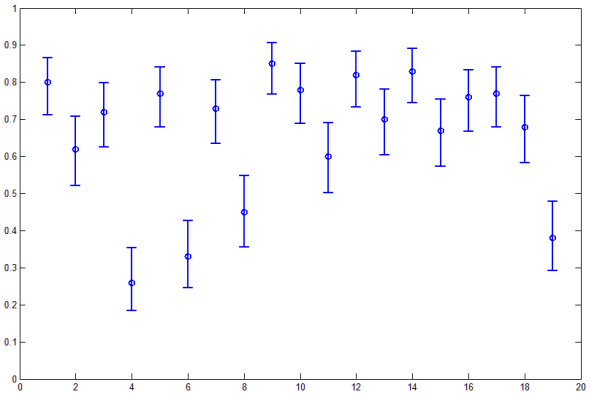
Διάγραμμα 41:9Α



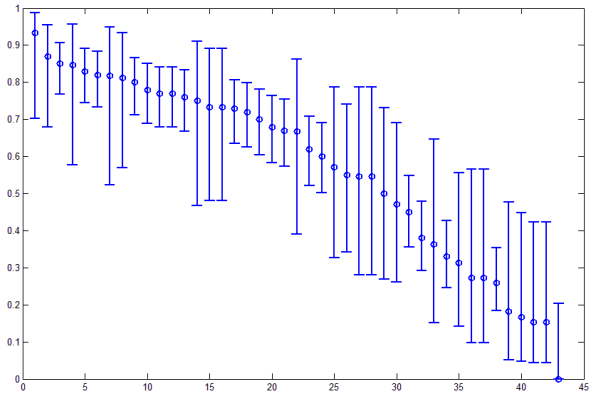
Διάγραμμα 42:9Β



Διάγραμμα 43:9Γ



Διάγραμμα 44:9Δ



Διάγραμμα 45:9Ε

Η συμπεριφορά της wrapper μεθόδου RFE-RR, είναι παρόμοια με αυτήν των μεθόδων RFE-FSVs-7DK, RFE-FLD, RFE-LNW-GD, RFE-LNW1 και RFE-SVM, καθώς:

- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 9Α (τυπική απόκλιση μέσης ακρίβειας 0.102) από ότι στο 9Β (τυπική απόκλιση μέσης ακρίβειας 0.098)
- το διάγραμμα 9Α έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 9Β
- το διάγραμμα 9Γ έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 9Δ
- η τυπική απόκλιση βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 9Γ (τυπική απόκλιση μέσης ακρίβειας 0.267) από ότι στο 9Δ (τυπική απόκλιση μέσης ακρίβειας 0.173)
- το διάγραμμα 9Α (μέση ακρίβεια 0.659) έχει μεγαλύτερες τιμές ακρίβειας από το 9Β (μέση ακρίβεια 0.620)
- Το διάγραμμα 9Δ (μέση ακρίβεια 0.659) έχει μεγαλύτερες τιμές ακρίβειας από το 9Γ (μέση ακρίβεια 0.509)

Παρατηρούμε ακόμα ότι στο διάγραμμα 9Β υπάρχουν 3 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.Τέλος, παρατηρούμε ότι στο διάγραμμα 9Ε υπάρχουν 9 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

Παρουσιάζουμε επίσης έναν συγκεντρωτικό πίνακα με τα μέτρα απόδοσης των παραπάνω αλγορίθμων ώστε να γίνει ευκολότερη η σύγκρισή τους:

	RFE-FSVs-7DK Genes 73	GSM Genes 61	RFE-SVM Genes 32	RFE-LSSVM Genes 45	FLD Genes 28	LNW-GD Genes 22	LNW1 Genes 44	LNW2 Genes 64	RFE-RR Genes 7
Per Subject CV Acc(Pj)	0.668	0.744	0.580	0.644	0.573	0.612	0.612	0.788	0.509
st.deviation for accpj	0.387	0.320	0.354	0.385	0.276	0.306	0.352	0.254	0.267
Mean ci for accpj	0.300	0.324	0.347	0.310	0.398	0.374	0.346	0.332	0.411
st.deviation ci for accpj	0.090	0.102	0.099	0.091	0.084	0.098	0.100	0.097	0.072
r Subject Test Acc(Tj)	0.735	0.660	0.584	0.658	0.614	0.719	0.753	0.669	0.659
st.deviation for accTj	0.221	0.230	0.294	0.358	0.240	0.277	0.236	0.217	0.173
Mean ci for accTj	0.143	0.155	0.146	0.110	0.161	0.128	0.136	0.157	0.170
st.deviation ci for accTj	0.040	0.043	0.049	0.055	0.031	0.052	0.037	0.039	0.015
Per Run CV Acc(Ri)	0.653	0.708	0.570	0.598	0.546	0.587	0.621	0.708	0.514
st.deviation for accRi	0.164	0.174	0.187	0.175	0.199	0.172	0.156	0.158	0.196
Mean ci for accRi	0.552	0.531	0.563	0.562	0.559	0.564	0.563	0.536	0.564
st.deviation ci for accRi	0.054	0.073	0.032	0.047	0.055	0.041	0.039	0.067	0.039
Per Run Test Acc(ri)	0.735	0.660	0.584	0.658	0.614	0.719	0.753	0.669	0.659
st.deviation for accri	0.089	0.102	0.095	0.077	0.116	0.099	0.073	0.104	0.102
Mean ci for accri	0.363	0.384	0.399	0.388	0.391	0.367	0.359	0.381	0.384
st.deviation ci for accri	0.034	0.027	0.016	0.017	0.022	0.030	0.030	0.025	0.028
Per Run All Acc(Rri)	0.713	0.673	0.580	0.642	0.596	0.684	0.717	0.680	0.620
st.deviation for accRri	0.077	0.091	0.080	0.090	0.105	0.083	0.068	0.087	0.098
Mean ci for accRri	0.324	0.333	0.351	0.340	0.346	0.332	0.324	0.332	0.343
st.deviation ci for accRri	0.024	0.025	0.010	0.018	0.016	0.019	0.022	0.021	0.017

Πίνακας 5: Μέτρα απόδοσης και όρια εμπιστοσύνης για τις 9 μεθόδους που δοκιμάστηκαν στα δεδομένα του καρκίνου του μαστού.

Για τρεις από τις παραπάνω μεθόδους γίνεται σύγκριση με τα δεδομένα που σχετίζονται με την λευχαιμία. Οι μέθοδοι αυτές εφαρμόστηκαν στα δεδομένα καρκίνου του μαστού και παρακάτω παραθέτουμε την μέση ακρίβεια και την εξέλιξή της για όλα τα τρεξίματα από τα 100 γονίδια και κάτω.

Η μέσες τιμές της ακρίβειας ανά αριθμό γονιδίων, θεωρούμε ότι ακολουθούν κανονική κατανομή επειδή ο μέσος ενός μεγάλου πληθυσμού τείνει να ακολουθεί κανονική κατανομή ακόμα και αν δεν προέρχεται από κανονικό δείγμα, όπως ισχύει στην συγκεκριμένη περίπτωση. ‘Ετσι για να υπολογίσουμε το διάστημα εμπιστοσύνης κάθε μέσου, χρησιμοποιούμε τον τύπο υπολογισμού του διαστήματος εμπιστοσύνης για τον μέσο κανονικού πληθυσμού: $\left[\bar{X} - \frac{\sigma}{\sqrt{n}} Z_{a/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} Z_{a/2} \right]$

Συνεπώς, από κάθε τιμή της μέσης ακρίβειας ανά αριθμό γονιδίων, προσθέτουμε και αφαιρούμε αντίστοιχα την ποσότητα 1.96*τυπική απόκλιση, ώστε να προκύψει το όριο εμπιστοσύνης σαν ένα παραπάνω μέτρο της απόδοσης της κάθε μεθόδου.

Παρακάτω παρατίθενται 7 διαγράμματα μέσων τιμών ακριβείας για κάθε έναν από τους 3 αλγορίθμους τα οποία είναι:

Α. Μέση ακρίβεια ασθενών του cross validation, στον υπολογισμό της οποίας συμμετέχουν μόνο όσοι ασθενείς του cross validation εμφανίζονται περισσότερες από 10 φορές στα test set και τυπική απόκλιση για τους ίδιου ασθενείς που έχουν όμως και ακρίβεια μεγαλύτερη από 0.4, ανά αριθμό γονιδίων

Β. Μέση ακρίβεια ασθενών του cross validation, στον υπολογισμό της οποίας συμμετέχουν όλοι οι ασθενείς και τυπική απόκλιση υπολογισμένη μόνο για όσους έχουν ακρίβεια μεγαλύτερη του 0.5, ανά αριθμό γονιδίων

Γ. Μέση ακρίβεια των ασθενών του ανεξάρτητου test set, υπολογισμένη μόνο για όσους έχουν ακρίβεια μεγαλύτερη του 0.4 και τυπική απόκλιση υπολογισμένη για τους ίδιους ασθενείς, ανά αριθμό γονιδίων

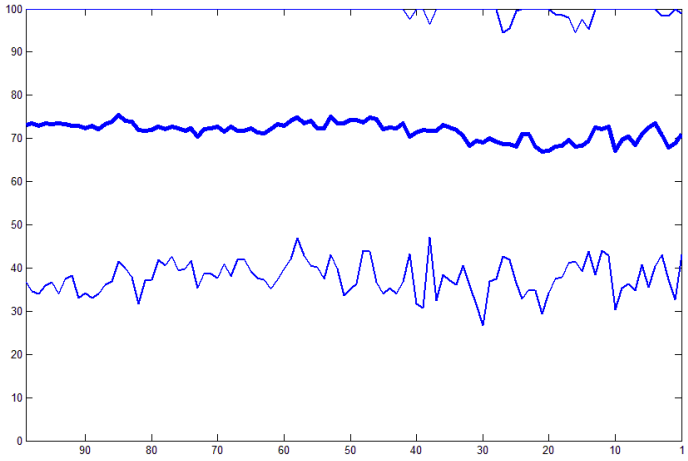
Δ. Μέση ακρίβεια των ασθενών του ανεξάρτητου test set, υπολογισμένη για όλους τους ασθενείς και τυπική απόκλιση υπολογισμένη μόνο για όσους ασθενείς του ανεξάρτητου test set έχουν ακρίβεια μεγαλύτερη του 0.4 για τον RFE-SVM, μεγαλύτερη του 0.5 για τον GSM και μεγαλύτερη του 0.6 για τον RFE-FSVs-7DK, ανά αριθμό γονιδίων

Ε. Μέση ακρίβεια των τρεξιμάτων του cross validation, ανά αριθμό γονιδίων

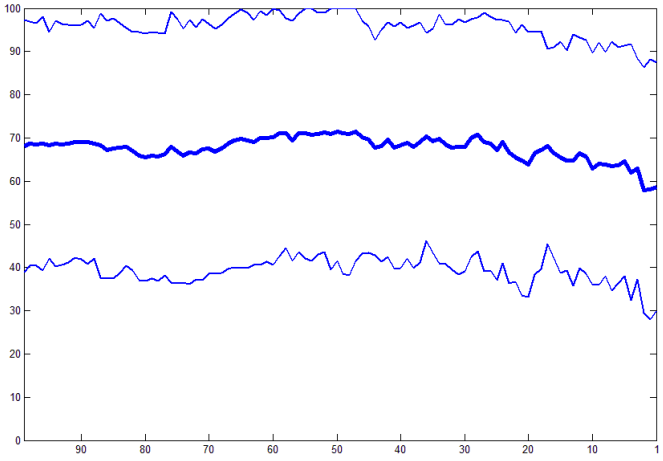
ΣΤ. Μέση ακρίβεια των τρεξιμάτων του ανεξάρτητου test set, ανά αριθμό γονιδίων

Ζ. Μέση ακρίβεια των τρεξιμάτων συνολικά, ανά αριθμό γονιδίων

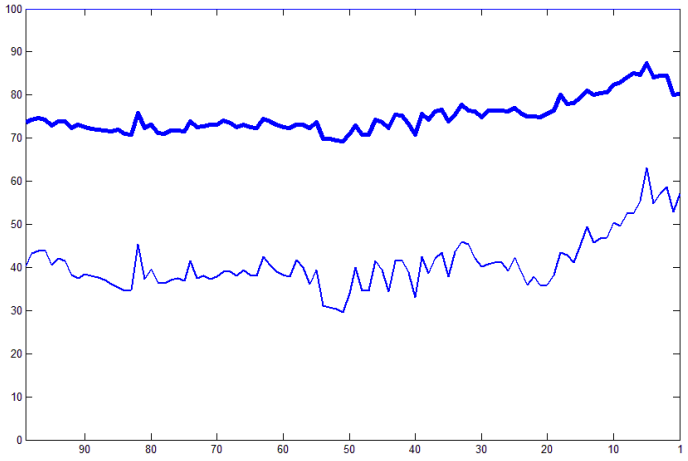
1. ΜΕΘΟΔΟΣ GSM:



