ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΈΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

«Μελέτη αναγνώρισης προτύπων σε αστρονομικά δεδομένα»

Συγγραφέας: Αγλαϊα-Ελένη Σκούλλου-Λάμπρου

Επιβλέπων καθηγητής: Κωνσταντίνος Κουσουρής, επίκουρος καθηγητής Ε.Μ.Π

Αθήνα, Σεπτέμβρης 2018

ΕΥΧΑΡΙΣΤΙΕΣ

Με την εκπόνηση της παρούσας διπλωματικής ολοκληρώνονται οι σπουδές μου στο 5ετές πρόγραμμα της Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Εθνικού Μετσόβιου Πολυτεχνείου.

Στις σπουδές μου ήταν καθοριστική η συμβολή των καθηγητών μου στα γνωστικά αντικείμενα που παρακολούθησα, στους οποίους οφείλω να εκφράσω τις ειλικρινείς μου ευχαριστίες για τη συμβολή τους στην ολοκλήρωση των σπουδών μου.

Ιδιαίτερα επιθυμώ να ευχαριστήσω τον καθηγητή μου και επιβλέποντα την παρούσα διπλωματική εργασία, κο Κωνσταντίνο Κουσουρή, για την επιστημονική και συμβουλευτική καθοδήγηση που μου προσέφερε σε όλα τα στάδια εκπόνησης της εργασίας με τις εύστοχες και πολύ εποικοδομητικές παρατηρήσεις του, καθώς και την ολική του υποστήριξη στο τελευταίο έτος των σπουδών μου.

Οφείλω να εκφράσω τις ευχαριστίες μου προς συναδέλφους μου με τους οποίους υπήρχε αλληλοϋποστήριξη ως προς την ολοκλήρωση των σπουδών μας.

Τέλος, οφείλω να ευχαριστήσω την οικογένειά μου, για τη συμπαράσταση και την υπομονή τους και ιδιαίτερα τον κ. Αντώνη Κωστόπουλο για την ανένδοτη και ειλικρινή υποστήριξη του ως προς την διεκπεραίωση των σπουδών μου .

ΠΕΡΙΕΧΟΜΕΝΑ

Περίληψη

Η ανακάλυψη και αυτοματοποίηση των πάλσαρ είναι ένα σημαντικό πεδίο της ραδιοαστρονομίας. Από την ανακάλυψή τους το 1967 η μελέτη των πάλσαρ έχει αποδειχθεί πολύ χρήσιμη ποικιλορόπως καθώς μπορούν να χρησιμοποιηθούν για τη μέτρηση της κοσμικής απόστασης, την ανακάλυψη εξωηλιακών πλανητών και βαρυτικών κυμάτων, καθώς και στη μελέτη της συμπυκνωμένης ύλης. Η εξέλιξη των αστρονομικών οργάνων είχε ως αποτέλεσμα την εκθετική αύξηση του όγκου των δεδομένων και κατά συνέπεια την ανάγκη χρήσης αλγορίθμων μηχανικής μάθησης για την εξόρυξη μεγάλων αστρονομικών συνόλων δεδομένων και την αυτόματη αναγνώριση υποψηφίων πάλσαρ. Το UCI Repository είναι μία συλλογή από βάσεις δεδομένων που έχει χρησιμοποιηθεί ευρέως από μαθητές, εκπαιδευτικούς και ερευνητές ως κύρια πηγή συνόλων δεδομένων μηχανικής μάθησης. Σε αυτή τη μελέτη εφαρμόσαμε, βελτιστοποιήσαμε και συγκρίναμε πολλαπλούς αλγορίθμους μηχανικής μάθησης υπό επίβλεψη για την ταξινόμηση των υποψηφίων πάλσαρ του συνόλου δεδομένων HTRU2 του αποθετηρίου UCI. Στην ανάλυση μας χρησιμοποιήσαμε Linear Classifier (Fisher), Boosted Decision Trees (BDT), Neural Networks and Support Vector Machine (SVM) και καταφέραμε να επιτύχουμε υψηλή απόδοση με τους περισσότερους από τους αλγορίθμους με καλύτερη απόδοση με τον αλγόριθμο Boosted Decision Trees.

Abstract

The discovering and identification of pulsars is a significant research field of radio astronomy. Since their discovery in 1967 pulsars have been proven very useful in the several ways as they have been used for the measurement of cosmic distance, the discovery of extrasolar planets and gravitational waves, as well as in the study of condensed matter. The development of astronomical instruments has resulted in the exponential growth of data volumes and the need to focus on machine learning algorithms for an automatic pulsar candidate identification to mine large astronomical data sets seems imperative. UCI Repository is a collection of databases, domain theories, and data generators that has been widely used by researchers, students and educators as a primary source of machine learning data sets. In this study we applied, optimized and compared multiple supervised machine learning algorithms for the classification of pulsar candidates of the UCI Repository's HTRU2 dataset. In our analysis we used Linear Classifier (Fisher), Boosted Decision Trees (BDT), Neural Networks and Support Vector Machine (SVM) and we were able to attain high performance using most of the algorithms with better performer the Boosted Decision Trees algorithm.

Γλωσσάριο

ακτίνες γ: Ηλεκτρομαγνητικές ακτινοβολίες που αποτελούνται από φωτόνια πολύ υψηλής ενέργειας. Από τη σχέση E=hν, που συνδέει την ενέργεια (E) ενός φωτονίου με τη συχνότητα (ν) της αντίστοιχης ακτινοβολίας, συμπεραίνουμε ότι οι ακτίνες γ έχουν πολύ μεγάλες συχνότητες και πολύ μικρά μήκη κύματος. Συνήθως παράγονται κατά τις πυρηνικές αντιδράσεις και κατά την αποδιέγερση των διεγερμένων πυρήνων των ατόμων.

