Schizophrenia Classification

Κατάρα Σωτηρία Μαρία, Μανάρα Χριστίνα
Πολυτεχνείο Κρήτης
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Χανιά, 2019

Abstract- Το ερευνητικό αυτό paper περιγράφει την διαδικασία που ακολουθήθηκε για την εκπαίδευση ενός μοντέλου με σκοπό την κατηγοριοποίηση μια ομάδας ασθενών σε πάσχοντες από σχιζοφρένεια και υγιείς κάνοντας χρήση διαφόρων αλγορίθμων για ταξινομητές και καταλήγοντας στον πιο αποδοτικό, ο οποίος δίνει και την μεγαλύτερη υπολογιστική ακρίβεια.

Index Terms- Schizophrenia, Normalization, Classification

I. Introduction

Η σχιζοφρένεια είναι μία από τις δέκα σημαντικότερες αιτίες μακροχρόνιας ανικανότητας των ασθενών παγκοσμίως, ενώ περίπου το 1% του πληθυσμού εκδηλώνει αυτή τη διαταραχή. Το κύριο κύριο γνώρισμα της σχιζοφρένειας είναι η διαταραχή στην σκέψη. Ανήκει στις ψυχωσικές διαταραχές και έχει ποικιλόμορφη κλινική εικόνα. Πιο συγκεκριμένα, ο ασθενής που πάσχει από σχιζοφρένεια μπορεί να παρουσιάσει θετικά (παραληρητικές ιδέες, ψευδαισθήσεις, αποδιοργάνωση της συμπεριφοράς), αρνητικά (συναισθηματική επιπέδωση, αλογία, αβουλία και ανηδονία) ή μεικτά συναισθήματα. Ωστόσο, είναι αρκετά δύσκολο να επιτευχθεί η ορθή διάγνωση της ασθένειας, καθώς τα συμπτώματά της επικαλύπτονται με άλλα διάφορων άλλων ψυχικών ασθενειών. Προκειμένου να πραγματωθεί η όσο το δυνατόν καλύτερη ανίχνευση της συγκεκριμένης πνευματικής ασθένειας, εφαρμόζονται κατάλληλοι ταξινομητές, οι οποίοι χρησιμοποιούν ιατρικά δεδομένα από 86 διαφορετικά άτομα.

Τα ιατρικά δεδομένα που λαμβάνονται υπόψη για τη διάγνωση της σχιζοφρένειας είναι τα Functional Network Connectivity (FNC) και τα Source-Based Morphometry (SBM). Τα πρώτα αφορούν σχετιζόμενες τιμές, οι οποίες περιγράφουν τη διασύνδεση των ανεξάρτητων τμημάτων του εγκεφάλου του εκάστοτε ατόμου. Οι πληροφορίες αυτές εξάγονται από εικόνες fMRI (functional MRI) μέσω της ανάλυσης GICA (Group Independent Component Analysis). Εν αντιθέσει, τα δεύτερα σχετίζονται με τη συγκέντρωση φαιάς ουσίας στα διάφορα τμήματα του εγκεφάλου. Ειδικότερα, η φαιά ουσία βρίσκεται στο εξωτερικό τμήμα του εγκεφάλου και ευθύνεται για την αποστολή εγκεφαλικών σημάτων. Τα αποτελέσματα των SBM χαρακτηριστικών εξάγονται μέσω της ανάλυσης ICA (Independent Component Analysis), η οποία εφαρμόζεται πάνω στις εικόνες που παράγονται από τις sMRI (structural MRI) σαρώσεις.

Τα ιατρικά δεδομένα έχουν παρθεί και βρίσκονται συγκεντρωτικά σε τρία αρχεία τύπου .CSV. Το πρώτο από αυτά (train_labels.csv) παρέχει την πληροφόρηση σχετικά με το αν τα διαφορετικά άτομα, τα οποία συμμετέχουν στο πείραμα είναι σχιζοφρενείς ή όχι. Επομένως, στο αρχείο περιέχεται το πλήθος των ατόμων μαζί με τον εκάστοτε χαρακτηριστικό αριθμό (Id) και τον δυαδικό αριθμό που είναι απαραίτητος για τη διάγνωση, με την τιμή 1 να αντιστοιχεί σε σχιζοφρενή και 0 σε υγιή. Το επόμενο αρχείο (train_FNC.csv) περιλαμβάνει τους ασθενείς με το Id τους, καθώς και 378 διαφορετικές τιμές για κάθε ασθενή. Οι τιμές αυτές αποτελούν τις συσχετισμένες τιμές στους διαφορετικούς χάρτες του εγκεφάλου, δίνουν δηλαδή τον βαθμό σχετικότητας μεταξύ δύο ζευγαριών εγκεφαλικών τμημάτων κατά τη διάρκεια μίας δραστηριότητας. Τέλος, στο τρίτο αρχείο (train_SBM.csv) δίνεται η πληροφορία για τη συγκέντρωσης φαιάς ουσίας για κάθε ασθενή κατά τη

