Schizophrenia Classification from fMRI data

Fotakis Tzanis,Kritharakis Emmanouil

***Abstract*—Η έγκυρη ανίχνευση ασθενειών αποτελεί σήμερα ένα από τα σημαντικότερα ζητήματα τόσο στον τομέα της ιατρικής όσο και σε αυτόν της αναγνώρισης προτύπων.Στην κατεύθυνση αυτή, μέσα από ιατρικά δεδομένα υλοποιήθηκαν μια σειρά από τεχνικές data preprocessing αλλά και ταξινομητές (classifiers) με σκοπό την ακριβέστερη ανίχνευση της εγκεφαλικής ασθένειας της σχιζοφρένειας.**

***Index Terms*—Classification,Schizophrenia,PCA.**

I. Introduction

Η σχιζοφρένεια αποτελεί μια από τις πιο συχνά εμφανίσιμες παθήσεις του εγκεφάλου.Χαρακτηριστικά, στις Η.Π.Α το 1.1% του συνολικού ενήλικου πληθυσμού πάσχει από σχιζοφρένεια σύμφωνα με μελέτες εν έτη 2018[1].Η διάγνωση και ανίχνευση της είναι ιδιαίτερα δύσκολη στον τομέα της ιατρικής καθώς τα συμπτώματα της ασθένειας επικαλύπτονται με άλλα διαφόρων άλλων ασθενειών.Στην προσπάθεια μας για την δημιουργία ταξινομητών ανίχνευσης της πνευματικής ασθένειας αυτής χρησιμοποιήθηκαν ιατρικά δεδομένα από 86 διαφορετικά άτομα,τα οποία είτε πάσχουν από την ασθένεια είτε όχι.

Ορισμένα εκ των ιατρικών δεδομένων αποτέλεσαν τα Functional Network Connectivity (FNC).Συγκεκριμένα πρόκειται για σχετιζόμενες τιμές, οι οποίες περιγράφουν την διασύνδεση των ανεξάρτητων τμημάτων του εγκεφάλου του εκάστοτε ατόμου.Οι πληροφορίες αυτές εξήχθησαν από εικόνες FMRI(functional magnetic resonance imaging) μέσω της ανάλυσης GICA (group independent component analysis). Το σύνολο των τιμών είναι 378 καθώς κάθε μέτρηση γίνεται σε μια συγκεκριμένη χρονική στιγμή. Με αυτό τον τρόπο εκφράζεται ο "συγχρονισμός" του εγκεφάλου σε ένα χρονικό διάστημα.Εν αντιθέσει με τα χαρακτηριστικά των FNCs τα οποία περιγράφουν την διασύνδεση των εγκεφαλικών περιοχών στο dataset δίνεται ακομα μια σειρά χαρακτηριστικών τα SBMs(Source-Based Morphometry), τα οποία αναλύουν την δομική σύνθεση των εγκεφαλικών περιοχών. Τα SBMs περιγράφουν τη συγκέντρωση της φαιάς ουσίας στα διάφορα τμήματα του εγκεφάλου.Η φαιά ουσία βρίσκεται στο εξωτερικό τμήμα του εγκεφάλου και είναι υπεύθυνη για την επεξεργασία των εγκεφαλικών σημάτων. Η εξαγωγή των χαρακτηριστικών αυτών προκύπτει από ανάλυση ICA(independent component analysis) πάνω σε Structural magnetic resonance imaging (SMRI).Το σύνολο των SBMs είναι 33 και όσο πιο μικρή είναι η τιμή τους τόσο μικρότερη είναι και η συγκέντρωση της φαιάς ουσίας στο συγκεκριμένο τμήμα του εγκεφάλου.

Ειδική μνεία οφείλεται να γίνει στο σύνολο του dataset ως προς το μέγεθος του. Συγκεντρωτικά, το dataset αποτελούνταν απο 3 CSVs.Στο πρώτο(label.csv) υπήρχαν μόνο οι 86 ασθενείς με το id τους και τον δυικό χαρακτηρισμό εάν είναι ή δεν είναι σχιζοφρενείς.Στο δεύτερο(FNC.csv) υπήρχαν οι ασθενείς με τα id τους και 378 διαφορετικές τιμές,οι οποίες και αναπαριστούσαν τις συσχετιζόμενες τιμές στα διαφορετικά brain maps σε 378 διαφορετικές χρονικές στιγμές.Τέλος, στο τρίτο(SBM.csv) ενυπήρχαν τα άτομα με τα id τους και 33 χαρακτηριστικά για την ποσόστωση της φαιάς ουσίας στα διάφορα brain maps.Με τα δεδομένα αυτά, το πρόβλημα που παρουσιάστηκε να επιλυθεί κατηγοριοποιήθηκε ως "High Dimensional Small Sample Size Data"[2],ελέω των 86 data points που είχαμε στην διάθεση μας,γεγονός που αποτέλεσε σημαντικό παράγοντα στο τρόπο διαχείρισης των δεδομένων με σκοπό την βέλτιστη εκπαίδευση του εκάστοτε ταξινομητή.