αστέρας νετρονίων: Αστέρας αποτελούμενος κατά κύριο λόγο από νετρόνια. Έχει εξαιρετικά μεγάλη πυκνότητα και μικρή διάμετρο -της τάξεως των 10 km. Ο σχηματισμός του είναι ένα ενδεχόμενο της τελευταίας φάσης της ζωής των αστέρων. Ο αστέρας νετρονίων εμφανίζει πολύ ισχυρό βαρυτικό και μαγνητικό πεδίο. Τα παγιδευμένα στο μαγνητικό του πεδίο ηλεκτρόνια εκπέμπουν ραδιοκύματα. Λόγω της γρήγορης περιστροφής του, η ένταση των παρατηρούμενων ραδιοκυμάτων παρουσιάζει περιοδικότητα (pulsar).

Βαρυτική κατάρρευση: Η συστολή ενός σώματος πολύ μεγάλης μάζας, που οφείλεται στις έλξεις μεταξύ των σωματιδίων που το απαρτίζουν. Η κίνηση της ύλης κατά τη βαρυτική κατάρρευση γίνεται προς το κέντρο της μάζας του σώματος που καταρρέει. Στο φαινόμενο αυτό οφείλεται κατά κύριο λόγο ο σχηματισμός των γαλαξιών, των αστέρων και των πλανητικών συστημάτων. Η βαρυτική κατάρρευση είναι βασικός δυναμικός παράγοντας στην εξέλιξη του Σύμπαντος.

Γενική Θεωρία της Σχετικότητας του Αϊνστάιν: Στη θεωρία αυτή αντιμετωπίζεται το πρόβλημα της σχέσης χώρου, χρόνου και βαρύτητας. Το βασικό της αξίωμα είναι η αρχή της ισοδυναμίας, σύμφωνα με την οποία δύο παρατηρητές, από τους οποίους ο ένας βρίσκεται μέσα σε ομογενές πεδίο βαρύτητας και ο άλλος επιταχύνεται με σταθερή επιτάχυνση, θα αντιληφθούν τα ίδια φυσικά φαινόμενα και θα διατυπώσουν τους ίδιους φυσικούς νόμους. Πειραματικά ελέγξιμες προβλέψεις της Θεωρίας είναι η καμπύλωση του φωτός, όταν διέρχεται κοντά από έναν αστέρα με ισχυρό πεδίο βαρύτητας, η μετατόπιση του περιηλίου του Ερμή και η μετατόπιση προς το ερυθρό του φάσματος της ακτινοβολίας που εκπέμπεται από πολύ απομακρυσμένα φωτεινά αντικείμενα.

διάγραμμα Hertzsprung-Russel: Παριστάνει τη σχέση μεταξύ του φασματικού τύπου -που συνδέεται άμεσα με τη θερμοκρασία- και της φωτεινότητας -που εξαρτάται από το απόλυτο μέγεθος- των αστέρων. Περίπου το 90% των αστέρων που έχουν καταγραφεί βρίσκονται σε μια ζώνη που διασχίζει διαγώνια το διάγραμμα H-R και ονομάζεται Κύρια Ακολουθία. Στην Κύρια Ακολουθία βρίσκονται οι αστέρες στους οποίους η βαρυτική κατάρρευση ισορροπείται από την πυρηνική καύση υδρογόνου. Εκτός από αυτή, συναντάμε αστέρες που διανύουν είτε τα πρώτα είτε τα τελευταία στάδια της ζωής τους. Από το διάγραμμα H-R μπορούμε να αντλήσουμε πολύ σημαντικές πληροφορίες που αφορούν την εξέλιξη των αστέρων.

διπλός αστέρας: Ένα σύστημα δύο αστέρων που βρίσκονται αρκετά κοντά, ώστε να αλληλεπιδρούν ισχυρά με βαρυτικές έλξεις. Αποτέλεσμα της αλληλεπίδρασης αυτής είναι να κινούνται σε ελλειπτικές τροχιές γύρω από το κέντρο της μάζας τους. Το φαινόμενο αυτό

είναι αρκετά συχνό στο Γαλαξία. Επειδή η απόσταση μεταξύ των αστέρων του ζεύγους είναι μικρή, δεν μπορούμε να τους διαχωρίσουμε με γυμνό μάτι ή με μικρό τηλεσκόπιο.

καινοφανής (nova): Σύμφωνα με τις σύγχρονες αντιλήψεις, ο όρος δηλώνει την απότομη και θεαματική αύξηση της λαμπρότητας ενός διπλού αστέρα που αποτελείται από ένα λευκό νάνο και έναν αστέρα που ανήκει στην Κύρια Ακολουθία του διαγράμματος H-R. Τα Βασικά σημεία του μηχανισμού που οδηγεί στην παρατηρούμενη έκρηξη του αστέρα είναι: α) η μεταφορά αστρικής ύλης προς το λευκό νάνο και β) η επιτάχυνση από το ισχυρό βαρυτικό του πεδίο και η επακόλουθη υπερθέρμανση της.

μεσοαστρική ύλη: Η ύλη που ανιχνεύεται μεταξύ των αστέρων του Γαλαξία. Αποτελείται κυρίως από υδρογόνο και αποτελεί περίπου το 10% της συνολικής γαλαξιακής μάζας.

πάλσαρ: Πηγή ραδιοκυμάτων που εκπέμπονται με τη μορφή παλμών σε πολύ κανονικά χρονικά διαστήματα. Η ένταση της ακτινοβολίας που εκπέμπεται από τα πάλσαρ μεταβάλλεται περιοδικά με το χρόνο. Είναι σχεδόν βέβαιο ότι τα πάλσαρ είναι περιστρεφόμενοι αστέρες νετρονίων που διαθέτουν ισχυρό μαγνητικό πεδίο.

τηλεσκόπιο: Διάταξη με τη βοήθεια της οποίας συλλέγουμε ακτινοβολίες μιας συγκεκριμένης περιοχής του φάσματος που εκπέμπονται από ένα ουράνιο αντικείμενο. Οι ακτινοβολίες που συλλέγονται από το τηλεσκόπιο εστιάζονται, καταγράφονται και αναλύονται με κατάλληλες διαδικασίες.

