Εργασία αναγνώρισης προτύπων

Code Readability Model

Μουτζόγλου Δημήτρης 8319

Dimitrpm@auth.gr

Δάμη Υπατία 8606

Ypatiapd@auth.gr

Περιγραφή του προβλήματος:

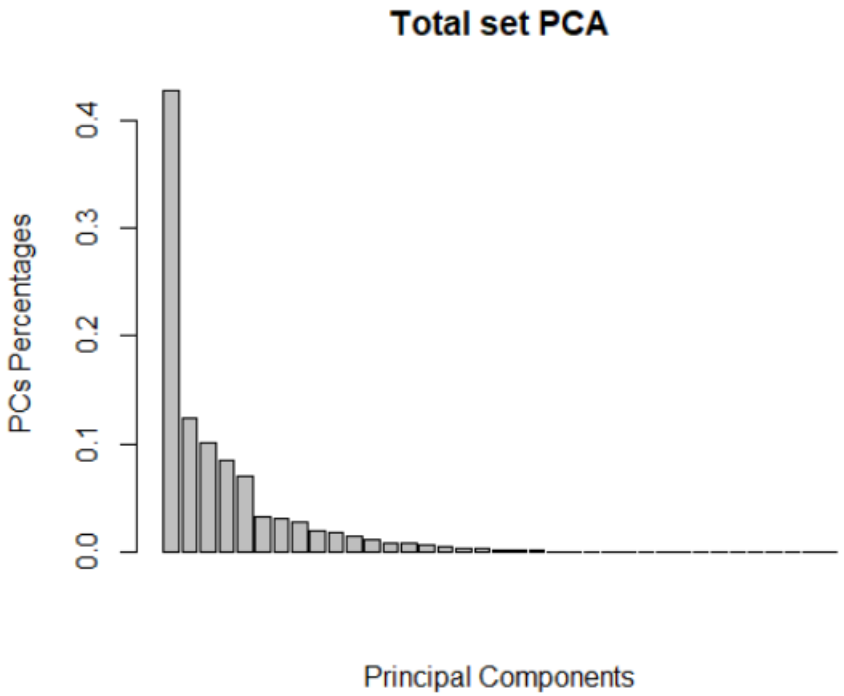
Βασικό ζητούμενο της εργασίας αποτελεί η εξέταση της αναγνωσιμότητας μιας μεθόδου δεδομένου ενός σετ μετρικών λογισμικού που το περιγράφουν και ενός σετ παραβιάσεων που έχουν ανιχνευτεί μέσω του PMD. Οι μετρικές περιγράφουν με ακρίβεια χαρακτηριστικά των μεθόδων που αφορούν την πολυπλοκότητα τους , το coupling , την πυκνότητα των σχολίων στον κώδικα , την συντηρισιμότητα και άλλους δείκτες που η συσχέτιση τους με την ανίχνευση της αναγνώσιμοτητς του κώδικα δεν είναι ούτε εγγυημένη ούτε προφανής. Παράλληλα , και η σχέση των παραβιάσεων που έχουν ανιχνευτεί από τον PMD είναι αδιόρατη και απαιτείται ανάλυση για την συσχέτιση τους με το αντικείμενο της εργασίας. Σκοπός μας , η δημιουργία ενός μοντέλου που δεδομένων των μετρικών και του μεγέθους μιας μεθόδου μπορεί να κατηγοριοποιήσει ως προς την αναγνωσιμότητα .

Πειραματισμοί:

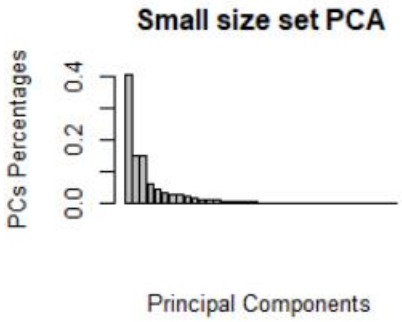
Αρχική μας λανθασμένη προσέγγιση ήταν να υπολογίσουμε το correlation matrix μεταξύ των μετρικών που ανήκουν στην ίδια κατηγορία προκειμένου να συμπεράνουμε ποιες μετρικές περιέχουν παρόμοια πληροφορία και μπορούν να συσχετιστούν μεταξύ τους. Στη συνέχεια, προσπαθήσαμε να βρούμε μία συσχέτιση μεταξύ μετρικών που ανήκουν σε διαφορετικές κατηγορίες , να τους αποδώσουμε αν υπάρχει μία φυσική σημασία , και στη συνέχεια να τις απεικονίσουμε σε δύο διαστάσεις προκειμένου να τις ομαδοποιήσουμε με την εφαρμογή αλγορίθμων (k-means, dbscan, gmms). Σε αντίθεση με τις στοχεύσεις μας , η φύση των δεδομένων και οι τιμές τους πριν και μετά την εφαρμογή pre-processing ήταν αδύνατον να ομαδοποιηθούν σε clusters με σαφή φυσική σημασία.