διάρκεια κάποιας δραστηριότητας, μέσω 75 χαρτογραφιμένων εγκεφάλων. Με τα δεδομένα αυτά, πραγματοποιείται η προσπάθεια της επίλυσης του προβλήματος, "High Dimensional Small Sample", καθώς διατίθενται 86 δείγματα, ενώ το σύνολο των χαρακτηριστικών αθροίζεται στα 410. Ο εντοπισμός του προβλήματος καθορίζει σημαντικά και την βέλτιστη εκπαίδευση του μοντέλου μέσω των χρησιμοποιούμενων ταξινομητών.

II. State Of The Art

Η αναγνώριση προτύπων είναι υποσύνολο της τεχνητής νοημοσύνης και περιλαμβάνει εργασίες όπως η αναγνώριση προσώπων, αντικειμένων, λόγου, γραφής, ασθενειών όπως στην συγκεκριμένη περίπτωση καθώς και πολλών ακόμη. Σημαντική είναι η συμβολή της σε διάγνωση ασθενειών από πολύπλοκη συνεκτίμηση αποτελεσμάτων ιατρικών εξετάσεων, όπως αυτή της σχιζοφρένειας. Πιο συγκεκριμένα πάνω στην ταξινομηση σχιζοφρενών και μη, πέντε χρόνια πριν η ιστοσελίδα Kaggle [1] ξεκίνησε έναν διαγωνισμό, ο οποίος προσκαλούσε τους συμμετέχοντες να διαγνώσουν αυτομάτως τα άτομα με σχιζοφρένεια που βασίζονται σε πολυτροπικά χαρακτηριστικά, τα οποία προέρχονται από εξετάσεις MRI (Magnetic Resonance Imaging). Η βέλτιστη λύση προήλθε από τον νικητή του διαγωνισμού Arno Solin, ο οποίος βασίστηκε στη χρήση ενός Gaussian Process Classifier με την θεώρηση ότι οι παρατηρήσεις προέρχονταν από την κατανομή Bernoulli, ενώ στη δεύτερη θέση βρέθηκε ο Alexander V. Lebedev με το SVM-RBF classifier και τέλος στην τρίτη ο Karolis Koncevičius με τον Distance Weighted Discriminant (DWD) classifier.

Άλλος πιο πρόσφατος διαγωνισμός και μάλιστα φετινός ήταν εκείνος της ταξινόμησης υγιών και λευχαιμικών κυττάρων, ο οποίος οργανώθηκε από το SBILab [2].

III. Περιγραφη Τεχνικων Υλοποιησης και Αλγοριθμων

Για την ανάγνωση των training .csv αρχείων (train_FNC, train_SBM, train_labels) έγινε χρήση της βιβλιοθήκης pandas της Python, με σκοπό την άντληση των μεθόδων *read*, *concat* και *drop*. Αφού διαβάστηκαν τα τρία αρχεία μέσω της συνάρτησης read_csv, έγινε συνένωση των αρχείων train_FNC και train_SBM μέσω της συνάρτησης concat. Από το αρχείο της συνένωσης αφαιρέθηκαν οι στήλες με τα id, έτσι ώστε να παραμείνουν μόνο οι τιμές FNC και SBM. Εν συνεχεία μέσω της συνάρτηση *pandas.DataFrame.values* έγινε διαχωρισμός του αρχείου train_labels, καθώς επιστράφηκαν μόνο τα labels (0,1) του αρχείου και όχι τα ονόματα των αξόνων. Η ίδια διαδικασία ακολουθήθηκε και για την εξαγωγή των features από το αρχείο συνένωσης. Ακολούθησε η διαδικασία του feature normalization μέσω της έτοιμης συνάρτησης Normalizer, με σκοπό την αλλαγή των τιμών στο dataset σε μια κοινή κλίμακα. Το feature extraction έγινε με principal component analysis (PCA) με σκοπό την μείωση των διαστάσεων του dataset από 410 σε 80, διατηρώντας έτσι τα πιο σημαντικά χαρακτηριστικά.