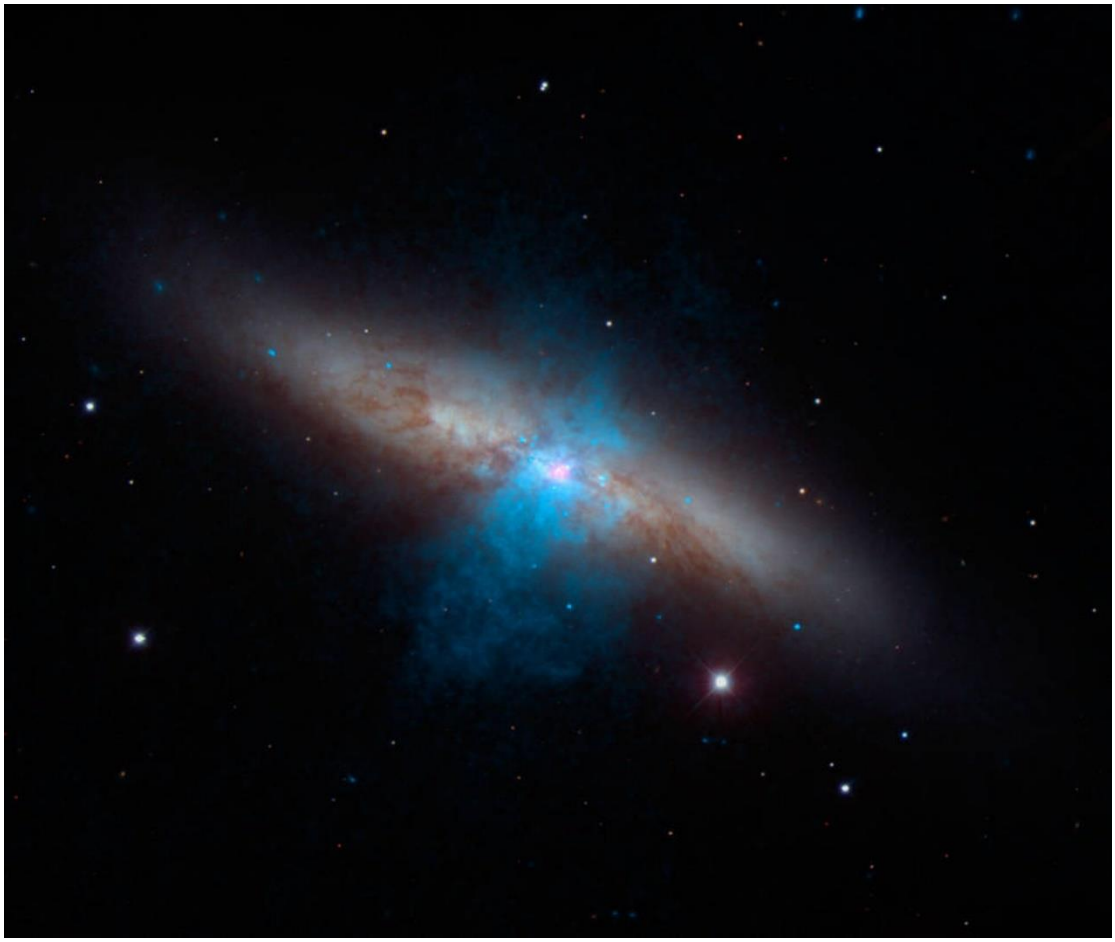
υπερκαινοφανής (supernova): Εξαιρετικά Βίαιη έκρηξη ενός αστέρα. Κατά την έκρηξη ενός υπερκαινοφανούς η λαμπρότητα του αστέρα αυξάνει αρκετές εκατοντάδες εκατομμύρια φορές. Το μεγαλύτερο μέρος της μάζας του αστέρα εκτοξεύεται προς το διάστημα με μεγάλες ταχύτητες. Με τον τρόπο αυτόν εμπλουτίζεται η μεσοαστρική ύλη με βαρέα στοιχεία (βλ. νουκλεοσύνθεση). Ο πυρήνας του αστέρα καταρρέει και καταλήγει σε έναν αστέρα νετρονίων ή μια μαύρη τρύπα.

φωτόσφαιρα: Το πρώτο στρώμα της ηλιακής ή αστρικής ατμόσφαιρας. Είναι το ορατό τμήμα της ηλιακής ή αστρικής δομής. Στη φωτόσφαιρα οφείλεται το φάσμα απορρόφησης της ακτινοβολίας που εκπέμπει ο Ήλιος. Από τη μορφή του ηλιακού φάσματος προσδιορίζεται η χημική της σύσταση. Το πάχος της φωτόσφαιρας είναι εκατοντάδες χιλιόμετρα και η θερμοκρασία της μεταβάλλεται από τους 6.000 K, που είναι κοντά στη ζώνη μεταφοράς, στους 4.000 K, όταν πλησιάζουμε τη χρωμόσφαιρα.