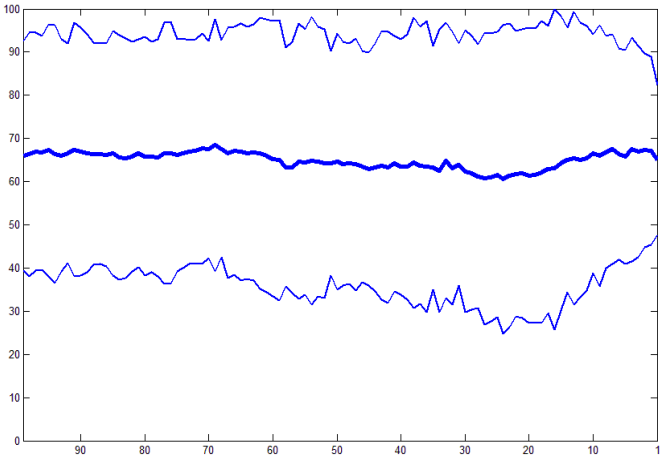
Διάγραμμα 46:1Α



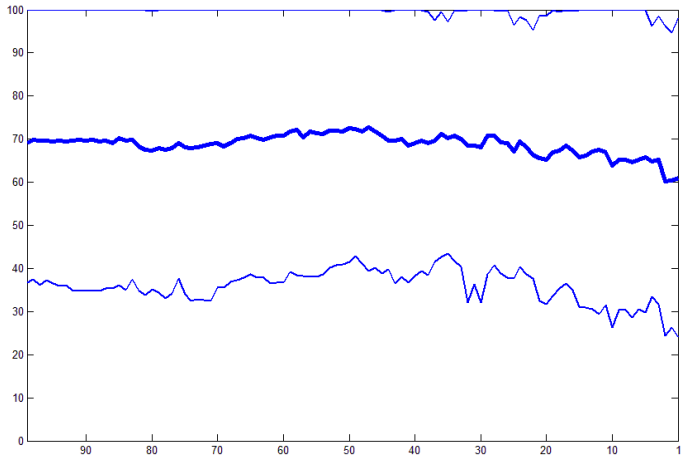
Διάγραμμα 47:1Β



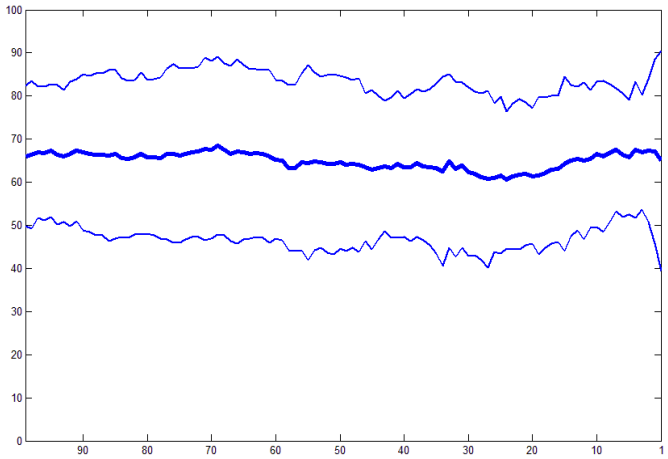
Διάγραμμα 48:1Γ



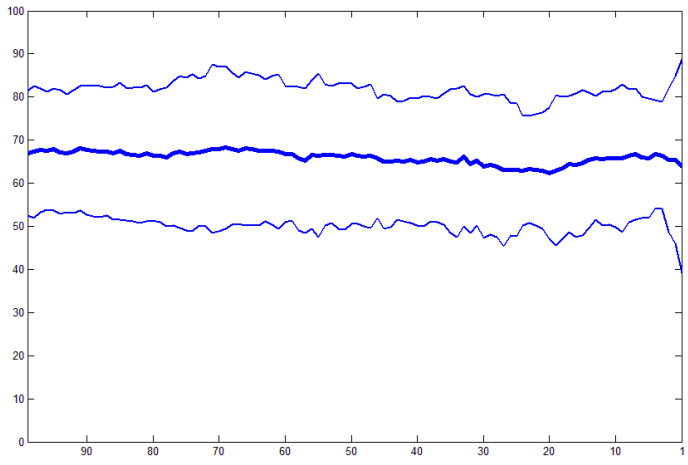
Διάγραμμα 49:1Δ



Διάγραμμα 50:1Ε



Διάγραμμα 51:1ΣΤ



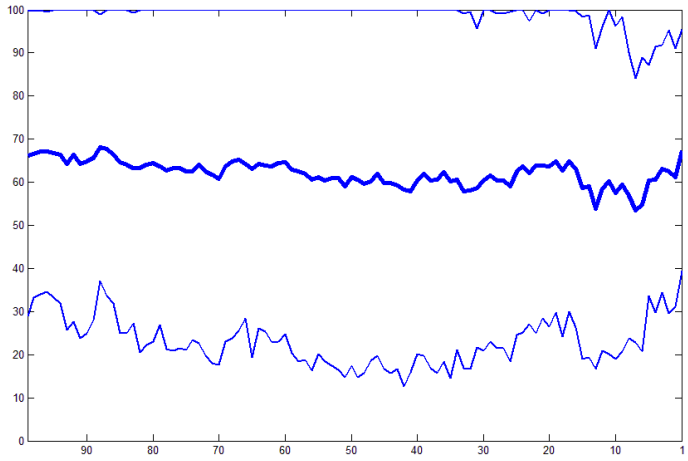
Διάγραμμα 52:1Ζ

Αυτό που μπορούμε να παρατηρήσουμε είναι ότι στο διάγραμμα 1Α επιτυγχάνονται μεγαλύτερες τιμές ακρίβειας απ' ότι στο διάγραμμα 1Β για τους ασθενείς του cross validation αν και έχουμε μεγαλύτερα διαστήματα εμπιστοσύνης λόγω μεγαλύτερης τυπικής απόκλισης.

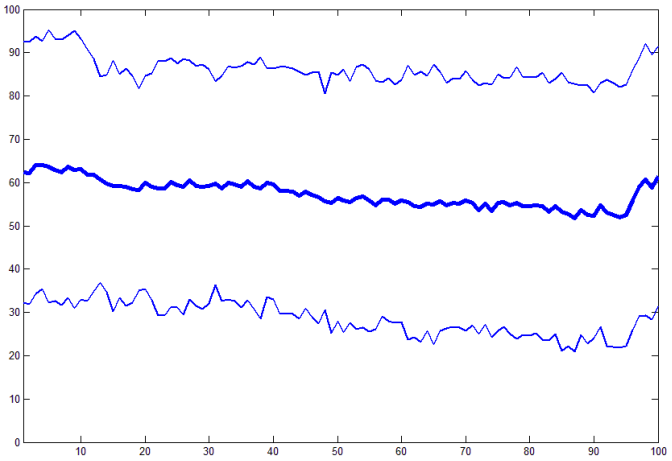
Σχετικά με τους ασθενείς του ανεξάρτητου test set, βλέπουμε ότι πετυχαίνουμε μεγαλύτερες τιμές ακρίβειας στο διάγραμμα 1Γ από αυτές του διαγράμματος 1Δ, αν και πάλι έχουμε μεγαλύτερες τυπικές αποκλίσεις, άρα και μεγαλύτερα διαστήματα εμπιστοσύνης στο διάγραμμα 1Γ.

Όσον αφορά τα διαγράμματα 1Ε,1ΣΤ,1Ζ που αναφέρονται στις μέσες ακρίβειες ανά τρέξιμο, βλέπουμε ότι η συγκεκριμένη μέθοδος πετυχαίνει καλύτερες ακρίβειες στα τρεξίματα του cross validation. Η απόδοση που πετυχαίνει στο ανεξάρτητο test set είναι μικρότερη με αποτέλεσμα να μειώνεται και η απόδοση της μεθόδου στα συνολικά τρεξίματα σε σχέση με την απόδοση που πετυχαίνει ξεχωριστά στα τρεξίματα του cross validation.

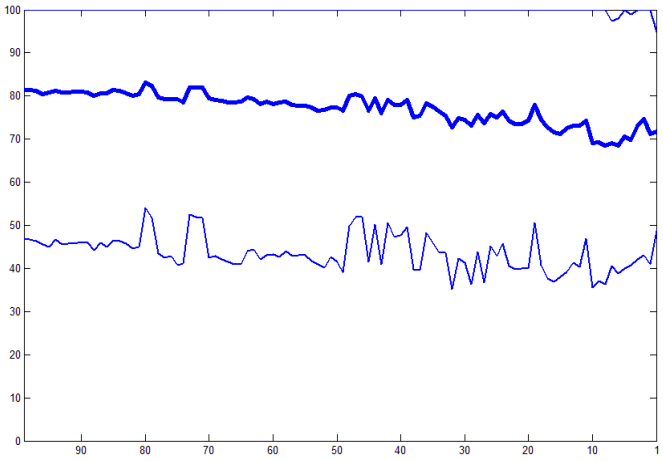
Επιπλέον, τα μικρότερα διαστήματα εμπιστοσύνης τα παρατηρούμε στο διάγραμμα 1Ζ όπου υπάρχουν και οι μικρότερες τυπικές αποκλίσεις. Στη συνέχεια, έχουμε την μικρότερη τυπική απόκλιση στα τρεξίματα του ανεξάρτητου test set (διάγραμμα 1ΣΤ) και το διάγραμμα 1Ε έχει τα μεγαλύτερα διαστήματα εμπιστοσύνης μεταξύ των διαγραμμάτων που απεικονίζουν μέσες τιμές ακρίβειας των τρεξιμάτων της μεθόδου.



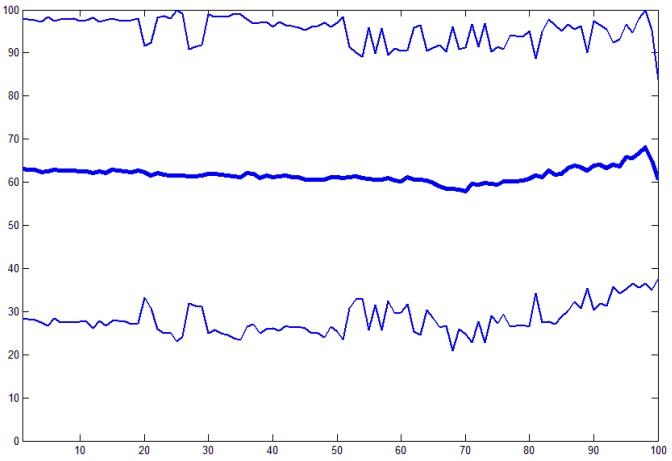
Διάγραμμα 53:2Α



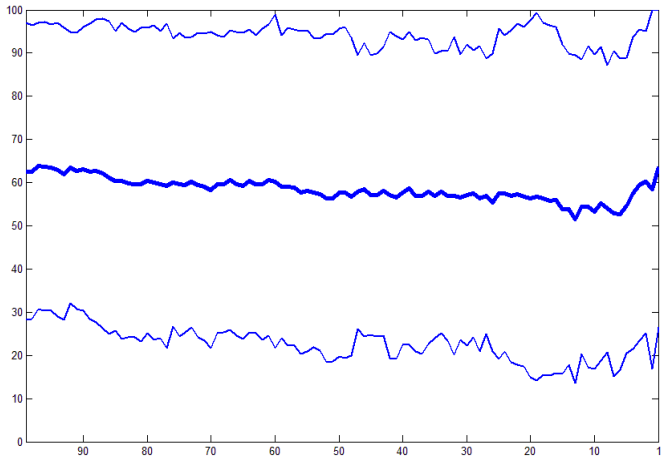
Διάγραμμα 54:2Β



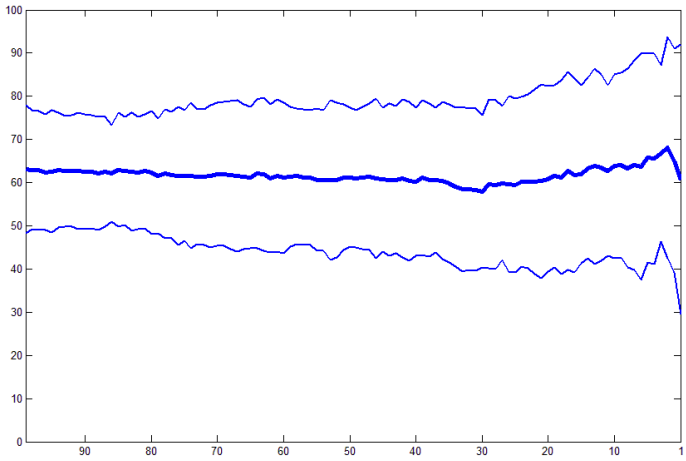
Διάγραμμα 55:2Γ



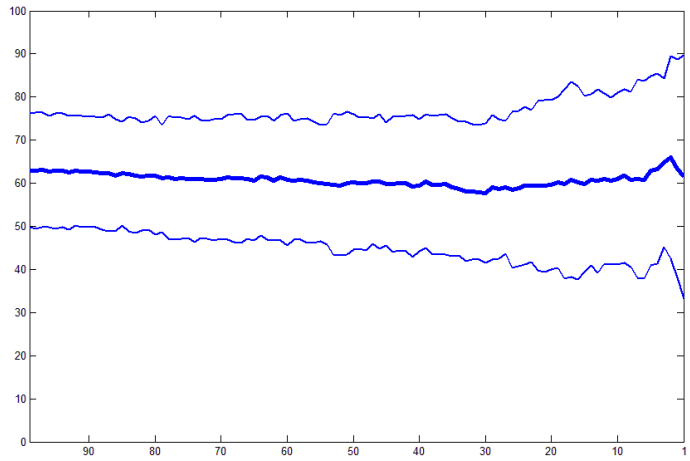
Διάγραμμα 56:2Δ



Διάγραμμα 57:2Ε



Διάγραμμα 58:2ΣΤ



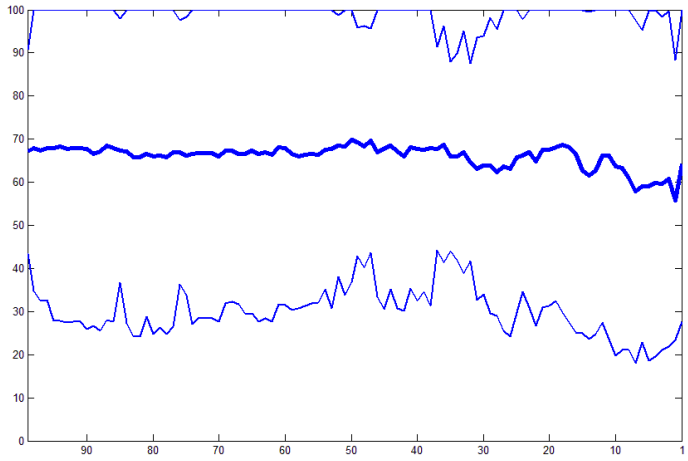
Διάγραμμα 59:2Ζ

Αυτό που μπορούμε να παρατηρήσουμε είναι ότι στο διάγραμμα 2Α επιτυγχάνονται μεγαλύτερες τιμές ακρίβειας απ' ότι στο διάγραμμα 2Β για τους ασθενείς του cross validation αν και έχουμε μεγαλύτερα διαστήματα εμπιστοσύνης λόγω μεγαλύτερης τυπικής απόκλισης.

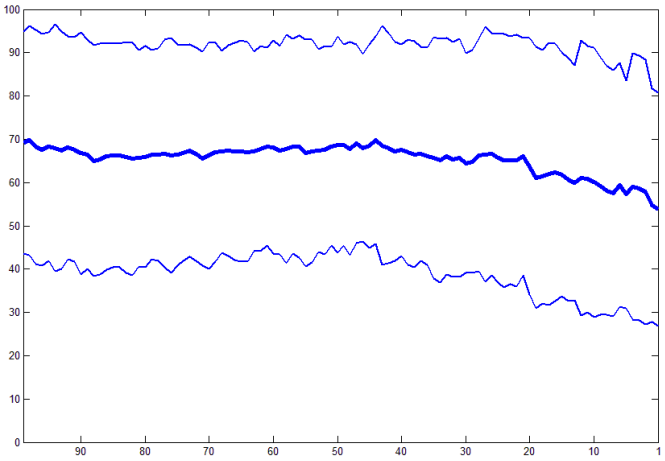
Σχετικά με τους ασθενείς του ανεξάρτητου test set, βλέπουμε ότι πετυχαίνουμε μεγαλύτερες τιμές ακρίβειας στο διάγραμμα 2Γ από αυτές του διαγράμματος 2Δ αλλά σε αυτήν την μέθοδο έχουμε σχεδόν ίδια διαστήματα εμπιστοσύνης για τα δύο διαγράμματα, καθώς έχουμε σχεδόν ίδιες τυπικές αποκλίσεις.

Όσον αφορά τα διαγράμματα 2Ε,2ΣΤ,2Ζ που αναφέρονται στις μέσες ακρίβειες ανά τρέξιμο, βλέπουμε ότι η συγκεκριμένη μέθοδος πετυχαίνει καλύτερες ακρίβειες στα τρεξίματα του ανεξάρτητου test set. Η απόδοση που πετυχαίνει στο cross validation είναι μικρότερη με αποτέλεσμα να μειώνεται και η απόδοση της μεθόδου στα συνολικά τρεξίματα σε σχέση με την απόδοση που πετυχαίνει ξεχωριστά στα τρεξίματα του ανεξάρτητου test set.

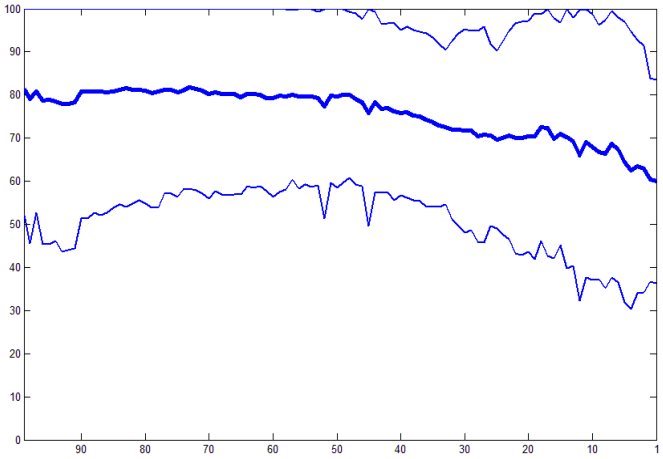
Επιπλέον, τα μικρότερα διαστήματα εμπιστοσύνης τα παρατηρούμε στο διάγραμμα 2Ζ όπου υπάρχουν και οι μικρότερες τυπικές αποκλίσεις. Στη συνέχεια, έχουμε την μικρότερη τυπική απόκλιση στα τρεξίματα του ανεξάρτητου test set (διάγραμμα 2ΣΤ) και το διάγραμμα 2Ε έχει τα μεγαλύτερα διαστήματα εμπιστοσύνης μεταξύ των διαγραμμάτων που απεικονίζουν μέσες τιμές ακρίβειας των τρεξιμάτων της μεθόδου.