Για την εύρεση του κατάλληλου ταξινομητή έγινε δοκιμή των παρακάτω αλγορίθμων.

Gaussian Naive Bayes Ο αλγόριθμος Gaussian Naive Bayes στοχεύει στην πιθανοτική προσέγγιση. Ουσιαστικά, περιλαμβάνει τον υπολογισμό των πιθανών (prior probability) και μεταγενέστερων πιθανοτήτων (posterior probability) των κλάσεων του dataset, καθώς και τα δεδομένων δοκιμών, που δίνονται σε μια κατηγορία αντίστοιχα[5]. Ο ταξινομητής Naive Bayes υποθέτει την υπό συνθήκη ανεξαρτησία των χαρακτηριστικών δεδομένης της κατηγορίας. Αυτή η ισχυρή υπόθεση συνήθως εξασφαλίζει αξιόπιστες εκτιμήσεις των υπό συνθήκη πιθανοτήτων οι οποίες απαιτούνται για ταξινόμηση, ακόμα και από πολύ μικρά σύνολα δεδομένων. Οι υλοποιήσεις του ταξινομητή Naive Bayes συχνά υποθέτουν ότι χρησιμοποιούνται μόνο διακριτά χαρακτηριστικά, επομένως τα συνεχή χαρακτηριστικά πρέπει να διακριτοποιηθούν εκ των προτέρων.

Support Vector Machine Το SVM αποτελεί μια από τις πιο ακριβείς προσεγγίσεις διακρινουσών συναρτήσεων για ταξινόμηση. Ο δυαδικός ταξινομητής SVM προσπαθεί να βρει ένα υπερεπίπεδο απόφασης το οποίο διαχωρίζει το σύνολο των παραδειγμάτων εκπαίδευσης με τέτοιο τρόπο ώστε τα παραδείγματα που ανήκουν στην ίδια κατηγορία να είναι στην ίδια πλευρά του