1. Κεφάλαιο 1: Εισαγωγή: Pulsars

1.1 Περί pulsars

Οι pulsars είναι σφαιρικά, συμπαγή αντικείμενα που έχουν μέγεθος περίπου μιας μεγάλης πόλης αλλά περιέχουν περισσότερη μάζα από τον ήλιο (Εικόνα 1.1). Οι επιστήμονες χρησιμοποιούν τους pulsars για να μελετήσουν τις ακραίες καταστάσεις της ύλης, να αναζητήσουν πλανήτες πέρα από το ηλιακό σύστημα της Γης και να μετρήσουν τις κοσμικές αποστάσεις. Οι pulsars θα μπορούσαν επίσης να βοηθήσουν στην εύρεση βαρυτικών κυμάτων και κατά συνέπεια την παρατήρηση ενεργητικών κοσμικών γεγονότων όπως οι συγκρούσεις μεταξύ υπερμεγέθων μαύρων τρυπών. Ανακαλύφθηκαν το 1967 και αποτελούν συναρπαστικά μέλη της κοσμικής κοινότητας. (Cofield, space.com, 2016)

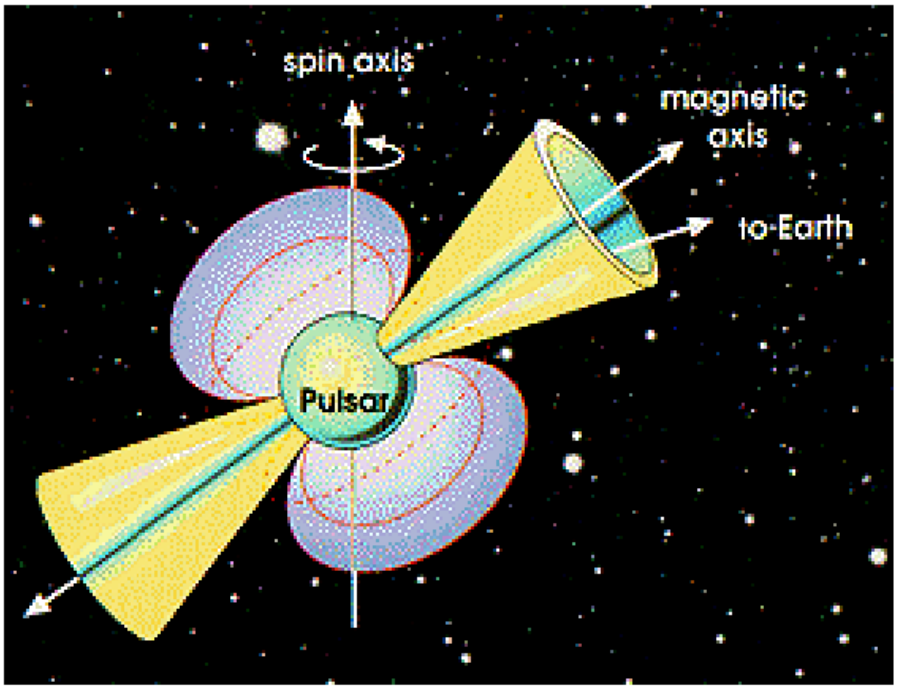


Εικόνα 0.1 Ένας pulsar (ροζ) μπορεί να φανεί στο κέντρο του γαλαξία Messier 82 σε αυτό το πορτραίτο πολλαπλών μηκών κυμάτος. Ο pulsar ανακαλύφθηκε από το NuSTAR

της NASA που ανίχνευσε την εκπομπή των ακτίνων X του pulsar. Πηγή: NASA/JPL-Caltech/SAO/NOAO

1.2 Τι είναι ένας pulsar

Οι Pulsars ακτινοβολούν δύο σταθερές, στενές δέσμες φωτός σε αντίθετες κατευθύνσεις. Παρόλο που το φως από τη δέσμη είναι σταθερό, οι pulsars φαίνεται να τρεμοπαίζουν επειδή κινούνται ταυτόχρονα. Είναι ο ίδιος λόγος που ένας φάρος φαίνεται να αναβοσβήνει όταν βλέπει κάποιος ναυτικός στον ωκεανό: Καθώς ο pulsar περιστρέφεται, η ακτίνα φωτός μπορεί να σαρώνει τη Γη, μετά να χάνεται από το οπτικό πεδίο και στη συνέχεια να ξαναεμφανίζεται. Για έναν αστρονόμο στο έδαφος, το φως εισέρχεται και εξέρχεται από την οπτικό πεδίο, δίνοντας την εντύπωση ότι ο pulsar αναβοσβήνει. Ο λόγος που μια ακτίνα φωτός του pulsar στρέφεται γύρω του σαν μια δέσμη φωτός είναι ότι η ακτίνα αυτή δεν είναι τυπικά ευθυγραμμισμένη με τον άξονα περιστροφής του pulsar (Εικόνα 1.2).



Εικόνα 0.2 Αυτό το διάγραμμα ενός pulsar δείχνει τον κίτρινο κώνο του φωτός που μπορεί να δουν οι αστρονόμοι στη Γη. Ο κώνος δεν είναι ευθυγραμμισμένος με τον άξονα περιστροφής, γι 'αυτό η δέσμη σαρώνει τον ουρανό αντί να δείχνει μόνο μία κατεύθυνση. Πηγή: Muijres, Lianne & Reisenegger, Andreas & Kuijpers, Jan. (2020). Decay modes of the magnetic field in magnetars.

Επειδή η "αναλαμπή" ενός pulsar προκαλείται από την περιστροφή του, ο ρυθμός των pulsars αποκαλύπτει επίσης τον ρυθμό περιστροφής τους. Συνολικά έχουν ανιχνευθεί πάνω από 2000 pulsars. Οι περισσότεροι από αυτούς περιστρέφονται με συχνότητα της τάξης 1 στροφή ανά δευτερόλεπτο (μερικές φορές αποκαλούνται "αργοί" pulsars), ενώ έχουν βρεθεί περισσότεροι από 200 pulsars που περιστρέφονται εκατοντάδες φορές ανά δευτερόλεπτο. Οι ταχύτεροι millisecond pulsars (χιλιοστών του δευτερολέπτου)

που έχουν ανακαλυφθεί μέχρι στιγμής μπορούν να περιστραφούν περισσότερο από 700 φορές ανά δευτερόλεπτο.

Οι pulsars δεν είναι πραγματικά αστέρια, καθώς ανήκουν σε μια οικογένεια αντικειμένων που ονομάζονται αστέρια νετρονίων που σχηματίζονται όταν ένα αστέρι με μάζα μεγαλύτερη από αυτή του ήλιου εξαντλεί τα καύσιμα στον πυρήνα του και καταρρέει. Αυτός ο αστρικός θάνατος συνήθως δημιουργεί μια τεράστια έκρηξη που ονομάζεται supernova. Το αστέρι νετρονίων είναι το πυκνό ψήγμα του υλικού που απομένει μετά από αυτόν τον εκρηκτικό θάνατο.