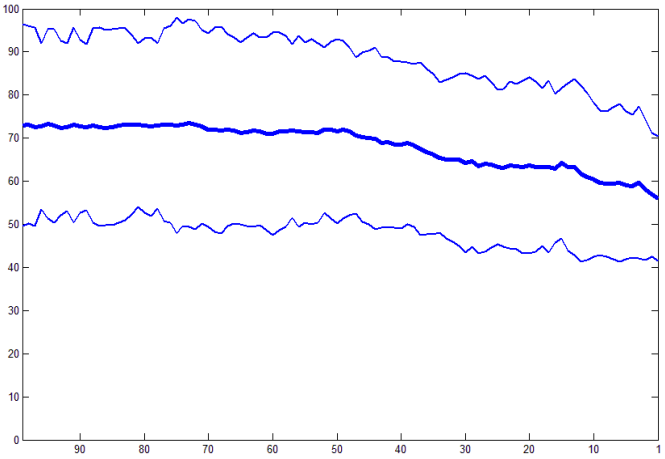
Διάγραμμα 60:3Α



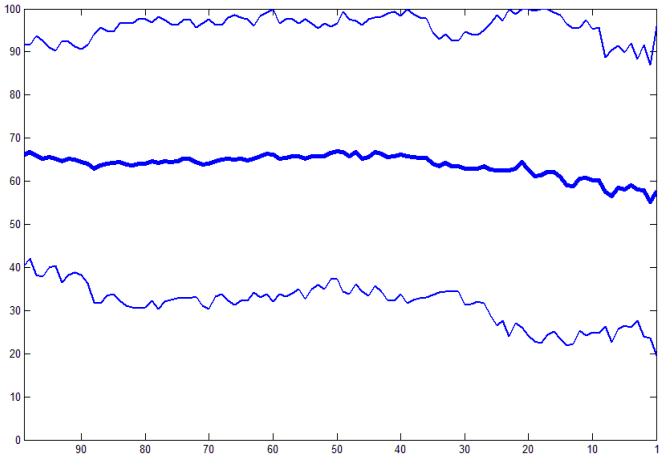
Διάγραμμα 61:3Β



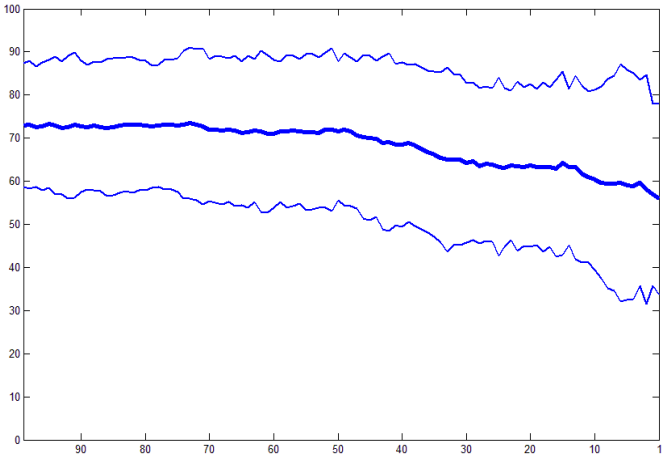
Διάγραμμα 62:3Γ



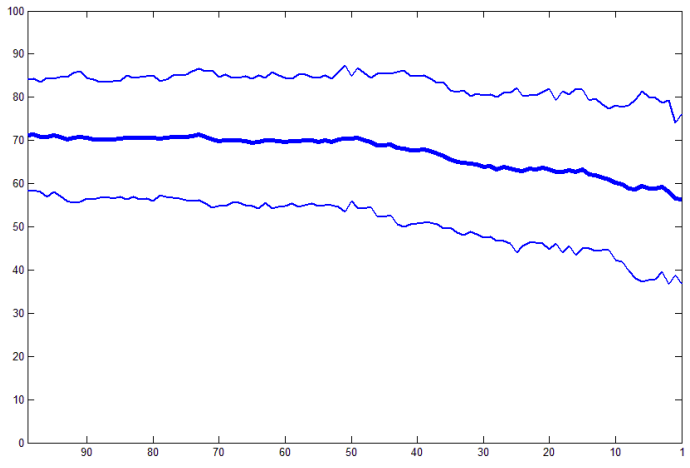
Διάγραμμα 63:3Δ



Διάγραμμα 64:3Ε



Διάγραμμα 65:3ΣΤ



Διάγραμμα 66:3Ζ

Αυτό που μπορούμε να παρατηρήσουμε είναι ότι στο διάγραμμα 3Α επιτυγχάνονται μεγαλύτερες τιμές ακρίβειας απ' ότι στο διάγραμμα 3Β για τους ασθενείς του cross validation αν και έχουμε μεγαλύτερα διαστήματα εμπιστοσύνης λόγω μεγαλύτερης τυπικής απόκλισης.

Σχετικά με τους ασθενείς του ανεξάρτητου test set, βλέπουμε ότι πετυχαίνουμε μεγαλύτερες τιμές ακρίβειας στο διάγραμμα 3Γ από αυτές του διαγράμματος 3Δ, αν και πάλι έχουμε μεγαλύτερες τυπικές αποκλίσεις, άρα και μεγαλύτερα διαστήματα εμπιστοσύνης στο διάγραμμα 3Γ.

Όσον αφορά τα διαγράμματα 3Ε, 3ΣΤ, 3Ζ που αναφέρονται στις μέσες ακρίβειες ανά τρέξιμο, βλέπουμε ότι η συγκεκριμένη μέθοδος πετυχαίνει καλύτερες ακρίβειες στα τρεξίματα του ανεξάρτητου test set. Η απόδοση που πετυχαίνει στο cross validation είναι μικρότερη με αποτέλεσμα να μειώνεται και η απόδοση της μεθόδου στα συνολικά τρεξίματα σε σχέση με την απόδοση που πετυχαίνει ξεχωριστά στα τρεξίματα του ανεξάρτητου test set.

Επιπλέον, τα μικρότερα διαστήματα εμπιστοσύνης τα παρατηρούμε στο διάγραμμα 3Ζ όπου υπάρχουν και οι μικρότερες τυπικές αποκλίσεις. Στη συνέχεια, έχουμε την μικρότερη τυπική απόκλιση στα τρεξίματα του ανεξάρτητου test set (διάγραμμα 3ΣΤ) και το διάγραμμα 3Ε έχει τα μεγαλύτερα διαστήματα εμπιστοσύνης μεταξύ των διαγραμμάτων που απεικονίζουν μέσες τιμές ακρίβειας των τρεξιμάτων της μεθόδου.

4.2.Αποτελέσματα δεδομένων λευχαιμίας

Για κάθε μέθοδο από τις 3 που υλοποιήσαμε υπολογίσαμε τέσσερα μεγέθη:

- Την ακρίβεια της μεθόδου ανά τρέξιμο για τους 34 ασθενείς του ανεξάρτητου test set(accri)
- Την ακρίβεια της μεθόδου ανά τρέξιμο για όλους ασθενείς του ανεξάρτητου test set και του test set(accRri)
- Την ακρίβεια της μεθόδου ανά ασθενή για τους ασθενείς του test set που εμφανίστηκαν πάνω από 10 φορές συνολικά στα test set των 4 ασθενών, στα 100 τρεξίματα της μεθόδου(accpj)
- την ακρίβεια της μεθόδου ανά ασθενή για όλους τους ασθενείς του ανεξάρτητου test set(acctj)

Για τα παραπάνω αυτά τα μεγέθη, υπολογίσαμε τα όρια εμπιστοσύνης, θεωρώντας ότι ακολουθούν διωνυμική κατανομή ως άθροισμα δοκιμών Bernouli, με βάση τον τύπο:

$$\left(\hat{p} + \frac{z_{a/2}^2}{2n} \pm z_{a/2} \sqrt{\left[\hat{p}\left(1-\hat{p}\right) + z_{a/2}^2 / 4n \right] / n} \right) / \left(1 + z_{a/2}^2 / n\right)$$

όπου, $\hat{p}$ είναι η μέση τιμή του κάθε μέτρου απόδοσης όπως αυτή προκύπτει από τις πειραματικές μετρήσεις, Z_α είναι η ποσότητα που προκύπτει από το άνω α-σημείο της κανονικής κατανομής και n είναι το πλήθος των στοιχείων.

Για κάθε μέθοδο, έχουμε υπολογίσει τον βέλτιστο αριθμό γονιδίων ο οποίος είναι αυτός στον οποίο πετυχαίνουμε απόλυτη ακρίβεια στο training set και την καλύτερη ακρίβεια στο ανεξάρτητο test set των 34 ασθενών. Όλα τα αποτελέσματα παρουσιάζονται για αυτόν τον αριθμό ο οποίο είναι διαφορετικός για κάθε αλγόριθμο:

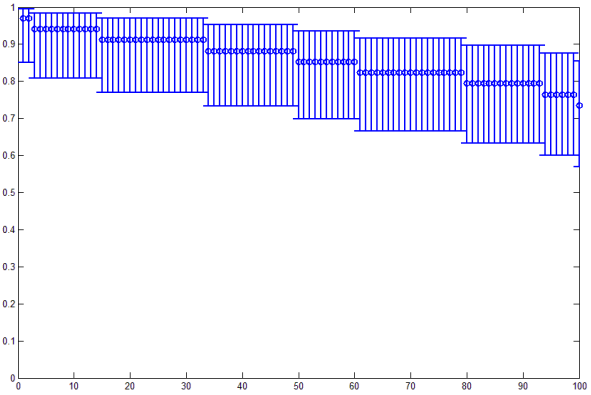
ΟΝΟΜΑ ΜΕΘΟΔΟΥ	ΒΕΛΤΙΣΤΟΣ ΑΡΙΘΜΟΣ ΓΟΝΙΔΙΩΝ
RFE-FSVs-7DK	45
GSM	32
SVM	24

Πίνακας 6: Μέθοδοι που υλοποιήθηκαν και δοκιμάστηκαν στο σύνολο δεδομένων της λευχαιμίας και ο αριθμός γονιδίων στον οποίο πετυχαίνουν την καλύτερη ακρίβεια.

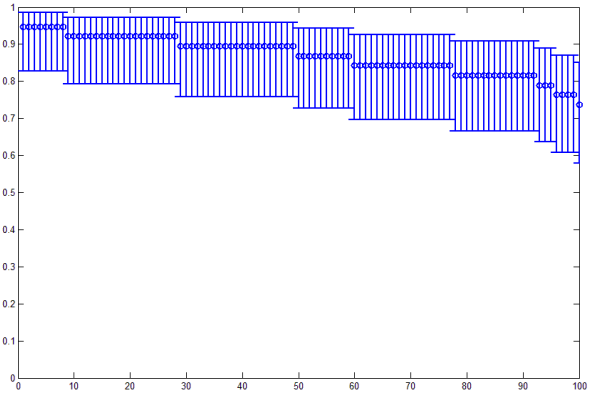
Παρακάτω, παρουσιάζονται για τους 3 αλγορίθμους, 5 διαγράμματα τα οποία είναι:

Α. Ακρίβεια για όλους τους ασθενείς του ανεξάρτητου test set ανά τρέξιμο (accri)
Β. Ακρίβεια για όλους τους ασθενείς του test set που προκύπτει από το cross validation και τους ασθενείς του ανεξάρτητου test set συνολικά, ανά τρέξιμο (accRri)
Γ. Ακρίβεια για όλα τα τρεξίματα, ανά ασθενή του cross validation, μόνο για όσους εμφανίζονται περισσότερες από 10 φορές σε test set(accpj με συχνότητα >10)
Δ. Ακρίβεια για όλα τα τρεξίματα, ανά ασθενή του ανεξάρτητου test set(acctj)
Ε. Ακρίβεια για όλα τα τρεξίματα ανά ασθενή, για όλους τους ασθενείς του ανεξάρτητου test set και όσους ασθενείς του cross validation εμφανίζονται περισσότερες από 10 φορές σε test set(acctj,accpj)

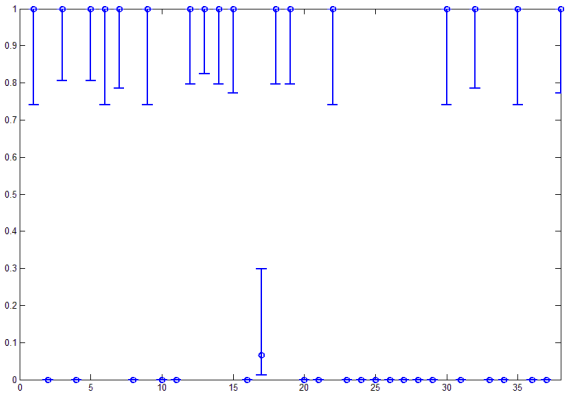
1.ΜΕΘΟΔΟΣ FSVs-4DK ΣΤΑ 45 GENES:



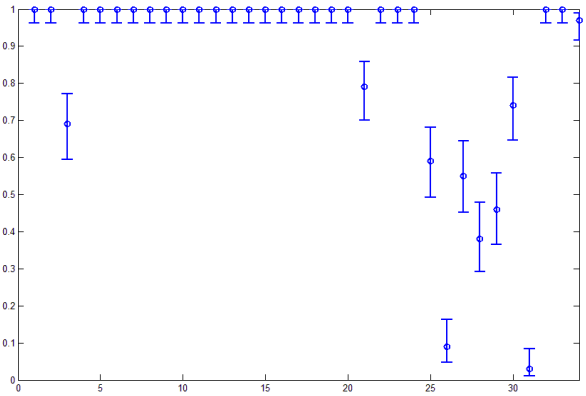
Διάγραμμα 67:1Α



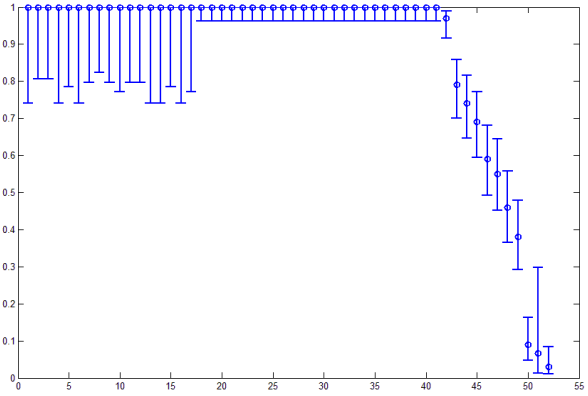
Διάγραμμα 68:1Β



Διάγραμμα 69:1Γ



Διάγραμμα 70:1Δ



Διάγραμμα 71:1Ε

Παρατηρούμε ότι η ακρίβεια για όλους τους ασθενείς του test set που προκύπτει από το cross validation και τους ασθενείς του ανεξάρτητου test set συνολικά, ανά τρέξιμο (διάγραμμα 1Β), έχει πιο μικρά διαστήματα εμπιστοσύνης από την ακρίβεια για τους ασθενείς του ανεξάρτητου test set ανά τρέξιμο (διάγραμμα 1Α).

Αυτό συμβαίνει γιατί, στο διάγραμμα 1Α έχουμε μικρότερο N_i στον υπολογισμό του διαστήματος εμπιστοσύνης καθώς πάντα έχουμε 34 δείγματα που κατηγοριοποιούνται, σε αντίθεση με το διάγραμμα 1Β που εξετάζει την κατηγοριοποίηση 34 δειγμάτων από το ανεξάρτητο test set και 4 δειγμάτων από το cross validation, σύνολο δηλαδή 38 δειγμάτων.

Επιπλέον, υπολογίζουμε την τυπική απόκλιση της κάθε ακρίβειας και βλέπουμε ότι έχει μεγαλύτερες τιμές στο διάγραμμα 1Α (τυπική απόκλιση μέσης ακρίβειας 0.057) από ότι στο 1Β (τυπική απόκλιση μέσης ακρίβειας 0.051). Παρατηρούμε ακόμα ότι στο διάγραμμα 1Β υπάρχουν 92 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

Το διάγραμμα 1Α (μέση ακρίβεια 0.861) έχει γενικά μικρότερες τιμές ακρίβειας από το 1Β (μέση ακρίβεια 0.870), γιατί η συγκεκριμένη μέθοδος παρουσιάζει μικρότερη ακρίβεια στην κατηγοριοποίηση του ανεξάρτητου test set από ότι στην κατηγοριοποίηση του cross validation, με αποτέλεσμα η απόδοση του cross validation να αυξάνει την απόδοση της μεθόδου στο σύνολο των ασθενών (διάγραμμα 1Β) σε σχέση με αυτήν στο ανεξάρτητο test set (διάγραμμα 1Α).