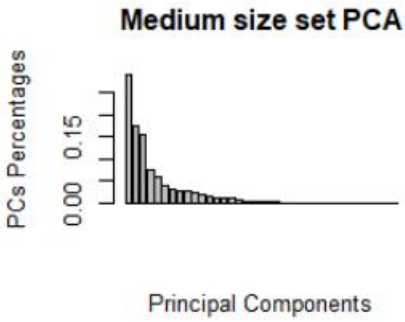
Εξαγωγή κριτηρίου διαχωρισμού dataset και εξαγωγή μετρικών:

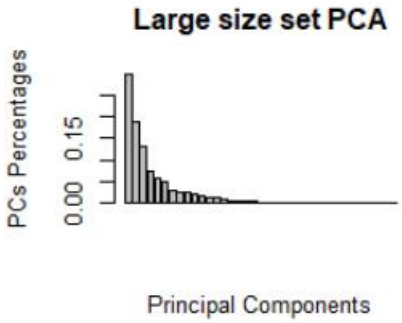
Η διαδικασία που ακολουθήσαμε φαίνεται στο script main.R όπου γίνεται και η φόρτωση των βιβλιοθηκών.Αρχικά , στο script *metrics_selection_PCA.R*, εφαρμόζουμε τον αλγόριθμο PCA για το σύνολο των δεδομένων, αφού έχουμε πρώτα απαλείψει όλες τις NA τιμές του dataset. Να σημειωθεί πως αρχικά έγινε η δοκιμή για την αντικατάσταση των NA τιμών με τη μέση τιμή της κάθε στήλης , αλλά απορρίφθηκε λόγω πόλωσης του dataset προς αυτές τις τιμές. Στην παρακάτω εικόνα φαίνεται η εφαρμογή του αλγορίθμου στο συνολικό dataset.