Τα αστέρια των νετρονίων έχουν συνήθως διάμετρο από 12,4 έως 14,9 μίλια (20 έως 24 χιλιόμετρα), αλλά μπορούν να περιέχουν μέχρι και διπλάσια από τη μάζα του ήλιου, η οποία έχει διάμετρο περίπου 1.392 εκατομμύρια χιλιόμετρα. Ένα κομμάτι μεγέθους ζάχαρης κύβου υλικού από ένα αστέρι νετρονίων θα ζυγίζει περίπου 1 δισεκατομμύριο τόνους (0,9 μετρικούς τόνους) - "περίπου το ίδιο με το όρος Everest", σύμφωνα με τη NASA. Η βαρυτική έλξη στην επιφάνεια ενός αστέρα νετρονίων θα ήταν περίπου 1 δισεκατομμύριο φορές ισχυρότερη από την βαρυτική έλξη στην επιφάνεια της Γης.

Το μόνο αντικείμενο με μεγαλύτερη πυκνότητα από ένα αστέρι νετρονίων είναι μια μαύρη τρύπα, η οποία σχηματίζεται επίσης όταν ένα αστέρι πεθαίνει. Το βαρύτερο αστέρι νετρονίων που μετρήθηκε ποτέ είναι 2.04 φορές η μάζα του ήλιου. Οι επιστήμονες δεν γνωρίζουν ακριβώς πόση μάζα μπορούν να πάρουν τα αστέρια νετρονίων πριν γίνουν μαύρες τρύπες.

Οι pulsars ως αστέρες νετρονίων είναι πολύ μαγνητικοί. Ενώ η Γη έχει ένα μαγνητικό πεδίο που είναι αρκετά ισχυρό ώστε να ασκεί μια απαλή ρυμούλκηση σε μια βελόνα πυξίδας, οι pulsars έχουν μαγνητικά πεδία που κυμαίνονται από 100 εκατομμύρια φορές έως 1 τετράκις εκατομμυριαστά (εκατομμύριο δισεκατομμύρια) φορές ισχυρότερα από τη Γη.

Για να εκπέμπει ένα αστέρι νετρονίων ως pulsar, πρέπει να έχει το σωστό συνδυασμό ισχύος μαγνητικού πεδίου και συχνότητας περιστροφής. Ορισμένα αστέρια νετρονίων μπορεί να ακτινοβολούσαν ως pulsar στο παρελθόν, αλλά όχι πλέον. Επίσης, η δέσμη ραδιοκυμάτων που εκπέμπεται από έναν pulsar μπορεί να μην περάσει από το οπτικό πεδίο ενός τηλεσκοπίου που βρίσκεται στη Γη, εμποδίζοντας τους αστρονόμους να το παρατηρήσουν.

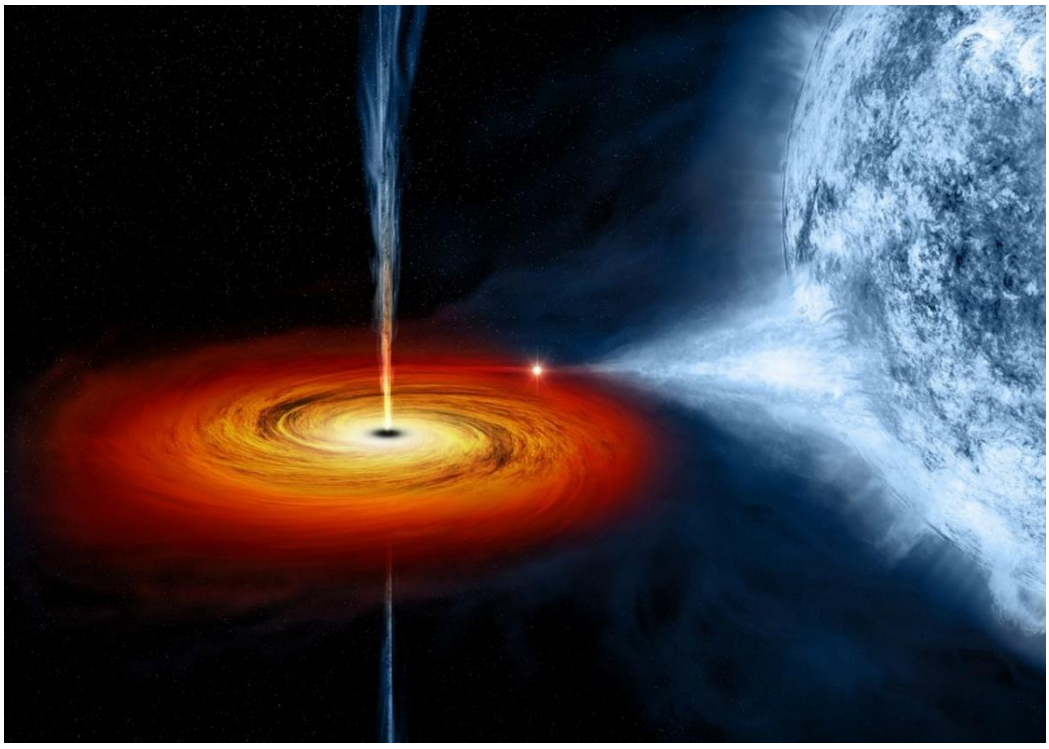
1.3 Γιατί περιστρέφονται οι pulsars

Οι πιο αργοί pulsars που ανιχνεύθηκαν ποτέ έχουν περιστροφή συχνότητας της τάξης μία στροφή ανά δευτερόλεπτο, και αυτοί συνήθως ονομάζονται αργοί pulsars. Οι γρηγορότεροι γνωστοί pulsars μπορούν να περιστρέφονται εκατοντάδες φορές ανά δευτερόλεπτο και είναι γνωστοί ως γρήγοροι pulsars ή millisecond pulsars (επειδή η περίοδος περιστροφής μετράται σε χιλιοστά του δευτερολέπτου). Περισσότερο από το ήμισυ αυτών των άστρων ταχείας περιστροφής έχουν συντροφικά αστέρια, ενώ οι βραδύτεροι ξάδελφοι τους τείνουν να εμφανίζονται μεμονωμένα. Τα υψηλά ποσοστά συντρόφων αστέρων υποδεικνύουν στους επιστήμονες ότι οι αλληλεπιδράσεις με ένα δεύτερο αστέρι μπορούν να επιταχύνουν την περιστροφή ενός φυσιολογικού pulsar.