II. State of art

Στο ευρύ σύνολο του ο κλάδος της αναγνώρισης προτύπων έχει χρησιμοποιήσει ιατρικά δεδομένα για ασθένειες αντίστοιχης δυσκολίας ανίχνευσης όπως η σχιζοφρένεια.Συγκεκριμένα πάνω στην ασθένεια αυτή το 2014 η ιστοσελίδα Kaggle δημοσίευσε έναν διαγωνισμό[3]ανάμεσα στους χρήστες της για την ανίχνευση της μέσω μοντέλων αναγνώρισης προτύπων.Οι χρήστες υλοποίησαν μια σειρά από διαφορετικούς ταξινομητές με νικητή τον Arno Solin ο οποίος με την χρήση Gaussian Processes[4] κατάφερε και πέτυχε ακρίβεια αποτελεσμάτων κοντά στο 80%.Στην δεύτερη θέση ακολούθησε ο Alex Lebedev με Support Vector Machines και την τριάδα συμπληρώνει ο Karolis Koncevičius με Distance Weighted Discriminant (DWD).

Παρεμφερείς εργασίες πάνω στην ανίχνευση επίκαιρων παθήσεων έχουν γίνει στην ιστοσελίδα της kaggle μέσω διαγωνισμών της. Συγκεκριμένα,το 2015 ξεκίνησε διαγωνισμός[5] με τίτλο "Diabetic Retinopathy Detection", όπου γίνεται classification για την ασθένεια του διαβήτη μέσα από δεδομένα που συλλέγονται από το μάτι του ασθενούς ενώ το 2016 αντίστοιχος διαγωνισμός[6] με τίτλο "Cervical Cancer Screening" όπου γίνεται classification για την ασθένεια του καρκίνου του τραχήλου της μήτρας μέσα από δεδομένα που συλλέγονται από την εν λόγω περιοχή της ασθενούς.Εργασίες και Paper πάνω στον τρόπο εξαγωγής των δεδομένων με τις τεχνικές των fMRI και sMRI παρατίθεται στην βιβλιογραφία.

III. ΠΕΡΙΓΡΑΦΗ ΤΕΧΝΙΚΩΝ ΥΛΟΠΟΊΗΣΗΣ

Οι τεχνικές που υλοποιήθηκαν στην προσπάθεια εύρεσης της καλύτερης δυνατής ακρίβειας της σχιζοφρένειας χωρίστηκαν σε 2 κατηγορίες. Στην πρώτη κατηγορία, κατηγοριοποιούνται οι τεχνικές data preproccesing,ενώ την δεύτερη κατηγορία αποτελούν οι διάφοροι ταξινομητές που υλοποιήθηκαν για την κατηγοριοποίηση των ατόμων σε ασθενείς ή όχι.

Ως προς το data processing , οι κύριες τεχνικές που υλοποιήθηκαν ήταν από την μία η κανονικοποίηση (normalization)

των τιμών των δεδομένων και από την άλλη η principal component analysis (PCA)[7] με στόχο αφενός την μείωση του συνόλου των διαστάσεων των χαρακτηριστικών αλλά και την διατήρηση των πλέον σημαντικών.Η κανονικοποίηση έγινε μέσω της συνάρτησης Minmaxscaler που παρέχεται από την βιβλιοθήκη sklearn.preprossesing και στόχο έχει να εξομαλύνει τις τιμές κάθε χαρακτηριστικού βάση της μικρότερης και μεγαλύτερης τιμής του στο σύνολο όλων των ατόμων. Όσο αναφορά την τεχνική του PCA ,αποτελούσε ένα προαιρετικό κομμάτι της συνολικής εργασίας. Μέσα από το PCA έγινε το κατάλληλο feature selection ,όπου κάθε ένα από τα 86 άτομα με συνολικά 411 χαρακτηριστικά κατέληξε να μπορεί να περιγραφεί με βάση τα χαρακτηριστικά με μέγιστη διακύμανση μόλις με 64 από αυτά. Η μέθοδος PCA προσπαθεί να μειώσει τις διαστάσεις του συνόλου των χαρακτηριστικών εφαρμόζοντας γραμμικούς μετασχηματισμούς στα δεδομένα και δημιουργώντας νέα χαρακτηριστικά από τα οποία αυτά με την μεγαλύτερη διακύμανση επιλέγονται. Ο λόγος που επιλέγεται η διακύμανση ως παράγοντας για την εξαγωγή των πλέον σημαντικών χαρακτηριστικών βασίζεται στην θεωρία ότι χαρακτηριστικά με μεγαλύτερη διακύμανση διακριτοποιούν καλύτερα τα άτομα.