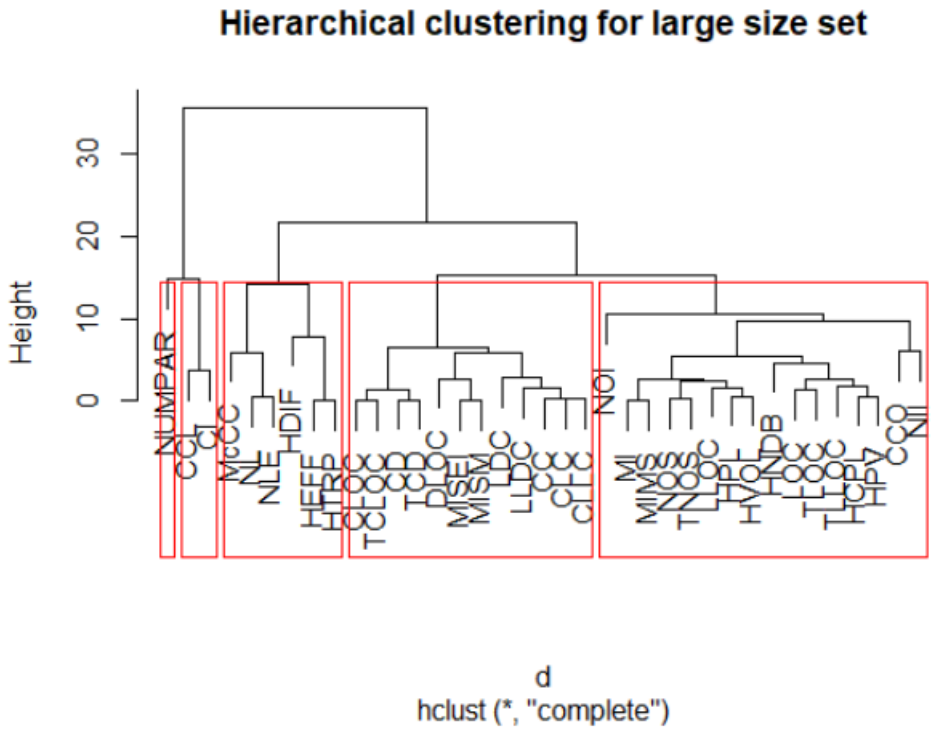
Με ένα for loop εμφανίζουμε τα ονόματα των μετρικών που έχουν τη μεγαλύτερη συνεισφορά στα 5 πρώτα principal components. Συμπεραίνουμε πως η μεταβλητή που συνεισφέρει περισσότερο στο πρώτο principal component , είναι η LLOC και αποφασίζουμε να χωρίσουμε το dataset με βάση την μεταβλητή αυτή (γραμμές κώδικα). Οι ομάδες που προκύπτουν από τον διαχωρισμό κατηγοριοποιούνται σε Far-Left Outliers,Small_Size,Medium_Size,Large_Size,Far-Right Outliers. Ο παραπάνω διαχωρισμός γίνεται με όρια που προέκυψαν από έρευνα στο ίντερνετ και την προσωπική μας εκτίμηση του τι αποτελεί μία μέθοδο με προφανή βαθμό readability ,όπως μέθοδοι με λιγότερες από 4 γραμμές κώδικα. Για το σύνολο far-Left Outliers κάνουμε την παραδοχή ότι εφόσον αποτελείται από μεθόδους από 1 έως 3 γραμμές κώδικα ,χαρακτηρίζεται απευθείας ως high-readable. Ωστόσο, για το σύνολο των Far-Right Outliers, θα πρέπει να γίνει ξεχωριστή ανάλυση. Έπειτα, στο ίδιο script, εφαρμόζουμε τον αλγορίθμου PCA για τα ξεχωριστά σύνολα δεδομένων που προέκυψαν προκειμένου να συμπεράνουμε ποιες μετρικές περιγράφουν τις περιεχόμενες μεθόδους με ικανοποιητική πληρότητα πληροφορίας.







Όπως φαίνεται και εκ πρώτης όψεως από τα παραπάνω ραβδογράμματα τα 5 βασικότερα Principal Components περιέχουν επαρκές ποσοστό πληροφορίας για την περιγραφή του συνολικού dataset. Υπολογίζοντας το ποσοστό απώλειας πληροφορίας πέρα από το 5ο principal component καταλήξαμε πως το ποσοστό πληροφορίας που περιέχεται στα 5 πρώτα principal components κρίνεται επαρκές. Με ένα for loop εμφανίζουμε τα ονόματα των μετρικών που έχουν τη μεγαλύτερη συνεισφορά στα 5 πρώτα principal components του κάθε συνόλου . Εφαρμόζουμε hierarchical clustering σε κάθε σύνολο έτσι ώστε να ομαδοποιήσουμε τις μετρικές σε 5 clusters και έπειτα επιλέγουμε από κάθε cluster μία μετρική εφόσον το κάθε cluster περιέχει μετρικές με ισχυρή συσχέτιση και θέλουμε να αντλήσουμε όσο περισσότερη πληροφορία μπορούμε από τις μετρικές μας. Οι μετρικές που επιλέγονται από το κάθε cluster είναι αυτές που έχουν τη μεγαλύτερη συνεισφορά στα principal components .Η ομαδοποίηση των μετρικών για το σετ μεγάλου μεγέθους φαίνεται στην παρακάτω εικόνα