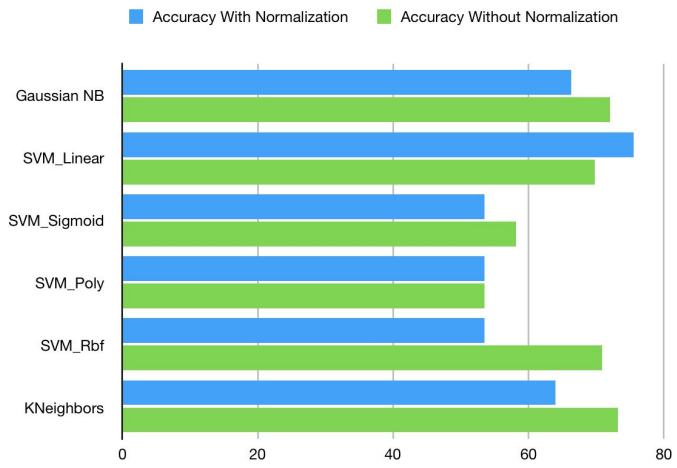
υπερεπιπέδου. Μεταξύ όλων των πιθανών υπερεπιπέδων αναζητά εκείνο για το οποίο η απόσταση από το κοντινότερο παράδειγμα είναι μέγιστη, δηλαδή αναζητά υπερεπίπεδο μέγιστου περιθωρίου (maximal margin hyperplane). Γενικά ο ταξινομητής SVM είναι μια μέθοδος βελτιστοποίησης πολλαπλών κριτηρίων και παραμέτρων: Μεγιστοποιεί την απόσταση μεταξύ των διανυσμάτων υποστήριξης (support vectors) και ενός υπερεπιπέδου απόφασης. Τα διανύσματα υποστήριξης είναι τα παραδείγματα εκπαίδευσης που βρίσκονται πιο κοντά στο υπερεπίπεδο και καθορίζουν το περιθώριο του (margin). Η μέθοδος μεγιστοποιεί το περιθώριο, το οποίο αποτελεί μέτρο της γενικευτικής ικανότητας του ταξινομητή, καθώς ταξινομητές που παράγουν όρια απόφασης με μικρά περιθώρια είναι ευάλωτοι σε φαινόμενα υπερεκπαίδευσης (model overfitting). Στα περισσότερα προβλήματα ταξινόμησης τα παραδείγματα εκπαίδευσης δεν είναι γραμμικά διαχωρίσιμα. Σ΄αυτή τη περίπτωση, τα SVM απεικονίζουν το αρχικό σύνολο χαρακτηριστικών σε ένα σύνολο μεγαλύτερης διάστασης μέσω ενός εσωτερικού γινομένου στην συνάρτηση απόφασης που χρησιμοποιείται το οποίο ονομάζεται πυρήνας (kernel)[6]. Ο πυρήνας μπορεί να είναι γραμμικός (linear), πολυωνυμικός (polynomial), ακτινικής συνάρτησης βάσης (radial basis function) και σιγμοειδής (sigmoid). Η παράμετρος gamma καθορίζει πόσο μακριά φτάνει η επιρροή ενός μόνο παραδείγματος εκπαίδευσης, με χαμηλές τιμές που σημαίνουν «μακριά» και υψηλές τιμές που σημαίνουν «κοντά». Με άλλα λόγια, με χαμηλό gamma, τα σημεία μακριά από πιθανή γραμμή διαχωρισμού λαμβάνονται υπόψη στον υπολογισμό της γραμμής διαχωρισμού. Με υψηλό γάμμα σημαίνει ότι τα σημεία μακριά από την γραμμή διαχωρισμού λαμβάνονται επίσης υπόψη στον υπολογισμό. Επειδή το πρόβλημα της ταξινόμησης μπορεί να επιλυθεί χρησιμοποιώντας σύνθετες διακρίνουσες συναρτήσεις, τα SVM επιδιώκουν να ελαχιστοποιούν το μέγεθος της λύσης (το άθροισμα των βαρών των χαρακτηριστικών). Η παράμετρος κανονικοποίησης (C in Python) ορίζει το πόσο επιθυμούμε να αποφύγουμε την εσφαλμένη ταξινόμηση κάθε παραδείγματος εκπαίδευσης. Για μεγάλες τιμές του C, η μέθοδος θα επιλέξει ένα υπερεπίπεδο μικρότερου περιθωρίου υπό την προϋπόθεση ότι θα ταξινομήσει σωστά τα δείγματα. Αντιστρόφως μια πολύ μικρή τιμή του C θα αναγκάσει τον ταξινομητή να αναζητήσει ένα υπερεπίπεδο με αρκετά μεγάλο περιθώριο ακόμη και αν αυτό οδηγεί σε εσφαλμένη ταξινόμηση των σημείων. Παρόλο που ο μετασχηματισμός των χαρακτηριστικών είναι μη γραμμικός, κάποια από τα παραδείγματα δεν ταξινομούνται σωστά. Τα SVM τείνουν να μειώνουν το πλήθος των εσφαλμένων ταξινομήσεων. Σ' αυτή την περίπτωση αναζητούν ένα υπερεπίπεδο «χαλαρού» περιθωρίου, δηλαδή μια επιφάνεια απόφασης η οποία διαχωρίζει τα δεδομένα εκπαίδευσης κάνοντας τα λιγότερα λάθη. Τα SVM χρησιμοποιούν διάφορους αλγόριθμους βελτιστοποίησης ώστε να ικανοποιούνται όλα τα κριτήρια ταυτόχρονα, παρόλα αυτά χρειάζεται να σταθμίσουμε κατάλληλα όλα τα παραπάνω κριτήρια ώστε να πετύχουμε την καλύτερη ταξινόμηση.

K-Nearest Neighbor Ο k-NN ταξινομητής ταξινομεί τα αντικείμενα με βάση τα k κοντινότερα σε αυτά δεδομένα εκπαίδευσης στο χώρο των χαρακτηριστικών. Στην ουσία, είναι ένας ταξινομητής όπου η συνάρτηση ανομοιότητας (ή ομοιότητας) προσεγγίζεται τοπικά, δηλαδή δεν λαμβάνει υπόψη την καθολική δομή των δεδομένων, αλλά μόνο των k κοντινότερων στο προς ταξινόμηση αντικείμενο. Ανήκει στους απλούστερους αλγορίθμους μηχανικής μάθησης[4]. Ο αλγόριθμος απαιτεί να ορίσει ο χρήστης τον αριθμό k των γειτόνων, ο οποίος πρέπει να βρίσκεται μεταξύ 10 και 50[3].