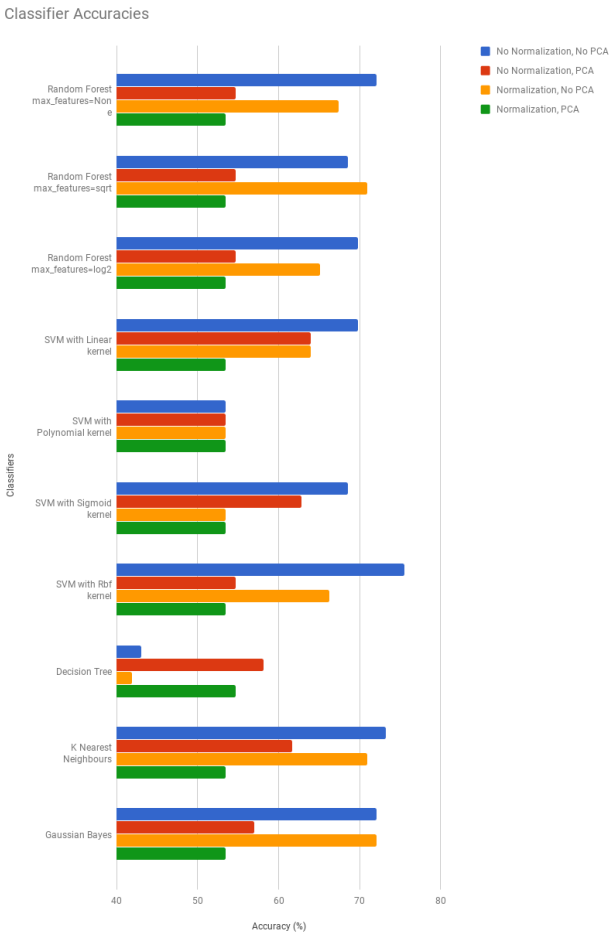
Ως προς τους classifiers[8],υλοποιήθηκαν μια σειρά από classifiers διαφορετικής λογικής για την εξεύρεση του πλέον κατάλληλου.Αρχικά υλοποιήθηκε ο **Decision Tree** classifier.Η λογική του Decision Tree αναφέρει πως από το σύνολο των χαρακτηριστικών δημιουργείται ένα binary search Tree με στόχο την τελική διακριτοποίηση των ατόμων σε ασθενής ή μη. Πρόκειται για έναν απλό αλγόριθμο κατάλληλο για binary classification. Ως συνέχεια της λογικής του Decision Tree classifier υλοποιήθηκε ο **Random Forest** classifier. Στον συγκεκριμένο αλγόριθμο,διατηρείται η λογική των binary trees για το classification αλλά αυξάνονται πλέον για την καλύτερη τελική εκτίμηση των data points και κάθε δέντρο λαμβάνει μια απόφαση.Η τελική απόφαση εξάγεται μέσα από ψηφοφορία ανάμεσα στα δέντρα (voting) και λαμβάνεται το αποτέλεσμα με τις περισσότερες ψήφους.Στον αλγόριθμο αυτόν, επηρεάστηκαν ορισμένα από τα hyperparameters του με στόχο την βέλτιστη απόδοση. Συγκεκριμένα, το πλήθος των δέντρων που θα συμμετέχουν στην τελική απόφαση ορίστηκε να είναι 1000 καθώς γύρω και πάνω από αυτή την τιμή η ακρίβεια απόφασης παραμένει σταθερή.Επιπλέον ελέγχθηκε η μέθοδος με την οποία τοποθετούνται οι ερωτήσεις για τα features πάνω στα trees. Το χαρακτηριστικό αυτό διακρίνεται είτε με την μέθοδο gini index είτε με την μέθοδο της εντροπίας όπου καλύτερα αποτελέσματα βρέθηκαν με την πρώτη.Τέλος, ένα ακόμα hyperparameter που χρησιμοποιήθηκε ήταν αυτό του πλήθους των features που θα υπάρχουν πάνω στα δέντρα.Οι διαφορετικές μας υλοποιήσεις έγιναν είτε με την τετραγωνική ρίζα του πλήθος είτε με την λογαριθμική τους τιμή ή το πλήρες σύνολο των χαρακτηριστικών.Στην συνέχεια,χρησιμοποιήθηκε μια διαφορετική ομαδά ταξινομητών(neighbour classifiers) με κύριο εκπρόσωπο της τον **K-nearest-neighbours** classifier. Η λογική του ταξινομητή αυτού έγκειται στην διακριτοποίηση ενός νέου data point με βάση τα κοντινότερα ήδη εκπαιδευμένα data points.Στον συγκεκριμένο αλγόριθμο επιδράσαμε πάνω του ελέγχοντας το βέλτιστο αριθμό γειτόνων που αυξάνουν την ακρίβεια των αποτελεσμάτων και καταλήξαμε στο πλήθος των 10 γειτόνων.Μια άλλη κατηγορία classifiers(bayes classifiers) υλοποιήθηκε με κύριο εκπρόσωπο της την **Gaussian Bayes** classifier. Η λογική πίσω από αυτή την κατηγορία είναι η μαθηματική έκφραση που εισήγαγε ο Bayes για την πιθανότητα εύρεσης της κατηγοριοποίησης των ατόμων με βάση τα χαρακτηριστικά που διαθέτουμε. Η διαφοροποίηση στην κλασσική αυτή σχέση είναι ότι η πιθανότητα των features δεδομένου των labels θεωρείται ότι ακολουθεί Gaussian κατανομή.