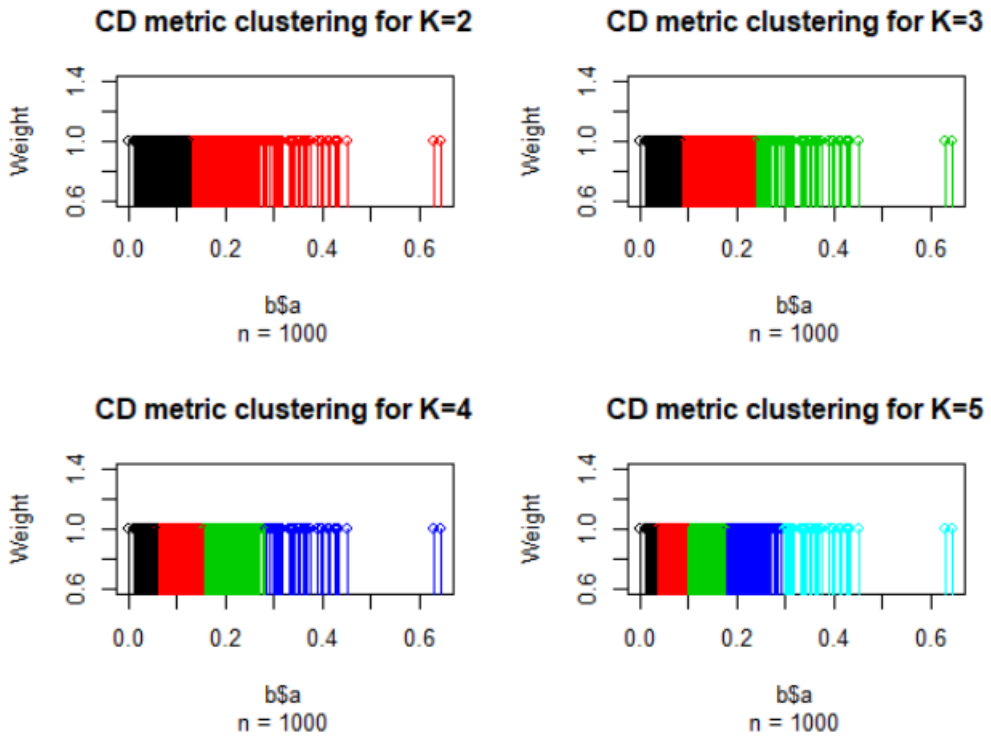


Για το σύνολο μεθόδων μικρού μεγέθους καταλήξαμε στην επιλογή των μετρικών MIMS,CD,CLLC,NL(+)NLE, TLOC . Η μεταβλητή NL+NLE προκύπτει από το άθροισμα των επί μέρους NL και NLE για να κερδίσουμε όλη την πληροφορία που αφορά το nesting level στη μέθοδο για τους if kai else if βρόχους. Για το σύνολο των μεθόδων μεσαίου μεγέθους καταλήξαμε στην επιλογή των μετρικών MI,CLC,CD,NL+NLE,HDIF . Τέλος, για το σύνολο των μεθόδων μεγάλου μεγέθους καταλήξαμε στην επιλογή των μετρικών HVOL,CLC,CD,NL(+)NLE,TLOC.

Διαχωρισμός της κάθε μετρικής σε clusters:

Έπειτα και τον διαχωρισμό του αρχικού σετ δεδομένων σε επιμέρους σετ και την εκλογή των μετρικών που τα περιγράφουν ακολουθεί η ομαδοποίηση των τιμών των επιμέρους μετρικών σε clusters (low,medium,high). Η διαδικασία που ακολουθούμε έχει ως σταθερά την απαλειφή των NA τιμών ενώ διαφέρει όσον αφορά των αριθμό των clusters από ομαδοποίηση σε ομαδοποίηση καθώς διακρίναμε πως στις μετρικές μας δεν αποδίδεται πάντα σαφές νόημα στα clusters για έναν προκαθορισμένο αριθμό ομάδων ενώ για την επιλογή του κατάλληλου αριθμού ομάδων βοηθάει και η εφαρμογή των 1d_optimal_silhouette.R , 2d_optimal_silhouette.R ανάλογα με το πλήθος των διαστάσεων των δεδομένων. Τα παραπάνω scripts υπολογίζουν την τιμή της μετρικής silhouette που συνδυάζει την συνοχή και την απόσταση των διάφορων clusters, για αριθμό κέντρων του Kmeans από 2 έως 5.Σε περιπτώσεις που η τιμή της silhouette διέφερε ελάχιστα επιλέγαμε τον μικρότερο αριθμό clusters για την αποφυγή της πολυπλοκότητας. Παράλληλα, όταν οι τιμές μιας μετρικής μας δεν είναι έντονα πολωμένες (συνήθως η πόλωση εντοπίζεται κοντά στο 0) εφαρμόζουμε με τη βοήθεια του script outlier_detection.R ανίχνευση και διαγραφή των εξωκείμενων τιμών. Η παραπάνω τακτική δεν εφαρμόζεται σαν κανόνας καθώς για έντονα πολωμένα σετ δεδομένων αναιρείται η φυσική σημασία της μετρικής και δημιουργείται