Οι pulsars γυρίζουν επειδή τα αστέρια από τα οποία σχηματίζονται περιστρέφονται και η κατάρρευση του αστρικού υλικού θα αυξήσει φυσικά την ταχύτητα περιστροφής του pulsar. (Η προσέγγιση μάζας πιο κοντά στο κέντρο ενός περιστρεφόμενου αντικειμένου αυξάνει την ταχύτητα περιστροφής του, γι 'αυτό και οι αθλητές καλλιτεχνικού πατινάζ μπορούν να γυρίσουν πιο γρήγορα τραβώντας τα χέρια τους προς τον κορμό τους.)

Οι pulsars έχουν το μέγεθος μικρών πόλεων, οπότε η αύξηση τους σε τόσο υψηλές ταχύτητες είναι αξιοθαύμαστη. Οι millisecond pulsars απαιτούν μια πρόσθετη πηγή ενέργειας για να φτάσουν σε ένα τέτοιο υψηλό ποσοστό περιστροφής.

Οι επιστήμονες πιστεύουν ότι οι millisecond pulsars πρέπει να έχουν σχηματιστεί κλέβοντας ενέργεια από έναν συντροφικό αστέρι (Εικόνα 1.3). Ύλη και ορμή από το συντροφικό αστέρι διοχετεύεται στον pulsar αυξάνοντας σταδιακά την ταχύτητα περιστροφής του, οδηγώντας πιθανώς στην πλήρη απορρόφηση του από τον pulsar . Αυτό θα εξηγούσε γιατί έχουν ανακαλυφθεί millisecond pulsars χωρίς ορατό σύντροφο κοντά. Τα συστήματα όπου παρατηρείται ένας pulsar να απορροφά γειτονικό του αστέρι ονομάζονται black widow ή redback αστέρια από τα επονομαζόμενα επικίνδυνα είδη αραχνών.



Εικόνα 0.3 Καλλιτεχνική απεικόνιση ενός αστεριού που απορροφάται από έναν pulsar, οδηγώντας στο σχηματισμό ενός millisecond pulsar.

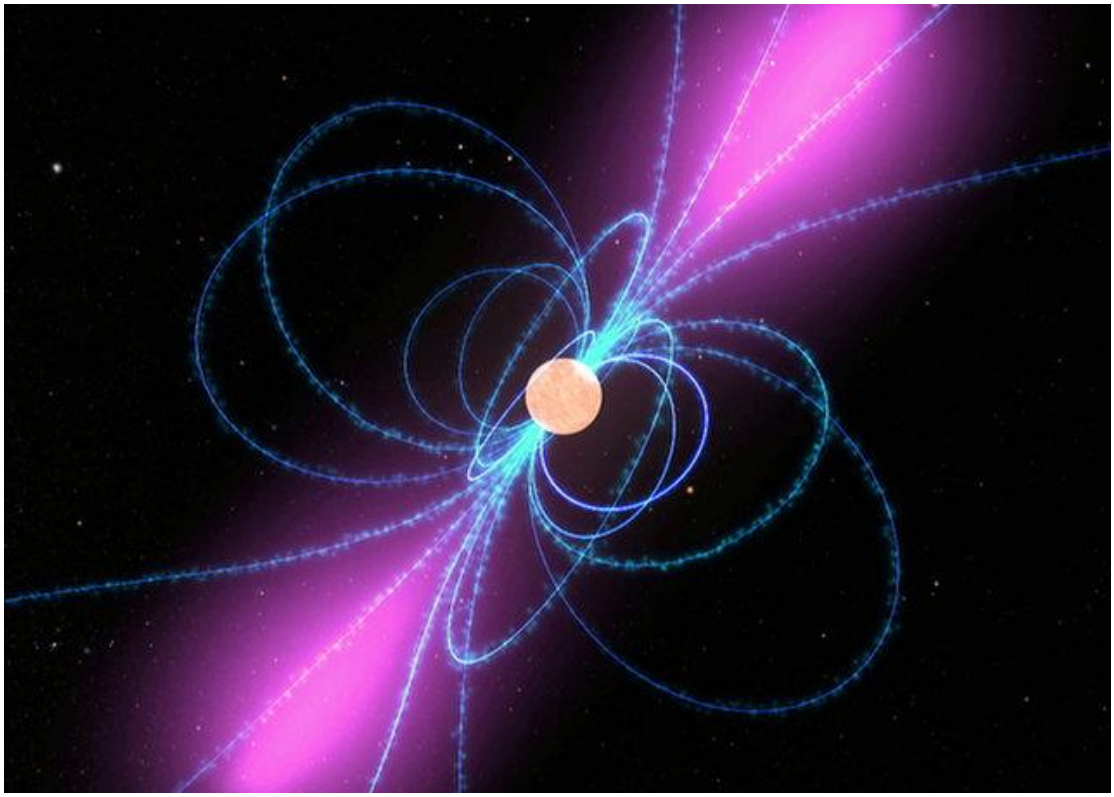
1.4 Γιατί οι pulsars εκπέμπουν ακτινοβολία

Οι pulsars μπορούν να ακτινοβολούν το φως σε πολλαπλά μήκη κύματος, από τα ραδιοκύματα μέχρι τις ακτίνες γάμα, την πιο ενεργητική μορφή του φωτός στο σύμπαν.