Όσον αφορά τα διαγράμματα που απεικονίζουν την ακρίβεια για όλα τα τρεξίματα, ανά ασθενή του cross validation, μόνο για όσους εμφανίζονται περισσότερες από 10 φορές σε test set (διάγραμμα 1Γ) και την ακρίβεια για όλα τα τρεξίματα, ανά ασθενή του ανεξάρτητου test set (διάγραμμα 1Δ), παρατηρούμε ότι το διάγραμμα 1Γ έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 1Δ.

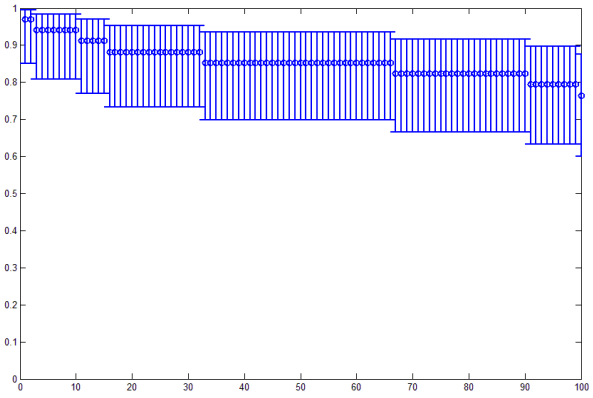
Αυτό συμβαίνει γιατί, στο διάγραμμα 1Γ έχουμε μικρότερο N_i στον υπολογισμό του διαστήματος εμπιστοσύνης καθώς το N_i είναι η συχνότητα με την οποία εμφανίζεται ένας ασθενής του cross validation στα test sets σε 100 τρεξίματα, σε αντίθεση με το διάγραμμα 1Δ στο οποίο κάθε ασθενής κατηγοριοποιείται όλες τις φορές στα 100 τρεξίματα της μεθόδου, επομένως η συχνότητα εμφάνισής του είναι πάντα 100. Το μέγεθος N_i επηρεάζει καθοριστικά το μέγεθος του διαστήματος εμπιστοσύνης ακόμα και αν στον υπολογισμό του τελευταίου για το διάγραμμα 1Γ συμμετέχουν μόνο οι ασθενείς με N_i μεγαλύτερο του 10.

Επιπλέον, υπολογίζουμε την τυπική απόκλιση της κάθε ακρίβειας και βλέπουμε ότι έχει μικρότερες τιμές στο διάγραμμα 1Γ (τυπική απόκλιση μέσης ακρίβειας 0.220) από ότι στο 1Δ (τυπική απόκλιση μέσης ακρίβειας 0.268).

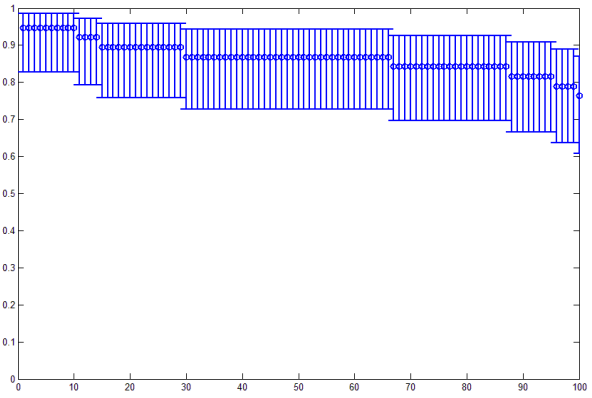
Το διάγραμμα 1Δ (μέση ακρίβεια 0.861) έχει γενικά μικρότερες τιμές ακρίβειας από το 1Γ (μέση ακρίβεια 0.948), γιατί η συγκεκριμένη μέθοδος παρουσιάζει μικρότερη ακρίβεια στην κατηγοριοποίηση του ανεξάρτητου test set (διάγραμμα 1Δ) από ότι στην κατηγοριοποίηση του cross validation (διάγραμμα 1Γ).

Τέλος, παρατηρούμε ότι στο διάγραμμα 1Ε υπάρχουν 42 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

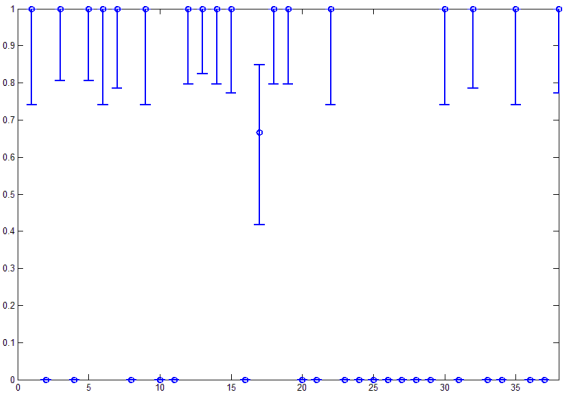
2. ΜΕΘΟΔΟΣ GSM ΣΤΑ 32 GENES:



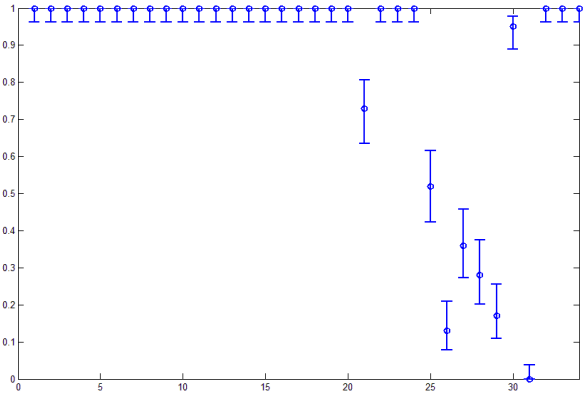
Διάγραμμα 72:2Α



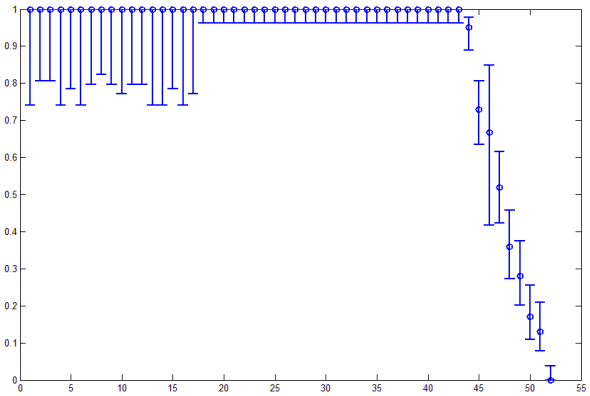
Διάγραμμα 73:2Β



Διάγραμμα 74:2Γ



Διάγραμμα 75:2Δ



Διάγραμμα 76:2Ε

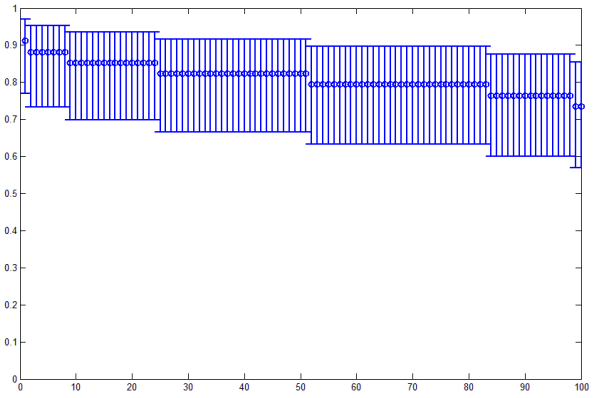
Η μέθοδος φίλτρου GSM έχει παρόμοια συμπεριφορά με την συνδυαστική μέθοδο RFE-FSVs-4DK καθώς:

- το διάγραμμα 2Α (μέση ακρίβεια 0.857) έχει γενικά μικρότερες τιμές ακρίβειας από το 2Β (μέση ακρίβεια 0.868)
- το διάγραμμα 2Γ έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 2Δ.
- η τυπική απόκλιση έχει μικρότερες τιμές στο διάγραμμα 2Γ (τυπική απόκλιση μέσης ακρίβειας 0.079) από ότι στο 2Δ (τυπική απόκλιση μέσης ακρίβειας 0.301).
- Το διάγραμμα 2Δ (μέση ακρίβεια 0.857) έχει γενικά μικρότερες τιμές ακρίβειας από το 2Γ (μέση ακρίβεια 0.981)

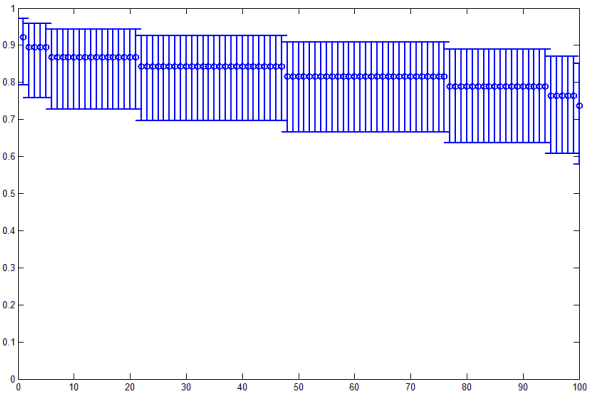
Η μόνη διαφοροποίηση μεταξύ των δύο μεθόδων είναι στην τυπική απόκλιση, η οποία έχει μεγαλύτερες τιμές στο διάγραμμα 2Α (τυπική απόκλιση μέσης ακρίβειας 0.043) από ότι στο 2Β (τυπική απόκλιση μέσης ακρίβειας 0.040) στην μέθοδο GSM, σε αντίθεση με την RFE-FSVs-4DK. Συνεπώς στην μέθοδο φίλτρου παρατηρούμε και μεγαλύτερα διαστήματα εμπιστοσύνης στο διάγραμμα 2Α από ότι στο 2Β.

Παρατηρούμε ακόμα ότι στο διάγραμμα 2Β υπάρχουν 95 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.Τέλος, παρατηρούμε ότι στο διάγραμμα 2Ε υπάρχουν 44 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

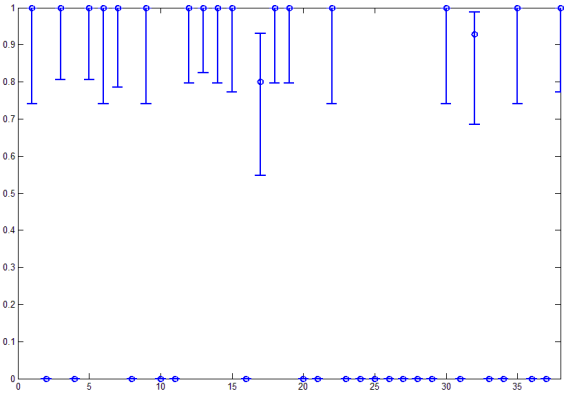
3. ΜΕΘΟΔΟΣ SVM ΣΤΑ 24 GENES:



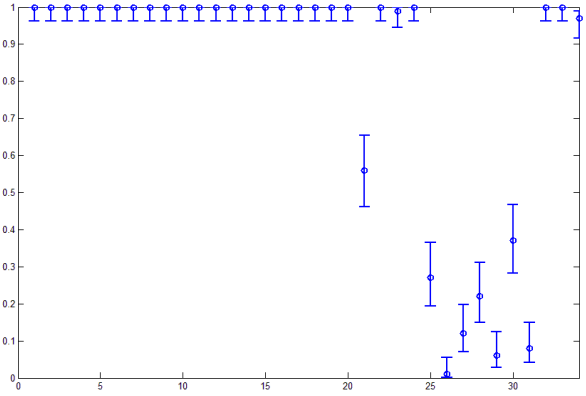
Διάγραμμα 77:3Α



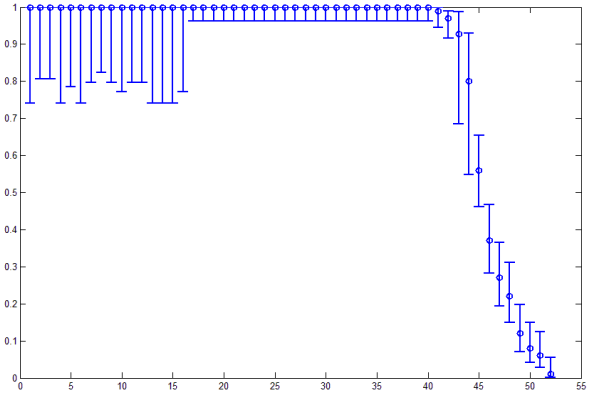
Διάγραμμα 78:3Β



Διάγραμμα 79:3Γ



Διάγραμμα 80:3Δ



Διάγραμμα 81:3Ε

Η wrapper μέθοδος RFE-SVM έχει παρόμοια συμπεριφορά με την συνδυαστική μέθοδο RFE-FSVs-4DK καθώς:

- το διάγραμμα 3Α (μέση ακρίβεια 0.813) έχει γενικά μικρότερες τιμές ακρίβειας από το 3Β (μέση ακρίβεια 0.827)
- το διάγραμμα 3Γ έχει μεγαλύτερα διαστήματα εμπιστοσύνης από το 3Δ
- η τυπική απόκλιση έχει μικρότερες τιμές στο διάγραμμα 3Γ (τυπική απόκλιση μέσης ακρίβειας 0.049) από ότι στο 3Δ (τυπική απόκλιση μέσης ακρίβειας 0.350)
- το διάγραμμα 3Δ (μέση ακρίβεια 0.813) έχει γενικά μικρότερες τιμές ακρίβειας από το 3Γ (μέση ακρίβεια 0.985)
- η ακρίβεια στο διάγραμμα 3Β, έχει πιο μικρά διαστήματα εμπιστοσύνης από την ακρίβεια στο διάγραμμα 3Α
- η τυπική απόκλιση στο διάγραμμα 3Α (τυπική απόκλιση μέσης ακρίβειας 0.036) είναι μεγαλύτερη από ότι στο 3Β (τυπική απόκλιση μέσης ακρίβειας 0.035)

Παρατηρούμε ακόμα ότι στο διάγραμμα 3Β υπάρχουν 76 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.Τέλος, παρατηρούμε ότι στο διάγραμμα 3Ε υπάρχουν 44 δείγματα που έχουν ακρίβεια μεγαλύτερη ή ίση του 0.8.

Από τα παραπάνω διαγράμματα βλέπουμε ότι στο σύνολο δεδομένων που σχετίζεται με τον καρκίνο του μαστού γίνεται καλύτερη κατηγοριοποίηση στο test set του cross validation με τις περισσότερες μεθόδους, ενώ στο σύνολο που σχετίζεται με την λευχαιμία γίνεται καλύτερη κατηγοριοποίηση στο ανεξάρτητο test set με όλες τις μεθόδους.

Παρουσιάζουμε επίσης έναν συγκεντρωτικό πίνακα με τα μέτρα απόδοσης των παραπάνω αλγορίθμων, ώστε να είναι ευκολότερη η σύγκρισή τους:

	RFE-FSVs-4DK Genes 45	GSM Genes 32	RFE-SVM Genes 24
Per Subject CV Acc(Pj)	0.948	0.981	0.985
st.deviation for accpj	0.220	0.079	0.049
Per Subject Test Acc(Tj)	0.861	0.857	0.813
st.deviation for accTj	0.268	0.301	0.350
Per Run CV Acc(Ri)	0.945	0.965	0.945
st.deviation for accRi	0.104	0.087	0.110
Per Run Test Acc(ri)	0.861	0.857	0.813
st.deviation for accri	0.057	0.043	0.036
Per Run All Acc(Rri)	0.870	0.868	0.827
st.deviation for accRri	0.051	0.040	0.035

Πίνακας 7: Μέτρα απόδοσης και όρια εμπιστοσύνης για τις 3 μεθόδους που δοκιμάστηκαν στα δεδομένα της λευχαιμίας.

Για τις τρεις παραπάνω μεθόδους γίνεται σύγκριση με τα δεδομένα που σχετίζονται με τον καρκίνο του μαστού. Οι μέθοδοι αυτές εφαρμόστηκαν στα δεδομένα της λευχαιμίας και παρακάτω παραθέτουμε την μέση ακρίβεια και την εξέλιξή της για όλα τα τρεξίματα από τα 100 γονίδια και κάτω.