λανθασμένη αντίληψη περί των τιμών της. Για τους ίδιους λόγους , ακολουθήσαμε και την μέθοδο της δειγματοληψίας κατά προτίμηση. Οι βασικοί αλγόριθμοι ταξινόμησης που χρησιμοποιήσαμε είναι ο K-means (στην περίπτωση των μετρικών NL-NLE που συνδυάστηκαν για τα small_size και large_size σύνολα δεδομένων, όπου η απεικόνιση και η 2d ομαδοποίηση τον ευνοεί) και ο Ckmeans που αφορά univariate clustering και χρησιμοποιήθηκε στο σύνολο των υπολοίπων μετρικών. Ξεχωριστή αναφορά μπορεί να γίνει στην αποπομπή της μετρικής CLLC από τις μετρικές που περιγράφουν τα small_size σετ δεδομένων καθώς παρατηρήσαμε πως οι τιμές της ήταν εξαιρετικά πολωμένες στο 0 και δεν υπήρχε επομένως ο διαχωρισμός σε clusters (όλες ανήκουν σε μία ομάδα).Στην παρακάτω εικόνα φαίνεται η διαδικασία του Script 1d_optimal_silhouette.R στη μεταβλητή CD του μεγάλου μεγέθους σετ όπου για κάθε clustering έχει υπολογιστεί το αντίστοιχο silhouette.



Στο συγκεκριμένο παράδειγμα προέκυψε μεγαλύτερο silhouette για 2 clusters και έτσι επιλέχθηκαν 2 κέντρα για την τελική εφαρμογή του K-means.

Δημιουργία ANN για την πρόβλεψη της αναγνωσιμότητας μεθόδων:

Η περιγραφή της διαδικασίας που ακολουθεί αφορά την κατασκευή ενός νευρωνικού ταξινομητή για την πρόβλεψη της αναγνωσιμότητας μιας μεθόδου συναρτήσει των μετρικών που την περιγράφουν σε επαρκή βαθμό. Αρχικώς, δημιουργούμε ένα σετ δεδομένων που αποτελείται από το πλήθος των κατηγοριών παραβιάσεων που έχουμε υποθέσει εξαρχής πως σχετίζονται με την αναγνωσιμότητα ενός αρχείου κώδικα. Εν συνεχεία , πολλαπλασιάζουμε το περιεχόμενο των κελιών των κατηγοριών παραβιάσεων ανάλογα με την βαρύτητα του σφάλματος (*5 για major τύπου παραβιάσεις , *10 για critical τύπου παραβιάσεις, απαράλλαχτο το περιεχόμενο για minor τύπου παραβιάσεις). Υπολογίζουμε στη συνέχεια το βεβαρυμμένο άθροισμα για κάθε μέθοδο και το αποθηκεύουμε σε ένα διάνυσμα (sums) . Το διάνυσμα αυτό αποτελεί τις τιμές 'Y' με τις οποίες θα εκπαιδεύσουμε το μοντέλο , κάνοντας την παραδοχή πως τα αθροίσματα των παραβιάσεων συνδέονται άμεσα με την αναγνωσιμότητα .Κατόπιν, αντιστοιχίζουμε το παραπάνω διάνυσμα με το σύνολο των μετρικών της μεθόδου που την αφορά. Από τον συνολικό πίνακα αφαιρούμε όσες σειρές περιέχουν NA τιμές και στη συνέχεια διαιρούμε σε 3 μεγεθών σύνολα δεδομένων με κριτήριο την τιμή της μετρικής LLOC. Προσθέτουμε , στην συνέχεια τις στήλες NL,NLE που περιγράφουν το nestling της μεθόδου και εξετάζουμε για κάθε μεγέθους σετ δεδομένων. Ακολούθως , δημιουργούμε από τα παραπάνω σετ δεδομένων 3 data frames στα οποία περιέχονται οι μετρικές που μας χρησιμεύουν για την περιγραφή του κάθε σετ μεθόδων αναλόγως το μέγεθος του (πχ MIMS, CD, NL-NLE,TLOC για small_size μεθόδους). Παράγουμε 3 διανύσματα (violations) με τα βεβαρυμμένα αθροίσματα των παραβιάσεων μέσω της αποκοπής της στήλης sums από τους πίνακες των διαφόρων μεγεθών. Με τη χρήση των scipts (small/medium/large_size_violations_clustering.R) εφαρμόζουμε τον Ckmeans για univariate clustering 2 ομάδων στις τιμές των αθροισμάτων των παραβιάσεων προκειμένου να αναγάγουμε το πρόβλημα της κατηγοριοποίησης μιας μεθόδου σε ένα πρόβλημα δυαδικής ταξινόμησης . Ανάλογα την μέση τιμή των παραβιάσεων ενός cluster του αποδίδουμε και χαρακτηρισμό που αφορά το readability του και τιμή -1 ή 1 (πχ για large_size_violations οι χαρακτηρισμοί που προκύπτουν είναι very high readability και medium readability) . Στη συνέχεια, κανονικοποιούμε τις τιμές των μετρικών στο πεδίο [0,1] και δημιουργούμε έναν ενοποιημένο πίνακα με τις παραπάνω τιμές και το διάνυσμα που περιέχει τις -1,1 ανάλογα το cluster που τοποθετήθηκε η κάθε μέθοδος. Σπάμε τον παραπάνω πίνακα σε δύο με κριτήριο τις τιμές του χαρακτηριστικού κλάσης και κάνουμε sampling τον πίνακα με το κυρίαρχο χαρακτηριστικό κλάσης για να επιτύχουμε ισορροπία εγγραφών (50-50). Ενοποιούμε τους παραπάνω πίνακες , εφαρμόζουμε shuffling και μετονομάζουμε τις στήλες του πίνακα για την αύξηση της αναγνωσιμότητας του κώδικα μας. Εν τέλει , με τη χρήση της neuralnet προσπαθούμε να παράξουμε ένα αξιόπιστο εργαλείο ταξινόμησης μεθόδων σε τύπους readability συναρτήσει των μετρικών που τις περιγράφουν.Το μοντέλο φαίνεται στην παρακάτω εικόνα