Οι επιστήμονες ακόμα δεν έχουν μία λεπτομερή απάντηση για το μηχανισμό ακτινοβολίας των pulsars. Έχουν διαπιστώσει ότι διάφοροι μηχανισμοί μπορεί να είναι υπεύθυνοι για την παραγωγή διαφορετικών μηκών κύματος φωτός από την περιοχή πάνω από την επιφάνεια του pulsar. Οι ακτίνες φωτός τύπου φάρου που οι επιστήμονες διαπίστωσαν για πρώτη φορά στη δεκαετία του 1960 είναι της τάξεως των ραδιοκυμάτων. Αυτές οι δέσμες φωτός είναι αξιοσημείωτες επειδή είναι εξαιρετικά φωτεινές και στενές και έχουν ιδιότητες παρόμοιες με αυτές μιας δέσμης λέιζερ. Το φως λέιζερ είναι "συνεκτικό", σε αντίθεση με το μη συνεκτικό φως που ακτινοβολείται, για παράδειγμα, από έναν λαμπτήρα. Σε μια δέσμη συνεκτικού φωτός, τα σωματίδια φωτός κινούνται μεταξύ τους με την ίδια φάση, δημιουργώντας μια ομοιόμορφη εστιασμένη δέσμη. Όταν τα σωματίδια του φωτός συνεργάζονται με αυτόν τον τρόπο, μπορούν να παράγουν μια δέσμη φωτός που είναι εκθετικά φωτεινότερη από μια πηγή διάχυτου φωτός που χρησιμοποιεί την ίδια ποσότητα ισχύος.

Αυτό που φαίνεται σαφές στους επιστήμονες είναι ότι οι εκπομπές pulsar τροφοδοτούνται από την περιστροφική του κίνηση και από το μαγνητικό του πεδίο. Οι ταχύτεροι pulsars έχουν πιο αδύναμα μαγνητικά πεδία από ό, τι οι πιο αργά περιστρεφόμενοι pulsar, αλλά η αύξηση της ταχύτητας περιστροφής εξακολουθεί να είναι αρκετή για να προκαλέσει τους γρήγορους pulsars να ακτινοβολούν παρόμοια με τους πιο αργούς pulsars.

Η παρακάτω καλλιτεχνική αναπαράσταση (Εικόνα 1.4) παρέχει μια ιδέα για το πώς οι γραμμές μαγνητικού πεδίου από ένα pulsar τυλίγονται γύρω από αυτό και συνδέονται στους δύο πόλους. Ωστόσο, στην πραγματικότητα, καθώς η περιστροφική κίνηση του pulsar «σβήνει» το μαγνητικό πεδίο, δημιουργείται μια πολύ πιο μπερδεμένη εικόνα. Ένα περιστρεφόμενο μαγνητικό πεδίο δημιουργεί ένα ηλεκτρικό πεδίο, το οποίο, με τη σειρά του, μπορεί να προκαλέσει κίνηση φορτισμένων σωματιδίων (δημιουργία ηλεκτρικού ρεύματος). Η περιοχή πάνω από την επιφάνεια του pulsar που κυριαρχείται από το μαγνητικό πεδίο ονομάζεται μαγνητόσφαιρα. Στην περιοχή αυτή, τα φορτισμένα σωματίδια, όπως τα ηλεκτρόνια και τα πρωτόνια, ή τα φορτισμένα άτομα επιταχύνονται σε εξαιρετικά υψηλές ταχύτητες από το πολύ ισχυρό ηλεκτρικό πεδίο. Κάθε φορά που φορτισμένα σωματίδια επιταχύνονται (δηλαδή είτε αυξάνουν την ταχύτητά τους είτε αλλάζουν κατεύθυνση) εκπέμπουν φως. Στη Γη, όργανα που ονομάζονται συγχροτρόνια (κυκλικοί επιταχυντές σωματιδίων) επιταχύνουν τα σωματίδια σε πολύ υψηλές ταχύτητες και χρησιμοποιούν το φως που ακτινοβολούν για επιστημονικές μελέτες. Στη μαγνητόσφαιρα του pulsar, αυτή η βασική διαδικασία μπορεί να παράγει φως στο οπτικό φάσμα και ακτίνες Χ.



Εικόνα 0.4 Καλλιτεχνική αναπαράσταση σχετικά με τις γραμμές μαγνητικού πεδίου που στροβιλίζονται γύρω από ένα pulsar. Η μωβ λάμψη αντιπροσωπεύει το φως ακτίνων γάμμα. Ο άξονας περιστροφής του παλμογράφου δεν φαίνεται και δεν ευθυγραμμίζεται με τον άξονα του μαγνητικού πεδίου. Πηγή: NASA/Goddard Space Flight Center Conceptual Image Lab

Οι παρατηρήσεις δείχνουν ότι οι ακτίνες γάμμα εκπέμπονται από μια διαφορετική θέση του περιβάλλοντα χώρου του pulsar από τις δέσμες των ραδιοκυμάτων και σε διαφορετικό υψόμετρο πάνω από την επιφάνεια. Αντί για μια στενή δέσμη, οι ακτίνες γάμμα εκπέμπονται σε σχήμα βεντάλιας. Αλλά όπως συμβαίνει και με τις εκπομπές ραδιοκυμάτων, οι επιστήμονες συζητούν ακόμη τον ακριβή μηχανισμό που είναι υπεύθυνος για τη δημιουργία ακτίνων γάμμα από έναν pulsar.