Η μέσες τιμές της ακρίβειας ανά αριθμό γονιδίων, θεωρούμε ότι ακολουθούν κανονική κατανομή επειδή ο μέσος ενός μεγάλου πληθυσμού τείνει να ακολουθεί κανονική κατανομή ακόμα και αν δεν προέρχεται από κανονικό δείγμα, όπως ισχύει στην συγκεκριμένη περίπτωση. 'Ετσι για να υπολογίσουμε το διάστημα εμπιστοσύνης κάθε μέσου, χρησιμοποιούμε τον τύπο υπολογισμού του διαστήματος εμπιστοσύνης για τον μέσο κανονικού πληθυσμού: $\left[\bar{X} - \frac{\sigma}{\sqrt{n}} Z_{a/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} Z_{a/2} \right]$

Συνεπώς, από κάθε τιμή της μέσης ακρίβειας ανά αριθμό γονιδίων, προσθέτουμε και αφαιρούμε αντίστοιχα την ποσότητα 1.96*τυπική απόκλιση,

ώστε να προκύψει το όριο εμπιστοσύνης σαν ένα παραπάνω μέτρο της απόδοσης της κάθε μεθόδου.

Παρακάτω παρατίθενται 7 διαγράμματα μέσων τιμών ακριβείας για κάθε έναν από τους 3 αλγορίθμους τα οποία είναι:

Α. Μέση ακρίβεια ασθενών του cross validation, στον υπολογισμό της οποίας συμμετέχουν μόνο όσοι ασθενείς του cross validation εμφανίζονται περισσότερες από 10 φορές στα test set και τυπική απόκλιση για τους ίδιου ασθενείς που έχουν όμως και ακρίβεια μεγαλύτερη από 0.4, ανά αριθμό γονιδίων

Β. Μέση ακρίβεια ασθενών του cross validation, στον υπολογισμό της οποίας συμμετέχουν όλοι οι ασθενείς και τυπική απόκλιση υπολογισμένη μόνο για όσους έχουν ακρίβεια μεγαλύτερη του 0.5, ανά αριθμό γονιδίων

Γ. Μέση ακρίβεια των ασθενών του ανεξάρτητου test set, υπολογισμένη μόνο για όσους έχουν ακρίβεια μεγαλύτερη του 0.4 και τυπική απόκλιση υπολογισμένη για τους ίδιους ασθενείς, ανά αριθμό γονιδίων

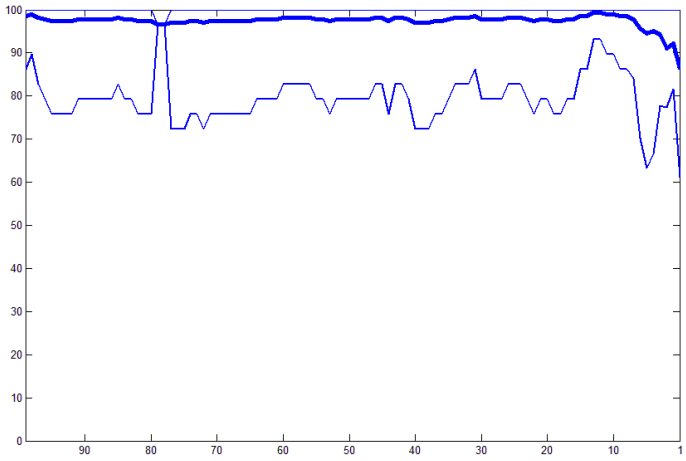
Δ. Μέση ακρίβεια των ασθενών του ανεξάρτητου test set, υπολογισμένη για όλους τους ασθενείς και τυπική απόκλιση υπολογισμένη μόνο για όσους ασθενείς του ανεξάρτητου test set έχουν ακρίβεια μεγαλύτερη του 0.4 για τον RFE-SVM, μεγαλύτερη του 0.5 για τον GSM και μεγαλύτερη του 0.6 για τον RFE-FSVs-7DK, ανά αριθμό γονιδίων

Ε. Μέση ακρίβεια των τρεξιμάτων του cross validation, ανά αριθμό γονιδίων

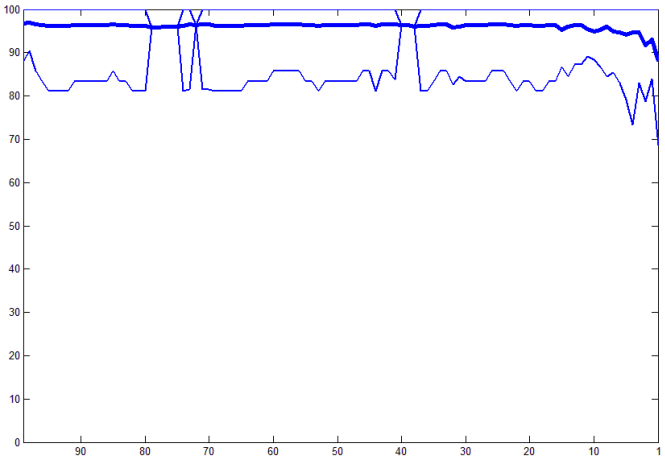
ΣΤ. Μέση ακρίβεια των τρεξιμάτων του ανεξάρτητου test set, ανά αριθμό γονιδίων

Ζ. Μέση ακρίβεια των τρεξιμάτων συνολικά, ανά αριθμό γονιδίων

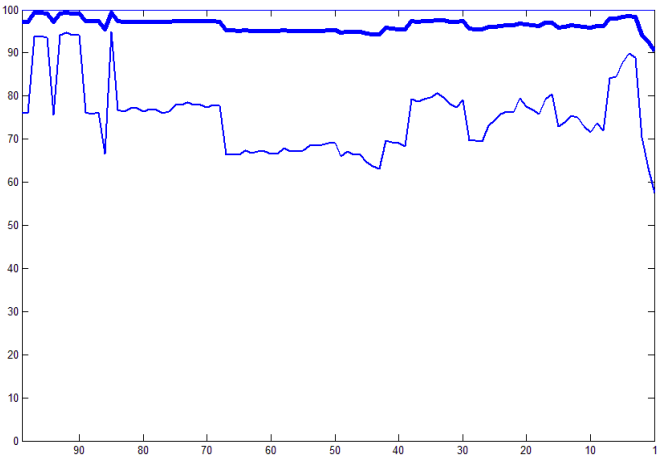
1. ΜΕΘΟΔΟΣ GSM:



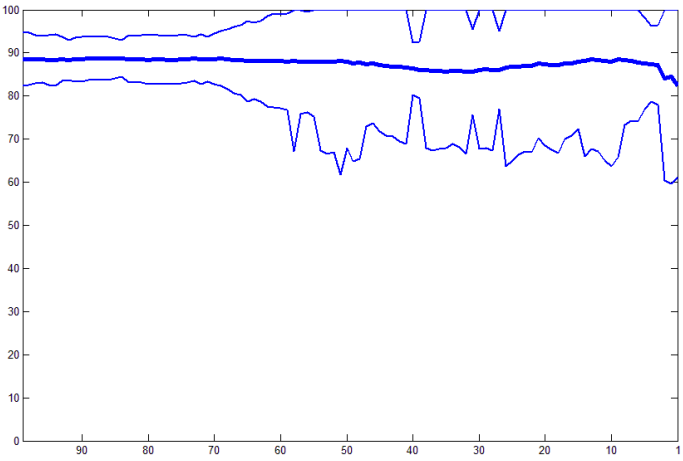
Διάγραμμα 82:1Α



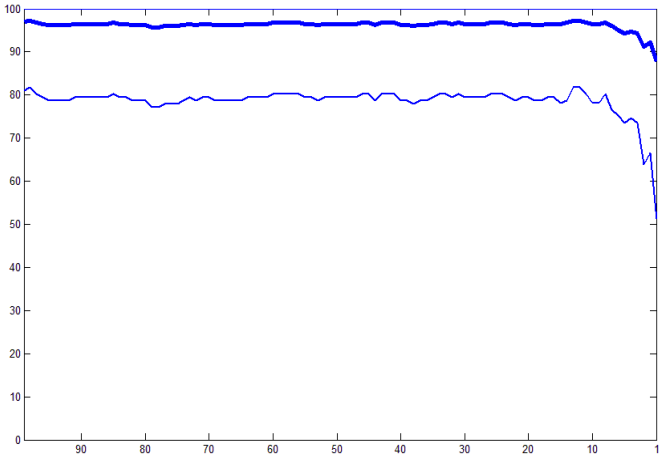
Διάγραμμα 83:1Β



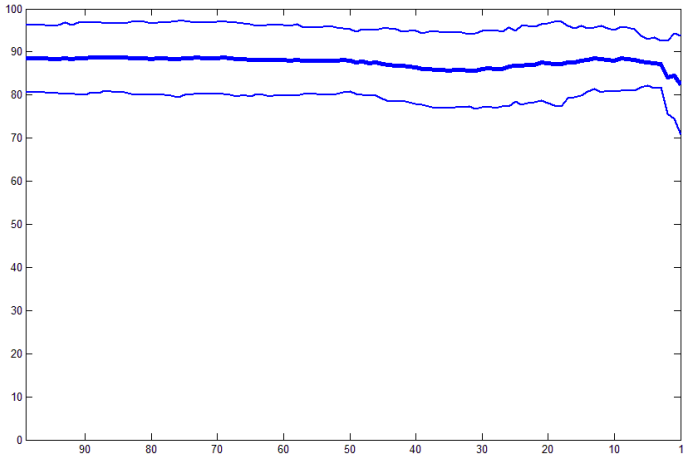
Διάγραμμα 84:1Γ



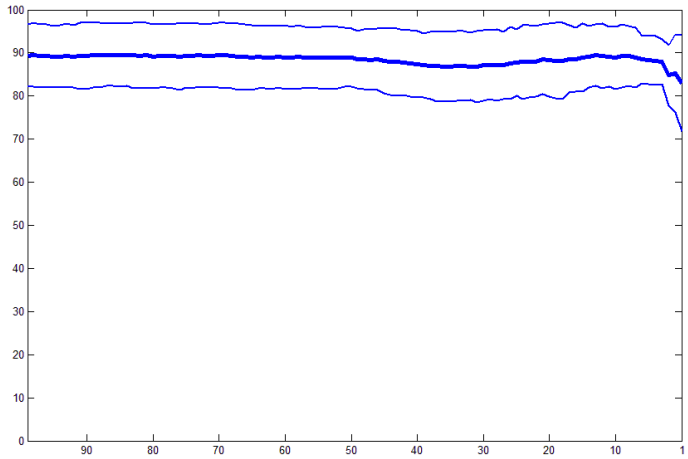
Διάγραμμα 85:1Δ



Διάγραμμα 86:1Ε



Διάγραμμα 87:1ΣΤ



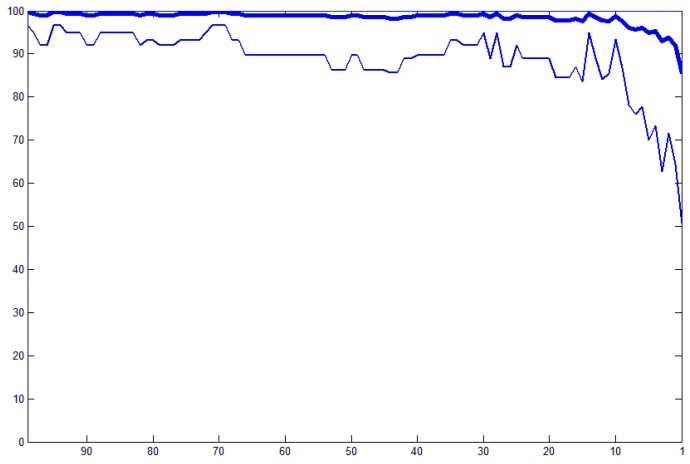
Διάγραμμα 88:1Ζ

Αυτό που μπορούμε να παρατηρήσουμε είναι ότι στο διάγραμμα 1Α επιτυγχάνονται μεγαλύτερες τιμές ακρίβειας απ' ότι στο διάγραμμα 1Β για τους ασθενείς του cross validation αν και έχουμε μεγαλύτερα διαστήματα εμπιστοσύνης λόγω μεγαλύτερης τυπικής απόκλισης.

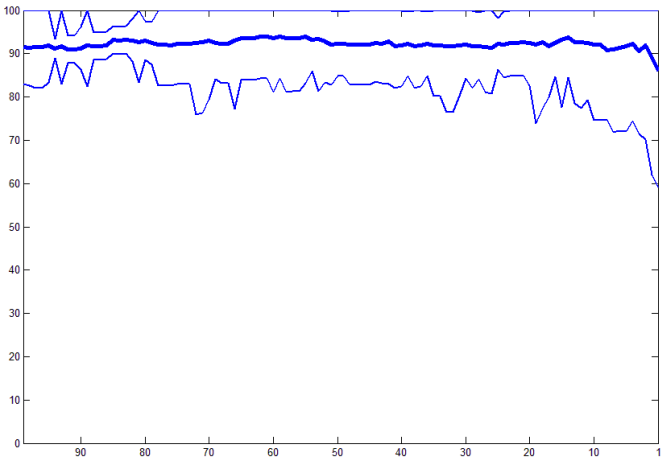
Σχετικά με τους ασθενείς του ανεξάρτητου test set, βλέπουμε ότι πετυχαίνουμε μεγαλύτερες τιμές ακρίβειας στο διάγραμμα 1Γ από αυτές του διαγράμματος 1Δ, αν και πάλι έχουμε μεγαλύτερες τυπικές αποκλίσεις, άρα και μεγαλύτερα διαστήματα εμπιστοσύνης στο διάγραμμα 1Γ.

Όσον αφορά τα διαγράμματα 1Ε,1ΣΤ,1Ζ που αναφέρονται στις μέσες ακρίβειες ανά τρέξιμο, βλέπουμε ότι η συγκεκριμένη μέθοδος πετυχαίνει καλύτερες ακρίβειες στα τρεξίματα του cross validation. Η απόδοση που πετυχαίνει στο ανεξάρτητο test set είναι μικρότερη με αποτέλεσμα να μειώνεται και η απόδοση της μεθόδου στα συνολικά τρεξίματα σε σχέση με την απόδοση που πετυχαίνει ξεχωριστά στα τρεξίματα του cross validation.

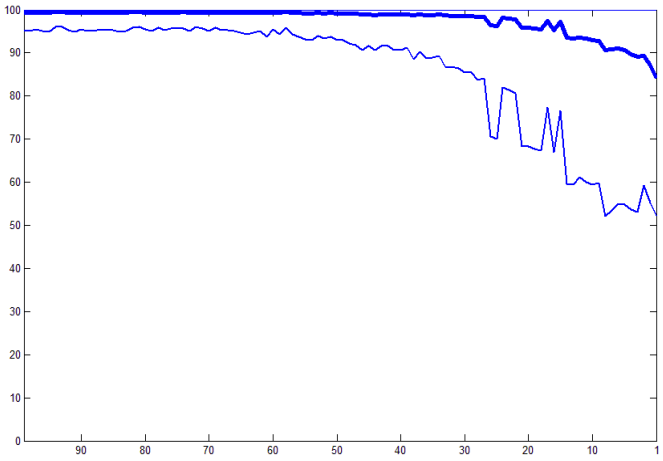
Επιπλέον, τα μικρότερα διαστήματα εμπιστοσύνης τα παρατηρούμε στο διάγραμμα 1Ζ όπου υπάρχουν και οι μικρότερες τυπικές αποκλίσεις. Στη συνέχεια, έχουμε την μικρότερη τυπική απόκλιση στα τρεξίματα του ανεξάρτητου test set (διάγραμμα 1ΣΤ) και το διάγραμμα 1Ε έχει τα μεγαλύτερα διαστήματα εμπιστοσύνης μεταξύ των διαγραμμάτων που απεικονίζουν μέσες τιμές ακρίβειας των τρεξιμάτων της μεθόδου.