Πρόβλεψη τιμών του τελικού μοντέλου:

Μετά την εκπαίδευση του μοντέλου, το καλούμε να προβλέψει την κλάση readability (high ή medium) για τιμές των μεταβλητών του σετ μεγάλου μεγέθους μεθόδων . Αρχικά κατασκευάζουμε ένα data frame με τιμές που θέλουμε να αντιστοιχούν σε MI=’LOW’, CLC=’HIGH’, CD=’LOW’, NL+NLE=’HIGH’ ,HEFF=’HIGH’. Είναι προφανές πως με αυτές τις τιμές το μοντέλο θα έπρεπε να προβλέψει medium readability , με πιθανές τιμές για το readability high kai medium, εφόσον το dataset αποτελείται από αρκετά readable μεθόδους. Έτσι σύμφωνα με το clustering που έχει προηγηθεί στις μεταβλητές μας ορίζουμε το data frame X1=c(0.013,0.0003,0,0.013,0.9).Το μοντέλο μας όντως ταξινομεί την εγγραφή στην κλάση ‘medium’. Έπειτα, κάνουμε predict με ένα data frame με τιμές που θέλουμε να αντιστοιχούν σε MI=’HIGH’, CLC=’LOW’, CD=’HIGH’, NL+NLE=’LOW’ ,HEFF=’LOW’. Είναι προφανές πως με αυτές τις τιμές το μοντέλο θα έπρεπε να προβλέψει ‘high‘ readability. . Έτσι σύμφωνα με το clustering που έχει προηγηθεί στις μεταβλητές μας ορίζουμε το data frame X1=c(0.03,0,0.00023,0,0.0009). Το μοντέλο μας όντως ταξινομεί την εγγραφή στην κλάση ‘high’

Στον παρακάτω πίνακα φαίνονται οι προβλέψεις του μοντέλου για τις επιλεγμένες τιμές μεταβλητών.

MI	CLC	CD	NLALL	HDIF	READABILITY
LOW	HIGH	LOW	HIGH	HIGH	MEDIUM
HIGH	LOW	HIGH	LOW	LOW	HIGH

Το μοντέλο επιδέχεται βελτίωσης ,όμως λόγω περιορισμένου χρονικού περιθωρίου δεν είχαμε τη δυνατότητα να ορίσουμε με το βέλτιστο τρόπο τις επιμέρους μεταβλητές του ANN .