Η τελευταία μεγάλη κατηγορία ταξινομητών που ελέγθηκε είναι αυτή των **Support Vector Machines** γνωστών και ως SVMs.Η λογική πίσω από τα SVMs βασίζεται στην δημιουργία του κατάλληλου hyperplane για την διακριτοποίηση των δεδομένων.Το hyperplane βασίζεται στα λεγόμενα "support vectors",τα οποία είναι διανύσματα κατάλληλης διάστασης που οριοθετούν τα εκπαιδευμένα data points που είναι σχιζοφρενής ή μη με τέτοιο τρόπο έτσι ώστε τα επόμενα data points του test να μπορούν να διακριτοποιηθούν ανάμεσα στο hyperplane σε σχιζοφρενή ή όχι.Τα hyperparameters που επηρεάστηκαν ήταν τόσο τα kernels όσο και οι τιμές gamma και C.Αρχικά τα kernels διακρίνονται σε linear,polynomial,sigmoid και radis basis function. Τα kernels αποτελούν συναρτήσεις που μετασχηματίζουν το επίπεδο των διαστάσεων που είναι τοποθετημένα τα εκπαιδευμένα data points με σκοπό την καλύτερη διακριτοποίηση των ατόμων μέσω του hyperplane που θα σχηματιστεί. Πέραν των kernels, σημαντική παράμετρος στην οικογένεια των SVMs αποτελεί το C.Εν γένει το hyperplane διακριτοποιεί απόλυτα τις 2 περιοχές των data points σε σχιζοφρενείς ή μη.Παρόλα αυτά, ανάλογα την τιμή του C(στην συγκεκριμένη περίπτωση C=2) επιτρέπονται data points που βρίσκονται κοντά στο hyperplane αλλά σε λανθασμένες περιοχές κατηγοριοποίησης να ανιχνεύονται και να κατηγοριοποιούνται σωστά.Γι αυτό και στην θεωρία η παράμετρος C ορίζεται και ως soft margin καθώς είναι ένα "χαλαρό" όριο απόφασης για το εκάστοτε data point γύρω απο το hyperplane.Όσον αφορά την gamma(στην συγκεκριμένη περίπτωση gamma=0.01) παράμετρο είναι υπεύθυνη στο να καθορίσει ποία data points θα ορίσουν το hyperplane ανάμεσα στους σχιζοφρενείς και στους υγιείς ανθρώπους. Η θέση του hyperplane εξαρτάται από τα data points που επιλέγονται με βάση την τιμή gamma. Όσο μεγαλύτερη είναι αυτή η τιμή τόσο πιο απομακρυσμένα από το σημείο διαχωρισμού είναι τα data points που επιλέγονται.