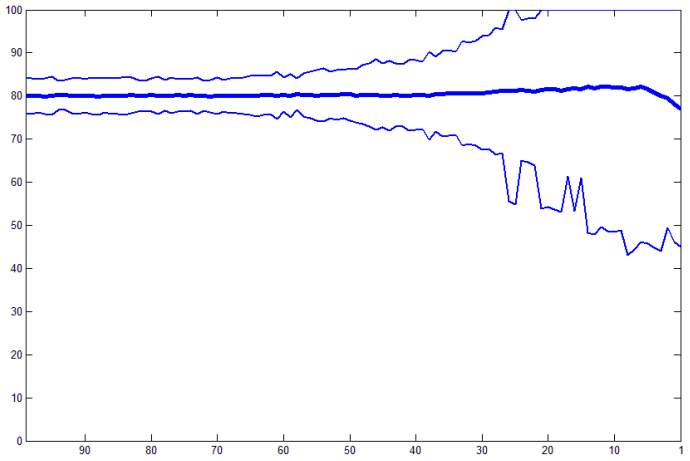
Διάγραμμα 89:2Α



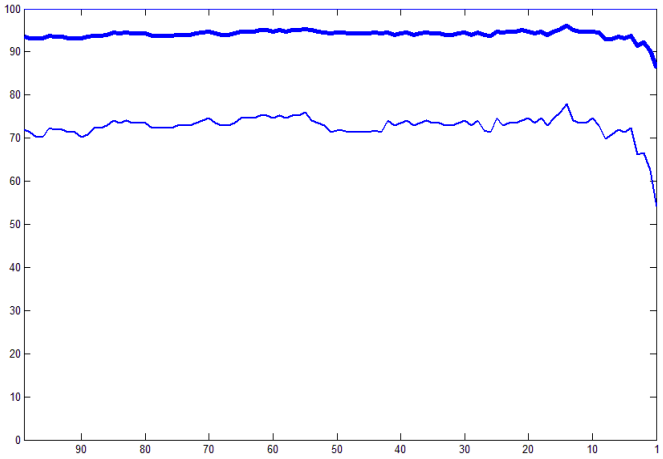
Διάγραμμα 90:2Β



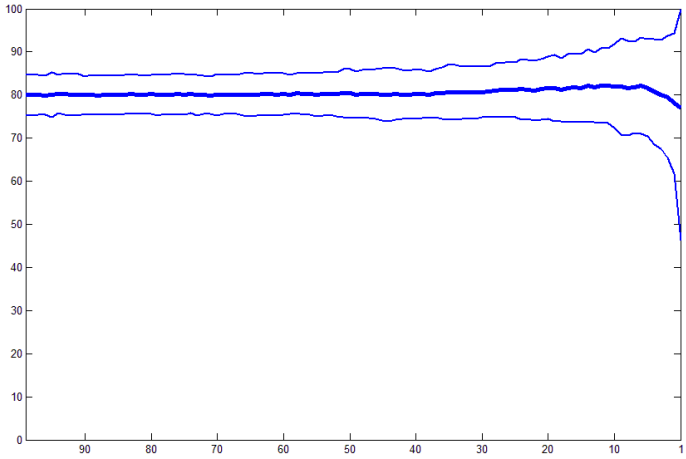
Διάγραμμα 91:2Γ



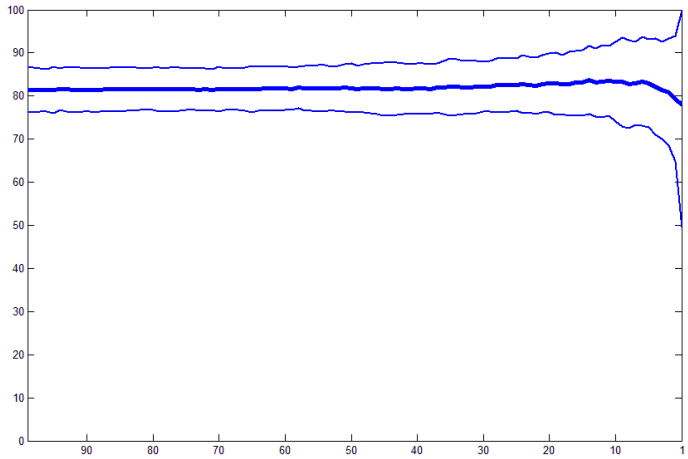
Διάγραμμα 92:2Δ



Διάγραμμα 93:2Ε



Διάγραμμα 94:2ΣΤ



Διάγραμμα 95:2Ζ

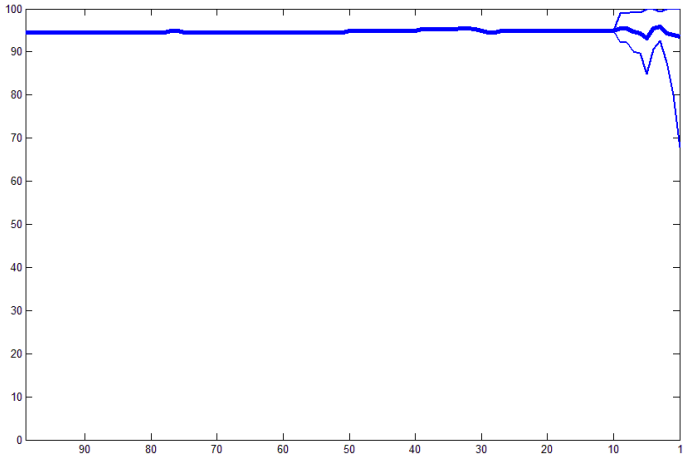
Αυτό που μπορούμε να παρατηρήσουμε είναι ότι στο διάγραμμα 2Α επιτυγχάνονται μεγαλύτερες τιμές ακρίβειας απ' ότι στο διάγραμμα 2Β για τους ασθενείς του cross validation και έχουμε μεγαλύτερα διαστήματα εμπιστοσύνης στο διάγραμμα 2Β λόγω μεγαλύτερης τυπικής απόκλισης από αυτήν του 2Α.

Σχετικά με τους ασθενείς του ανεξάρτητου test set, βλέπουμε ότι πετυχαίνουμε μεγαλύτερες τιμές ακρίβειας στο διάγραμμα 2Γ από αυτές του διαγράμματος 2Δ αλλά σε αυτήν την μέθοδο έχουμε σχεδόν ίδια διαστήματα εμπιστοσύνης για τα δύο διαγράμματα, καθώς έχουμε σχεδόν ίδιες τυπικές αποκλίσεις.

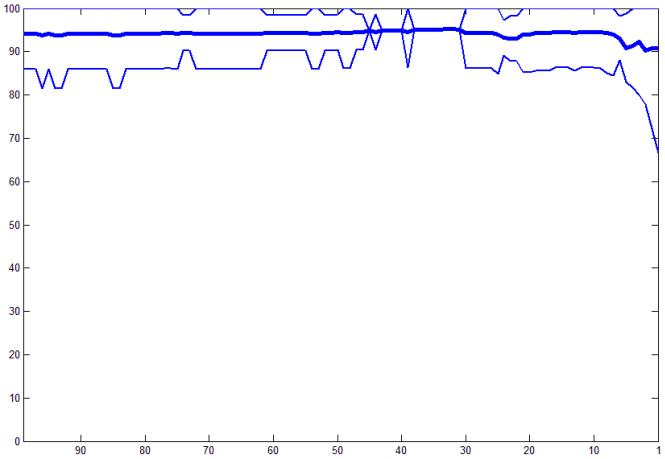
Όσον αφορά τα διαγράμματα 2Ε, 2ΣΤ, 2Ζ που αναφέρονται στις μέσες ακρίβειες ανά τρέξιμο, βλέπουμε ότι η συγκεκριμένη μέθοδος πετυχαίνει καλύτερες ακρίβειες στα τρεξίματα του cross validation. Η απόδοση που πετυχαίνει στο ανεξάρτητο test set είναι μικρότερη με αποτέλεσμα να μειώνεται και η απόδοση της μεθόδου στα συνολικά τρεξίματα σε σχέση με την απόδοση που πετυχαίνει ξεχωριστά στα τρεξίματα του cross validation.

Επιπλέον, τα μικρότερα διαστήματα εμπιστοσύνης τα παρατηρούμε στο διάγραμμα 2Ζ όπου υπάρχουν και οι μικρότερες τυπικές αποκλίσεις. Στη συνέχεια με μικρή διαφορά, έχουμε την μικρότερη τυπική απόκλιση στα τρεξίματα του ανεξάρτητου test set (διάγραμμα 2ΣΤ) και το διάγραμμα 2Ε έχει τα μεγαλύτερα διαστήματα εμπιστοσύνης μεταξύ των διαγραμμάτων που απεικονίζουν μέσες τιμές ακρίβειας των τρεξιμάτων της μεθόδου.

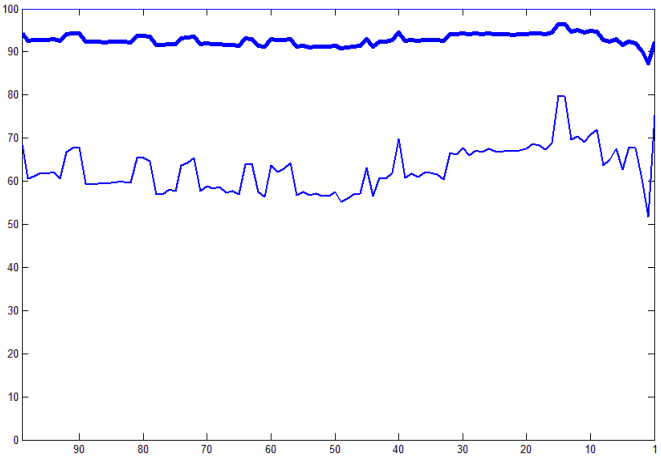
3. ΜΕΘΟΔΟΣ RFE-FSVs-4DK:



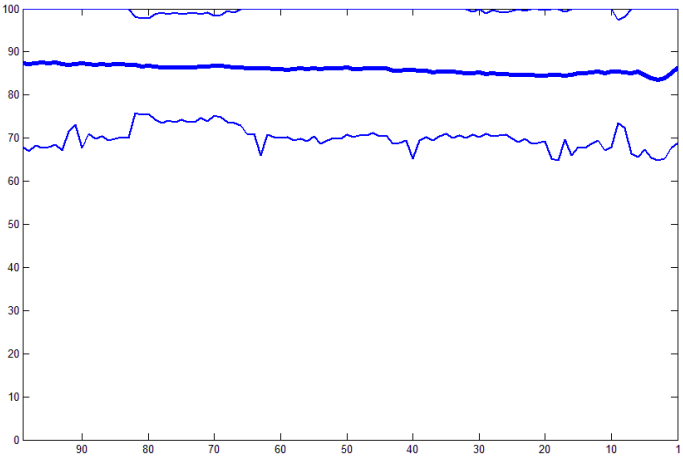
Διάγραμμα 96:3Α



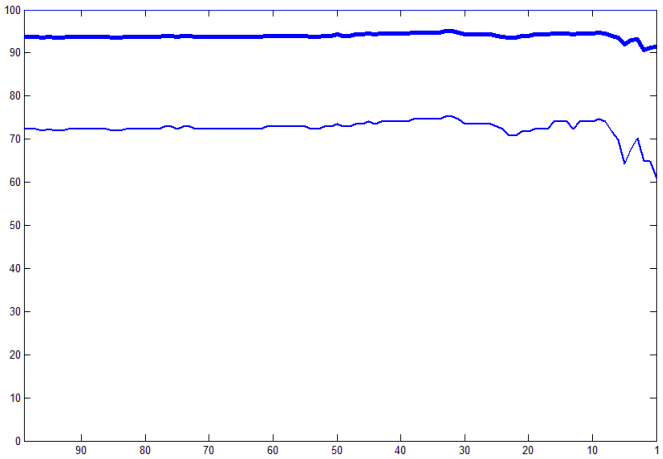
Διάγραμμα 97:3Β



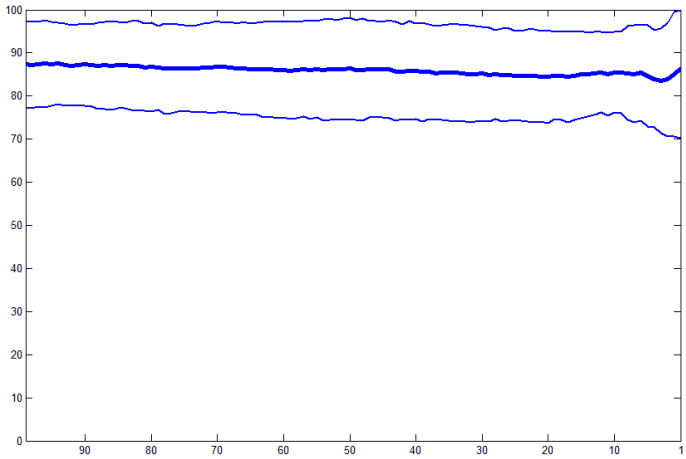
Διάγραμμα 98:3Γ



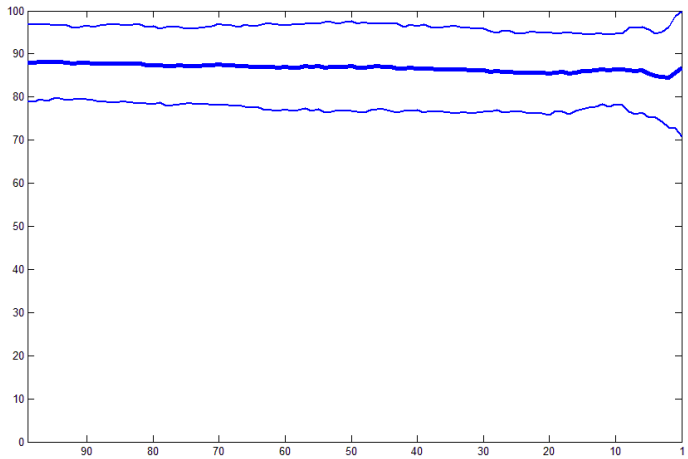
Διάγραμμα 99:3Δ



Διάγραμμα 100:3Ε



Διάγραμμα 101:3ΣΤ



Διάγραμμα 102:3Ζ

Αυτό που μπορούμε να παρατηρήσουμε είναι ότι στο διάγραμμα 3Α επιτυγχάνονται μεγαλύτερες τιμές ακρίβειας απ' ότι στο διάγραμμα 3Β για τους ασθενείς του cross validation αν και έχουμε μεγαλύτερα διαστήματα εμπιστοσύνης λόγω μεγαλύτερης τυπικής απόκλισης.

Σχετικά με τους ασθενείς του ανεξάρτητου test set, βλέπουμε ότι πετυχαίνουμε μεγαλύτερες τιμές ακρίβειας στο διάγραμμα 3Γ από αυτές του διαγράμματος 3Δ, αν και πάλι έχουμε μεγαλύτερες τυπικές αποκλίσεις, άρα και μεγαλύτερα διαστήματα εμπιστοσύνης στο διάγραμμα 3Γ.

Όσον αφορά τα διαγράμματα 3Ε, 3ΣΤ, 3Ζ που αναφέρονται στις μέσες ακρίβειες ανά τρέξιμο, βλέπουμε ότι η συγκεκριμένη μέθοδος πετυχαίνει καλύτερες ακρίβειες στα τρεξίματα του cross validation. Η απόδοση που πετυχαίνει στο ανεξάρτητο test set είναι μικρότερη με αποτέλεσμα να μειώνεται και η απόδοση της μεθόδου στα συνολικά τρεξίματα σε σχέση με την απόδοση που πετυχαίνει ξεχωριστά στα τρεξίματα του cross validation.