IV. Πειραματικο Μερος

Οι παραπάνω αλγόριθμοι εφαρμόζονται κατάλληλα πάνω στο δεδομένο dataset. Για τον εκάστοτε ταξινομητή έχει επιλεχθεί η μέθοδος Leave One Out, η οποία ανακουφίζει από την έλλειψη ανεξαρτησίας μεταξύ των συνόλων εκπαίδευσης και δοκιμής. Η εκπαίδευση επιτυγχάνεται χρησιμοποιώντας N-1 δείγματα, και η δοκιμή εκτελείται χρησιμοποιώντας το δείγμα που εξαιρέθηκε. Αν αυτό ταξινομηθεί εσφαλμένα, προσμετράται ως ένα σφάλμα. Η διαδικασία επαναλαμβάνεται N φορές, εξαιρώντας κάθε φορά ένα διαφορετικό δείγμα. Με αυτό τον τρόπο χρησιμοποιούνται όλα τα δείγματα και ταυτόχρονα διατηρείται η ανεξαρτησία μεταξύ συνόλου εκπαίδευσης και δοκιμής. Στον παρακάτω πίνακα περιγράφονται τα αποτελέσματα που

προκύπτουν από την εφαρμογή των διαφορετικών αλγορίθμων, όπως αυτοί αναλύονται προηγουμένως[7].



V. Συμπερασματα

Τα συμπεράσματα που εξάγονται από το παραπάνω γράφημα χωρίζονται σε δύο βασικές κατηγορίες. Η μία από αυτές σχετίζεται με τα αποτελέσματα που προκύπτουν με την εφαρμογή της μεθόδου της κανονικοποίησης. Στην περίπτωση αυτή, ο αποδοτικότερος εκτιμητής είναι ο SVM με Linear Kernel με ποσοστό επιτυχίας 75.58%. Στη συνέχεια, ακολουθεί ο Gaussina-NB με accuracy 66.28%, και ακολουθούν οι k-NN, SVM Sigmoid, Poly και Rbf Kernel. Δεν είναι δυνατόν να επιτευχθεί μεγαλύτερο ποσοστό επιτυχίας από αυτό που πετυχαίνει ο εκτιμητής SVM με Linear Kernel, καθώς το dataset, που χρησιμοποιείται δεν επαρκεί, ώστε να εκπαιδευτούν βέλτιστα οι εκτιμητές.

Παρατηρώντας το μικρό εύρος των τιμών των χαρακτηριστικών, προκύπτει το συμπέρασμα πως δεν είναι απαραίτητη η εφαρμογή της μεθόδου της κανονικοποίησης. Σε αυτή την περίπτωση, ο αποδοτικότερος εκτιμητής είναι ο k-NN με accuracy 73.26%. Στη συνέχεια, ακολουθούν οι Gaussina-NB με accuracy 72.09% και οι SVM Sigmoid, Poly και Rbf Kernel. Μια σημαντική παρατήρηση, που αξίζει να σημειωθεί, είναι το γεγονός ότι χωρίς την εφαρμογή της μεθόδου της κανονικοποίησης η απόδοση όλων των εκτιμητών αυξάνεται αρκετά εκτός από αυτή του SVM Poly, η οποία παραμένει η ίδια.

VI. References

[1].https://www.kaggle.com/c/mlsp-2014-mri/overview

[2].https://competitions.codalab.org/competitions/20429

[3].http://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/3777/Kiriakopoulou.pdf?sequence=2&isAllowed=y

[4].http://www.cs.uoi.gr/tech_reports//publications/MT-2012-01.pdf?fbclid=IwAR1KAJtwLis4y0oNV1R7xkpWUY2D4jO_RZbzawiuCQ3pZ9RyBxbqjqsWIC8

[5].https://hackernoon.com/implementation-of-gaussian-naive-bayes-in-python-from-scratch-c4ea64e3944d

[6].http://www.cs.uoi.gr/tech_reports//publications/MT-2012-01.pdf

[7].Βιβλίο “Αναγνώριση Προτύπων” Sergios Theodoridis & Konstantinos Koutroumbas σελίδες 486-487.