IV. ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ

Με βάση όλους τους παραπάνω αλγορίθμους εκτελέστηκαν μια σειρά από πειράματα πάνω στο δεδομένο dataset. Τα 2 CSVs με τα δεδομένα των FNCs και SBMs ενώθηκαν σε ένα CSV καθώς τα αποτελέσματα των accuracies ήταν αρκετά χαμηλά για κάθε CSV ξεχωριστά. Ως προς τον τρόπο εκπαίδευσης του εκάστοτε ταξινομητή επιλέχθηκε η μέθοδος Leave One Out.Στην συγκεκριμένη περίπτωση λόγω ότι έχουμε ένα "High Dimensional Small Sample Size Data" πρόβλημα με 86 data points ο καλύτερος τρόπος για να εκπαιδευτεί ο ταξινομητής είναι με το να κρατάμε 1 data point για testing και τα υπόλοιπα 85 για training.Εκτελώντας αυτή την

λογική για κάθε ένα από τους 86 δυνατούς συνδυασμούς και παίρνοντας την μέση τιμή των accuracies λαμβάνεται η καλύτερη δυνατή εκτίμηση του accuracy για κάθε εκτιμητή.Στον παρακάτω πίνακα αποδίδονται όλοι οι αλγόριθμοι για 4 διαφορετικές περιπτώσεις.Είτε θα εκτελεστούν οι εκτιμητές με normalization και PCA ,είτε χωρίς αυτά,είτε με PCA χωρίς normalization είτε με normalization χωρίς PCA.



V. Conclusion

Τα συμπεράσματα που εξάγονται από την παραπάνω εικόνα είναι πολλά και χρήσιμα.Αρχικά,ο καλύτερος εκτιμητής μας είναι ο SVM με rbf kernel χωρίς normalization και PCA,με το accuracy του να φτάνει κοντά στο 76%,ένω στην συνέχεια ακολουθούν ο random forest και ο K-nearest-neighbours.Ο λόγος που θεωρούμε πως οι αλγόριθμοι δεν πετυχαίνουν υψηλότερα ποσοστά ακρίβειας έγκειται στο γεγονός ότι διαθέτουμε μικρό dataset γεγονός που δεν μπορεί να εκπαιδεύσει τόσο καλά τους παραπάνω εκτιμητές.Επιπλέον παρατηρείται πως εν γένει στο συγκεκριμένο πρόβλημα η χρήση του PCA προκαλεί πτώση του accuracy στην πλειονότητα των estimators γεγονός που δείχνει πως γραμμικοί μετασχηματισμοί επηρεάζουν αρνητικά τα δεδομένα μας.Τρίτον, οφείλεται να γίνει αναφορά και στην μέθοδο Leave One Out.Σε αντίθεση με απλές μεθόδους train test split όπου η αναλογία train και test ήταν $80\% - 20\%$,απέτρεπαν τον εκτιμητή να εκπαιδευτεί με αρκετά data points χειροτερεύοντας του το accuracy.Ακόμα και βελτιώσεις της απλής train test split ,όπως k-fold cross validation,στο οποίο το συνολικό data set σπάει σε k τμήματα και για όλους τους δυνατούς συνδιασμούς των τμημάτων αυτών ως κομμάτια για train και ένα για test εξάγονται αρκετά accuracies που στο τέλος λαμβάνεται η μέση τιμή τους,δεν κατάφεραν να πετύχουν αξιόπιστα αποτελέσματα στην ανίχνευση της σχιζοφρένειας.Συνεπώς η Leave One Out αποτέλεσε την καλύτερη δυνατή μέθοδο διαχωρισμού των data points σε train και test.Τέταρτον,παρατηρούμε πως τα δεδομένα μας δεν χρειάζονται στην πλειονότητα των estimators μας normalization,στοιχείο που φαίνεται στα accuracies του πίνακα με κύριο εκφραστή την οικογένεια των SVMs στους οποίους και η πτώση των ακριβειών είναι αρκετά μεγάλη της τάξης του 20%.Τέλος, παρατηρούμε πως στην πλειονότητα των περιπτώσεων ο decision tree είναι κατά πολύ χειρότερος από τον random forest γεγονός που επιβεβαιώνει την θεωρία.

References

[1] https://www.medicinenet.com/script/main/art.asp?articlekey=41430
[2] Raudys 1991 Small sample size effects in statistical pattern recognition Recommendations for practitioners.pdf
[3] https://www.kaggle.com/c/mlsp-2014-mri
[4] Solin MLSP 2014 schizophrenia classification challenge Winning model documentation.pdf
[5] https://www.kaggle.com/c/diabetic-retinopathy-detection
[6] https://www.kaggle.com/c/cervical-cancer-screening
[7] Βιβλίο "Αναγνώριση Προτύπων" των συγγραφέων Sergios Theodoridis&Konstantinos Koutroumbas σελίδες 374-376
[8] http://scikit-learn.org/stable/index.html