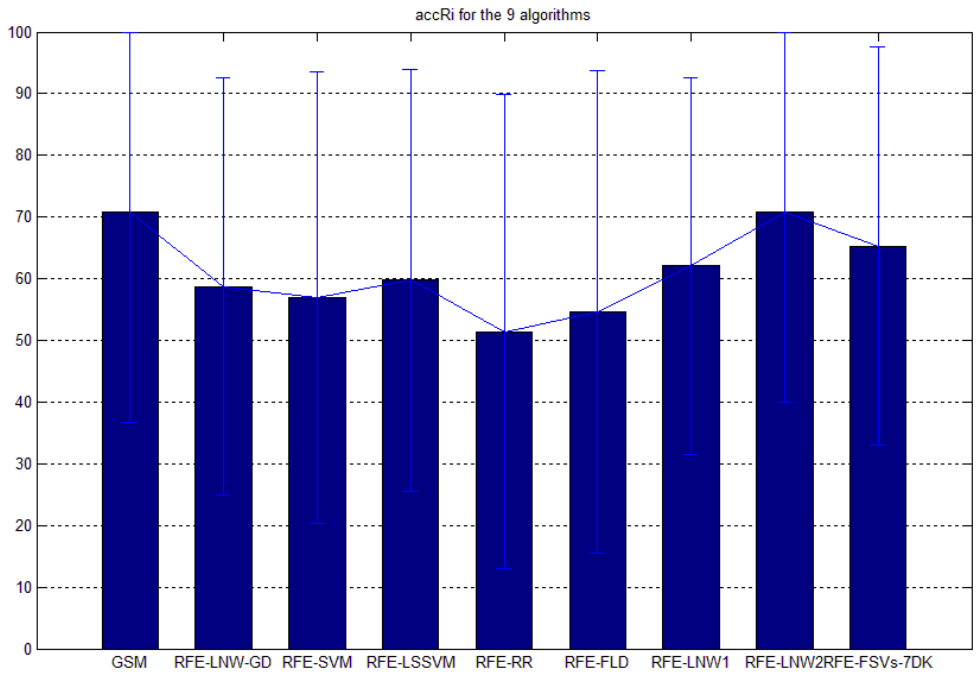
Επιπλέον, τα μικρότερα διαστήματα εμπιστοσύνης τα παρατηρούμε στο διάγραμμα 3Ζ όπου υπάρχουν και οι μικρότερες τυπικές αποκλίσεις. Στη συνέχεια, έχουμε την μικρότερη τυπική απόκλιση στα τρεξίματα του ανεξάρτητου test set (διάγραμμα 3ΣΤ) και το διάγραμμα 3Ε έχει τα μεγαλύτερα διαστήματα εμπιστοσύνης μεταξύ των διαγραμμάτων που απεικονίζουν μέσες τιμές ακρίβειας των τρεξιμάτων της μεθόδου.

4.3.Συζήτηση για τα δεδομένα που σχετίζονται με τον καρκίνο του μαστού

Από τα πειράματα που πραγματοποιήσαμε για τα δεδομένα που σχετίζονται με τον καρκίνο του μαστού, είδαμε μια διαφοροποίηση ανάμεσα στα αποτελέσματα που προέκυψαν από το cross validation και σε αυτά που προέκυψαν από το ανεξάρτητο test set.

Μερικά από τα παραπάνω αποτελέσματα συνοψίζονται καλύτερα στα παρακάτω διαγράμματα, όπου μπορεί κανείς να κάνει μια πιο άμεση σύγκριση των 9 μεθόδων.

Σε κάθε μέτρο ακρίβειας φαίνεται με λεπτή γαλάζια γραμμή και το αντίστοιχο όριο εμπιστοσύνης του, το οποίο έχει υπολογιστεί προσθέτοντας και αφαιρώντας αντίστοιχα την ποσότητα 1.96 * τυπική απόκλιση της ακρίβειας από την ακρίβεια αυτή.

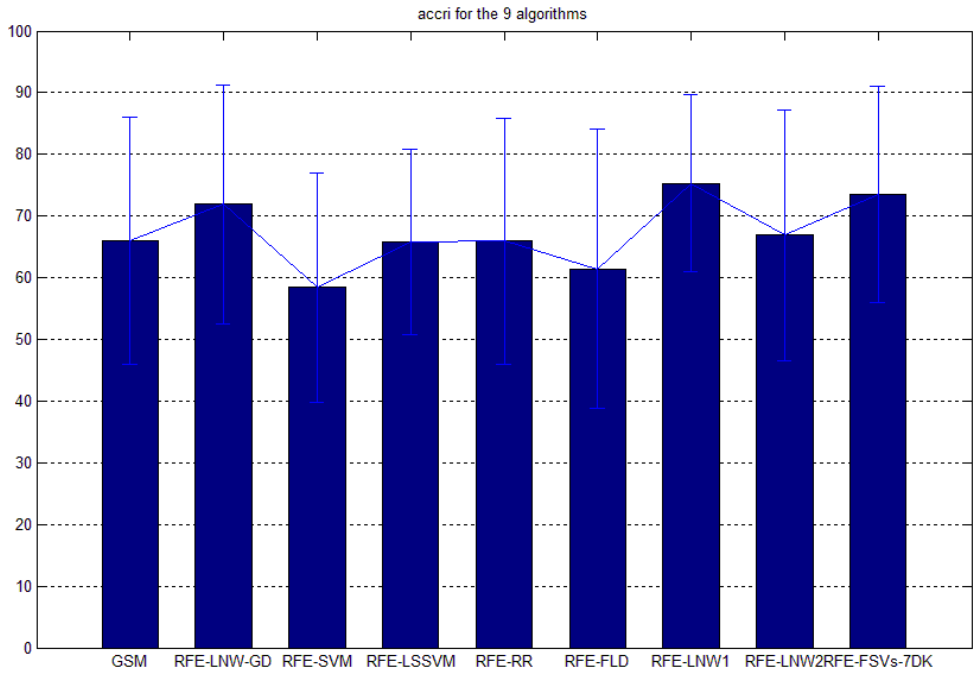


Διάγραμμα 103: Μέση accRi ανά τρέξιμο για τις 9 μεθόδους

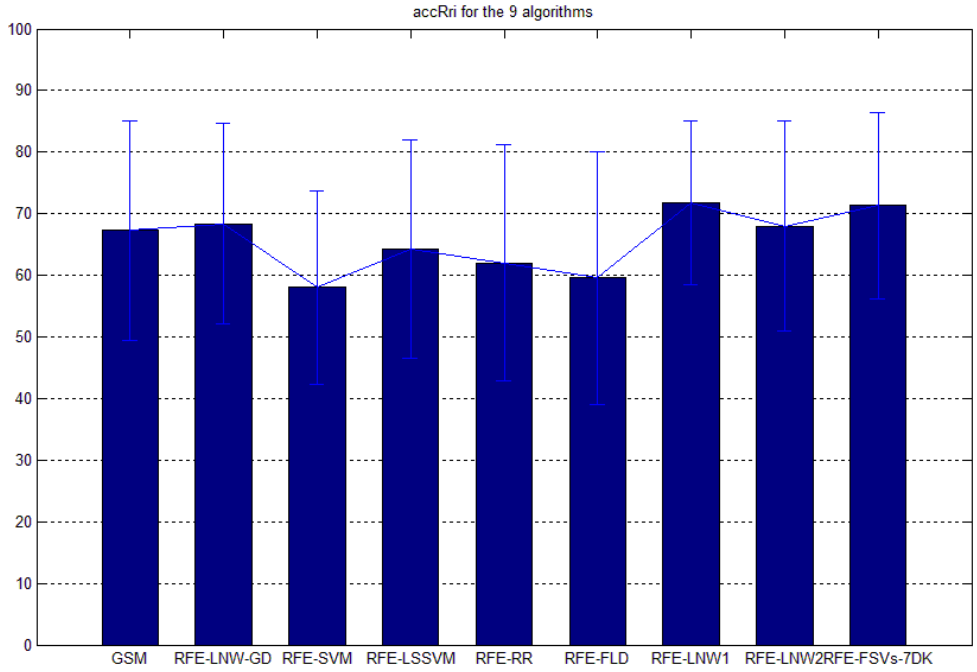
Για τα δεδομένα που προκύπτουν από το cross validation, κάνουμε μία σύγκριση μεταξύ των wrapper μεθόδων και της μεθόδου φίλτρου GSM.

Βλέπουμε ότι η μέθοδος φίλτρου έχει καλύτερη απόδοση από τις υπόλοιπες μεθόδους.

Στη συνέχεια, συγκρίνουμε τις μεθόδους που συνδυάζουν χαρακτηριστικά των wrapper και μεθόδων φίλτρου με αυτήν που είναι καθαρά μέθοδος φίλτρου. Βλέπουμε παρόμοια συμπεριφορά της GSM με την RFE-LNW2, η οποία έχει χαμηλό learning rate. Καταλαβαίνουμε ότι για χαμηλό learning rate, η συνδυαστική μέθοδος RFE-LNW, προσεγγίζει την μέθοδο φίλτρου. Οι υπόλοιπες όμως συνδυαστικές μέθοδοι δεν έχουν τόσο καλή απόδοση όσο η μέθοδος φίλτρου για την συγκεκριμένη διαδικασία του cross validation.



Διάγραμμα 104: Μέση accri ανά τρέξιμο για τις 9 μεθόδους.



Διάγραμμα 105: Μέση accRri ανά τρέξιμο για τις 9 μεθόδους.

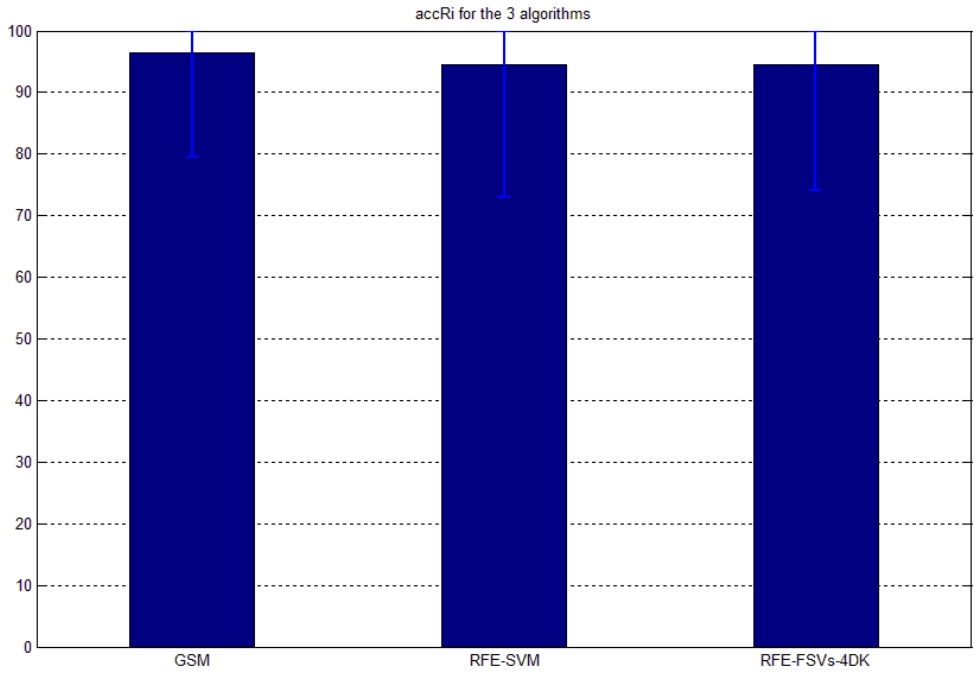
Αρχικά, αν κάνουμε μία σύγκριση μεταξύ των καθαρά wrapper μεθόδων και των μεθόδων φίλτρου(που για την δική μας περίπτωση είναι η μέθοδος GSM) για τα δεδομένα του ανεξάρτητου test set, παρατηρούμε μία σαφώς καλύτερη απόδοση της μεθόδου φίλτρου GSM σε σχέση με τις wrapper μεθόδους. Μοναδική εξαίρεση μπορούμε να πούμε ότι είναι η wrapper μέθοδος LNW-GD, που μας οδηγεί στο συμπέρασμα ότι ένα γραμμικό νευρωνικό μπορεί να χρησιμοποιηθεί αποτελεσματικά ως επιλογέας χαρακτηριστικών (γονιδίων).

Στη συνέχεια συγκρίνουμε τις μεθόδους που συνδυάζουν τα wrapper με τα χαρακτηριστικά φίλτρου με την καθαρά μέθοδο φίλτρου GSM. Για το ίδιο σύνολο ασθενών με παραπάνω, παρατηρούμε μια σαφώς καλύτερη απόδοση των συνδυαστικών μεθόδων σε σχέση με την καθαρή μέθοδο φίλτρου. Προφανώς, οι συνδυαστικές μέθοδοι είναι αποτελεσματικότερες και από τις wrapper με βάση τα παραπάνω.

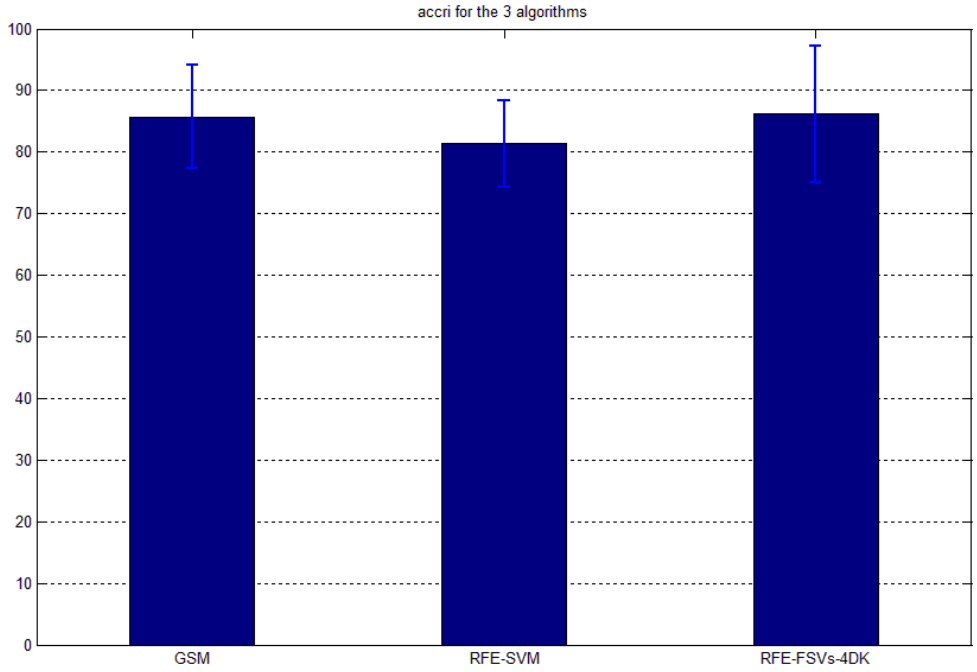
4.4.Συζήτηση για τα δεδομένα που σχετίζονται με τη λευχαιμία

Στα δεδομένα που σχετίζονται με την λευχαιμία παρατηρήσαμε την ίδια συμπεριφορά στα μέτρα απόδοσης των τριών μεθόδων επιλογής χαρακτηριστικών με τα δεδομένα που σχετίζονται με τον καρκίνο του μαστού.

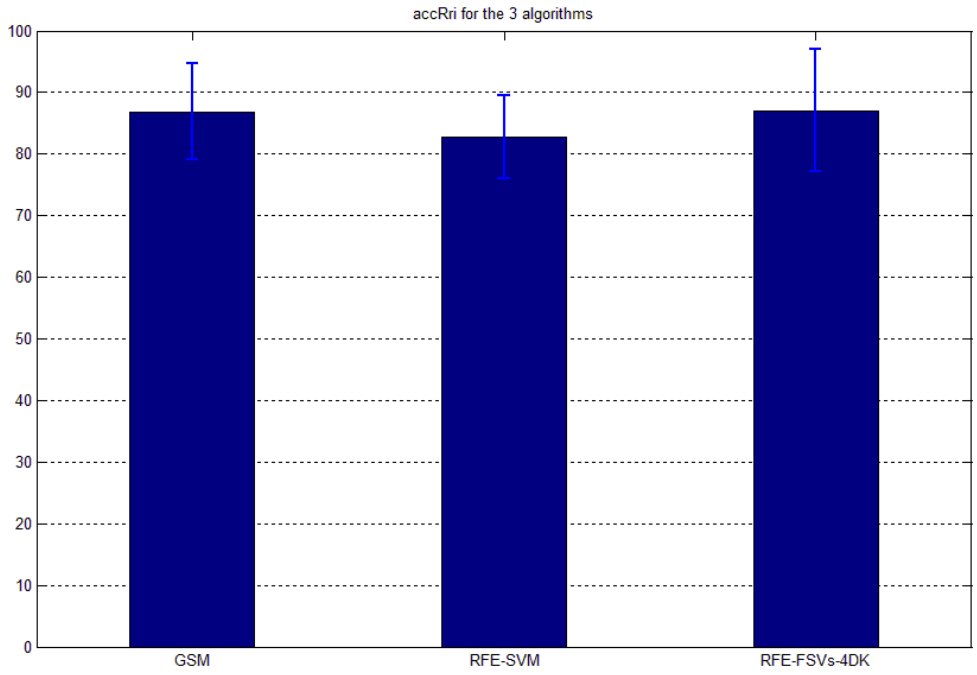
Αυτό φαίνεται και από τα παρακάτω διαγράμματα στα οποία μπορεί αν γίνει ευκολότερη σύγκριση μεταξύ των 3 μεθόδων που εφαρμόσαμε και σε αυτό το σύνολο δεδομένων:



Διάγραμμα 106: Μέση accRi ανά τρέξιμο για τις 3 μεθόδους



Διάγραμμα 107: Μέση accri ανά τρέξιμο για τις 3 μεθόδους.



Διάγραμμα 108: Μέση accRri ανά τρέξιμο για τις 3 μεθόδους.

Και εδώ, παρατηρήθηκε σαφώς καλύτερη απόδοση της συνδυαστικής μεθόδου (RFE-FSVs-4DK) στα δεδομένα του ανεξάρτητου test set, σε σχέση

με την μέθοδο φίλτρου (GSM) που εμφανίζει με τη σειρά της καλύτερη απόδοση σε σχέση με την wrapper μέθοδο(RFE-SVM).

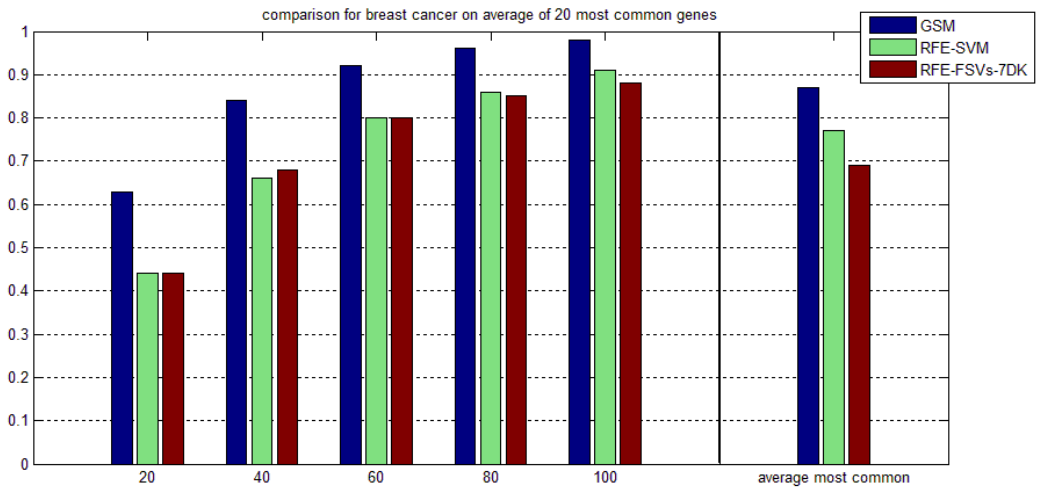
Από την άλλη μεριά, επεξεργαζόμενοι τα δεδομένα μέσω της διαδικασίας του cross validation παρατηρήσαμε την υπεροχή της μεθόδου φίλτρου σε σχέση με την συνδυαστική, η οποία πέτυχε όμως απόδοση καλύτερη από την wrapper μέθοδο.

4.5. Συζήτηση για την επικάλυψη γονιδίων

Ακόμα, υπολογίσαμε για κάθε μέθοδο από τις 3 που χρησιμοποιήθηκαν και για τα δεδομένα λευχαιμίας και για τα δεδομένα καρκίνου του μαστού,

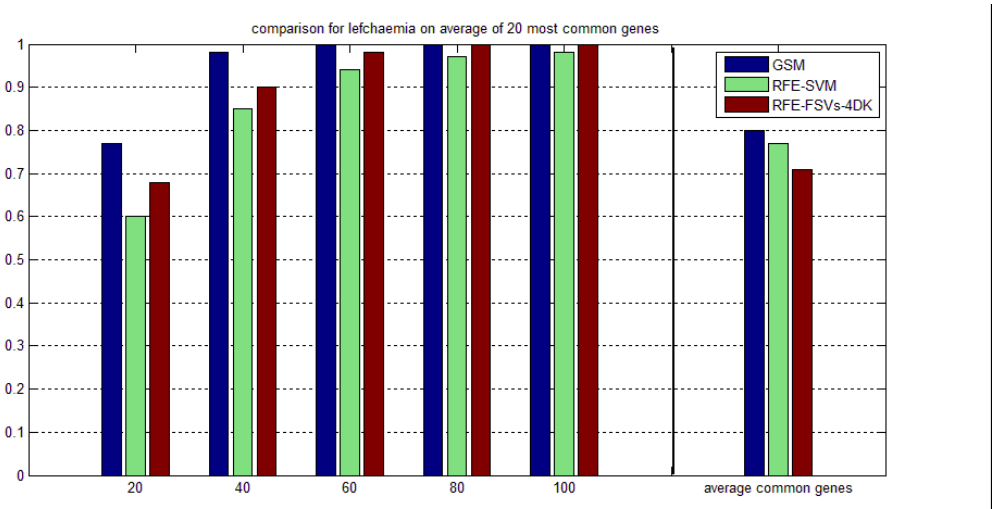
Σταματήσαμε την εκτέλεση κάθε μεθόδου στα 20,40,60,80 και 100 γονίδια και βρήκαμε την μέση συχνότητα των 20 πιο συχνά εμφανιζόμενων γονιδίων. Από κάθε ένα από αυτά τα πέντε σταματήματα της κάθε μεθόδου, πήραμε τα 20 γονίδια με την μεγαλύτερη συχνότητα και βρήκαμε πόσες φορές εμφανίζονται στα πέντε σταματήματα. Υπολογίσαμε έτσι την συχνότητά τους και βρήκαμε την μέση τιμή της η οποία εκφράζει την αλληλοεπικάλυψη των συνόλων των 20 γονιδίων για την κάθε μέθοδο.

Τα αποτελέσματα αυτής της διαδικασίας φαίνονται στο παρακάτω διάγραμμα για τα δεδομένα καρκίνου του μαστού:



Διάγραμμα 109: Μέση συχνότητα των 20 πιο συχνά εμφανιζόμενων γονιδίων σε σταματήματα των μεθόδων σε διαφορετικούς αριθμούς γονιδίων, καθώς και αλληλοεπικάλυψη του συνόλου των 20 αυτών γονιδίων για τα δεδομένα καρκίνου του μαστού.

Καθώς και παρακάτω για τα δεδομένα της λευχαιμίας:



Διάγραμμα 110: Μέση συχνότητα των 20 πιο συχνά εμφανιζόμενων γονιδίων σε σταματήματα των μεθόδων σε διαφορετικούς αριθμούς γονιδίων, καθώς και αλληλοεπικάλυψη του συνόλου των 20 αυτών γονιδίων για τα δεδομένα λευχαιμίας.

Παρατηρούμε λοιπόν, ότι και στα δύο σύνολα ασθενών, η μέθοδος φίλτρου εμφανίζει μεγαλύτερο ποσοστό αλληλοεπικάλυψης στα σύνολα των πιο

κοινών γονιδίων, ακολουθεί η wrapper μέθοδος και στη συνέχεια η συνδυαστική μέθοδος.

Αν εξετάσουμε ξεχωριστά το κάθε σταμάτημα, βλέπουμε ότι και στις τρεις μεθόδους, έχουμε μεγαλύτερη μέση συχνότητα γονιδίων στο σύνολο που σχετίζεται με την λευχαιμία παρά σε αυτό που σχετίζεται με τον καρκίνο του μαστού.

Συγκρίνοντας τώρα τις τρεις μεθόδους, μπορούμε να πούμε ότι η μέθοδος φίλτρου έχει ένα σαφές προβάδισμα και στα δύο σύνολα δεδομένων, ενώ στο μεν σύνολο καρκίνου του μαστού προηγείται μπορούμε να πούμε ελαφρώς η wrapper μέθοδος, στο δε σύνολο της λευχαιμίας προηγείται αισθητά η συνδυαστική μέθοδος.

Κεφάλαιο 5: Συμπεράσματα και μελλοντική εργασία

5.1.Συμπεράσματα

Σαν ένα γενικότερο συμπέρασμα μπορούμε να πούμε ότι οι συνδυαστικές μέθοδοι είναι καλύτερες όσον αφορά την απόδοσή τους σε σχέση με τις απλές wrapper μεθόδους και στις περιπτώσεις είναι καλύτερες ή έστω προσεγγίζουν κατά πολύ την απόδοση των μεθόδων φίλτρου.

Σαν εξαίρεση των παραπάνω μπορούμε να αναφέρουμε την μέτρηση της απόδοσης ανά ασθενή των μεθόδων κατά την διαδικασία του cross validation όπου στην περίπτωση των δεδομένων καρκίνου του μαστού παρατηρούμε μία καλύτερη απόδοση των συνδυαστικών μεθόδων σε σχέση με την μέθοδο φίλτρου, πράγμα που δεν συμβαίνει με τα υπόλοιπα μέτρα απόδοσης που σχετίζονται με την διαδικασία του cross validation.

Στα δεδομένα λευχαιμίας, παρατηρούμε για το παραπάνω μέτρο απόδοσης καλύτερη απόδοση από την πλευρά των καθαρά wrapper μεθόδων σε σχέση με την μέθοδο φίλτρου που έχει με τη σειρά της καλύτερη απόδοση από τις συνδυαστικές μεθόδους.

Όσον αφορά τα διαστήματα εμπιστοσύνης, μπορούμε να πούμε ότι είναι γενικά μικρότερα σε μέτρα απόδοσης που σχετίζονται με το ανεξάρτητο test set όπως είναι η απόδοση ανά ασθενή ή ανά τρέξιμο της μεθόδου και μεγαλύτερα για μέτρα που σχετίζονται με την διαδικασία του cross validation. Ένας λόγος για τον οποίο συμβαίνει αυτό είναι ότι τα δεδομένα του ανεξάρτητου test set συμμετέχουν περισσότερες φορές στα τρεξίματα της μεθόδου και εμφανίζουν επομένως μεγαλύτερη συχνότητα.

Αξίζει να αναφερθεί, ότι τα διαστήματα εμπιστοσύνης της απόδοσης ανά τρέξιμο για όλους τους ασθενείς συνολικά (και για αυτούς του cross validation αλλά και για αυτούς του ανεξάρτητου test set), είναι μικρότερα από τα

αντίστοιχα για την αποδόσεις ανά τρέξιμο για το ένα ή το άλλο σύνολο ασθενών ξεχωριστά.

5.2.Μελλοντική εργασία

Τα σχετικά μεγάλα διαστήματα εμπιστοσύνης που προκύπτουν σε μέτρα απόδοσης που αναφέρονται στους ασθενείς που επεξεργαζόμαστε μέσω της διαδικασίας του cross-validation είναι ένα ζήτημα που μπορεί μελλοντικά να επιλυθεί με την πραγματοποίηση περισσότερων εφαρμογών ανά μέθοδο. Έτσι, κάθε ασθενής θα συμμετέχει περισσότερες φορές στα test set που προκύπτουν από το cross-validation και το αποτέλεσμα θα είναι πιο αξιόπιστο και αντιπροσωπευτικό.

Θα μπορούσαμε για παράδειγμα να πραγματοποιήσουμε 500 ή 1000 εφαρμογές για κάθε μια από τις 9 μεθόδους, αντί για 100 που πραγματοποιήσαμε σε αυτήν την εργασία.

Μία ακόμα μελλοντική προσπάθεια, θα μπορούσε να είναι η εφαρμογή των ίδιων μεθόδων σε πολλά διαφορετικά σύνολα ασθενών. Εδώ δοκιμάσαμε τις 9 μεθόδους στο σύνολο των ασθενών του καρκίνου του μαστού και 3 από αυτές τις μεθόδους δοκιμάστηκαν επίσης και στο σύνολο ασθενών με λευχαιμία. Μελλοντικά λοιπόν, θα μπορούσαμε να κάνουμε την σύγκριση των μέτρων απόδοσης των μεθόδων όπως αυτά προέκυψαν από τα διάφορα σύνολα ασθενών και να βγάλουμε χρήσιμα συμπεράσματα για αυτές, σχετικά με την αξιοπιστία και την σταθερότητα τους.

Αναφορές:

[1] University of California Santa Cruz Genome Bioinformatics, Human Genome Working Draft, http://genome.ucsc.edu, (2001)

[2] R. T. Golub, K. D. Slonim, P. Tamayo, C. Huard, et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, 286 (1999) 531-536.

[3] R. Scalkoff, Pattern Recognition: Statistical, Structural and Neural Approaches, (John Wiley & Sons, inc., 1992).

[4] M. Blazadonakis and M. Zervakis, The Linear Neuron as Marker Selector and Clinical Predictor in Cancer Gene Analysis, Computer Methods and Programs in Biomedicine.

[5] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using Support vector machines, machine learning, 36 (2002) 389-422.

[6] Athanasios Papoulis, «Πιθανότητες, Τυχαίες μεταβλητές και Στοχαστικές διαδικασίες»,Κεφάλαιο 9.2:Εκτίμηση Παραμέτρων,3η έκδοση

[7] Boutsikas M.V.(2003), Σημειώσεις Στατιστικής ΙΙΙ, Τμήμα Οικονομικής Επιστήμης, Πανεπιστήμιο Πειραιώς. «5.Διαστήματα εμπιστοσύνης»

[8] Boutsikas M.V. (2005) Σημειώσεις μαθήματος:«Μέθοδοι Προσομοίωσης και Στατιστικές Υπολογιστικές Τεχνικές», “Εισαγωγή στη μέθοδο Bootstrap”

[9] Boutsikas M.V. (2004), Σημειώσεις μαθήματος «Στατιστικά Προγράμματα» Τμήμα Στατ. & Ασφ. Επιστήμης, Πανεπιστήμιο Πειραιώς, “Έλεγχοι καλής προσαρμογής”

[10] Calculating Confidence Intervals for Prediction Error in Microarray Classification Using Resampling. Wenyu Jiang*§, Sudhir Varma§ and Richard Simon§ Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Rockville MD 20892-7434, USA

[11] Ricardo Gutierrez-Osuna, Wright State University, “Statistical Pattern Recognition”, NOSE 2nd Summer school, LIoret de Mar, Spain, October 2-5, 2000

[12] Statistics Section in the Department of Mathematics at Imperial College Statistical theory II, «THE GLIVENKO-CANTELLI LEMMA»

Βιβλιογραφία:

1. Efron, B. (1987) Better bootstrap confidence intervals. (with discussion.) J. Am. Stat.Assoc., 82, 171-200.

2. Efron, B. and Tibshirani, R. (1998) An Introduction to the Bootstrap. Chapman and Hall.

3. Fu, W. Carroll, R. J. and Wang, S. (2005) Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics, 21, 1979-1986.

4. Jiang, W. and Simon, R. (2006) A comparison of bootstrap methods and an adjusted bootstrap approach for estimating prediction error in microarray classification. Technical Report. Biometric Research Branch, Division of Cancer Treatment and Diagnosis, NCI.

5. Lachenbruch P.A. and Mickey M. R. (1968) Estimation of error rates in discriminant analysis. Technometrics 10:1-11.

6. Martin JK, Hirschberg DS. (1996). Small sample statistics for classification error rates II:Confidence intervals and significance tests. Technical Report, ICS-TR-96-22.

7. Michiels S, Koscielny S and Hill C. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy.Lancet, 365, 488-492.

8. Molinaro, A. M., Simon, R. and Pfeiffer, R. M. (2005) Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21, 3301-3307.

9. Radmacher, M. D., McShane, L. M. and Simon, R. (2002) A paradigm for class prediction using gene expression profiles. J. Comput. Biology, 9, 505-511.