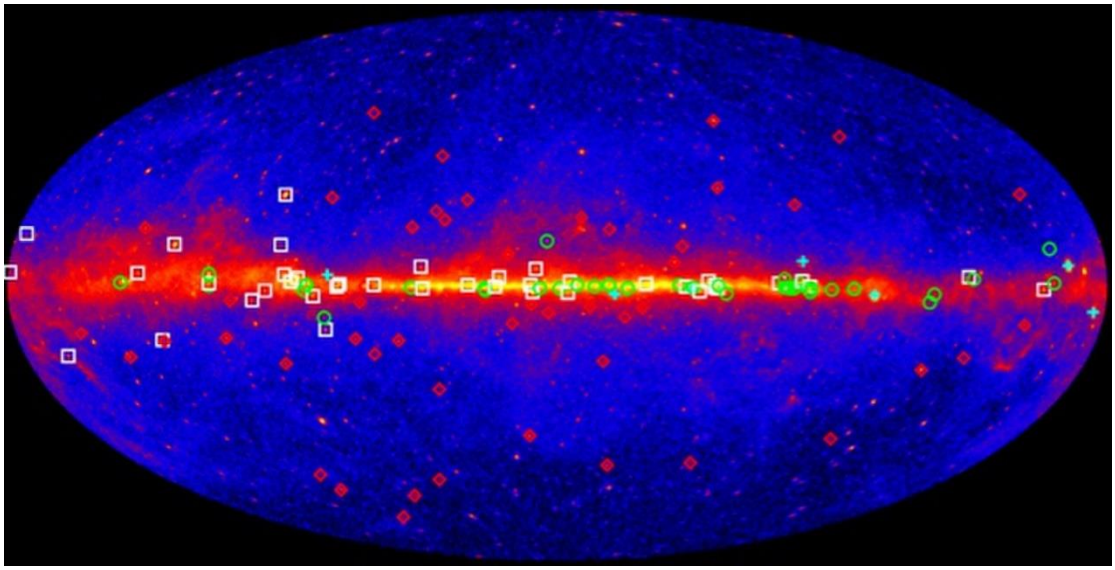
1.5 Ανακάλυψη pulsar

Τα pulsars ανακαλύφθηκαν με χρήση ραδιοτηλεσκοπίων, τα οποία εξακολουθούν να είναι το κύριο μέσο παρατήρησης και ανακάλυψης τους (Εικόνα 1.5).

Επειδή οι pulsars είναι μικροί και χαμηλής ορατότητας σε σύγκριση με πολλά άλλα ουράνια αντικείμενα, οι επιστήμονες διεξάγουν διερευνήσεις όλο το μήκος του ουρανού για την ανακάλυψη τους. Ένα τηλεσκόπιο σαρώνει ολόκληρο τον ουρανό και με την πάροδο του χρόνου οι επιστήμονες μπορούν να αναζητήσουν αντικείμενα που αναβοσβήνουν. Το ραδιοτηλεσκόπιο Parkes στην Αυστραλία έχει βρει την πλειονότητα των γνωστών pulsars. Άλλα τηλεσκόπια που έχουν συμβάλει σημαντικά στις έρευνες του pulsar είναι το ραδιοτηλεσκόπιο Arecibo στο Πουέρτο Ρίκο, το Green Bank Telescope στη Δυτική Βιρτζίνια, το τηλεσκόπιο Molonglo στην Αυστραλία και το τηλεσκόπιο της Jodrell Bank στην Αγγλία.

Χιλιάδες νέα pulsars μπορούν να ανιχνευθούν από δύο νέα τηλεσκόπια ραδιοτηλεοπτικών ερευνών, σύμφωνα με τον Scott Ransom, αστρονόμο του προσωπικού στο Εθνικό Παρατηρητήριο Αστρονομίας (NRAO) στο Charlottesville της Βιρτζίνια. Τα τηλεσκόπια είναι το Five Hundred Meter Aperture Spherical Telescope (FAST) της Κίνας, του οποίου η κατασκευή ολοκληρώθηκε το Σεπτέμβρη του 2016 και μέχρι τον Οκτώβρη του 2017 έχουν ήδη ανακαλυφθεί συνολικά 9 pulsars και το Square Kilometer Array (SKA), που χρηματοδοτείται από μια κοινοπραξία χωρών. Η κατασκευή στην SKA πρόκειται να αρχίσει το 2019, με τοποθεσίες τόσο στη Νότια Αφρική όσο και στην Αυστραλία. Ο ιστότοπος του οργανισμού λέει ότι οι πρώτες επιστημονικές παρατηρήσεις θα μπορούσαν να ξεκινήσουν το 2020, αλλά το συνολικό εγχείρημα δεν θα φθάσει σε πλήρεις επιστημονικές λειτουργίες μέχρι το 2030.

Το διαστημικό τηλεσκόπιο Fermi Gamma-ray, που ξεκίνησε τον Ιούνιο του 2008, ανίχνευσε 2.050 pulsars που εκπέμπουν ακτινοβολίες γάμμα, συμπεριλαμβανομένων των 93 millisecond pulsars. Το Fermi αποδείχτηκε ιδιαίτερα χρήσιμο καθώς σαρώνει ολόκληρο τον ουρανό, ενώ οι περισσότερες έρευνες με βάση ραδιοτηλεσκόπια τυπικά σαρώνουν μόνο τμήματα του ουρανού κατά μήκος του επιπέδου του Γαλαξία μας.



Εικόνα 0.5 Ένας χάρτης του ουρανού που δείχνει pulsars ακτίνων γάμμα που ανιχνεύονται με το όργανο LAT στο τηλεσκόπιο Ray Ray Gamma. Τα παραπάνω δείχνουν οι pulsars ακτίνων γάμμα που ανιχνεύθηκαν με τα LAT: CGRO PSRs (συν), τα νεαρά ραδιοεπιλεγμένα (κύκλος), τα νεαρά επιλεγμένα γάμμα (τετράγωνα) και τα MSPs (διαμάντι). Πηγή: NASA/DOE/Fermi LAT Collaboration

Η ανίχνευση διαφορετικών μηκών κύματος φωτός από έναν pulsar μπορεί να είναι δύσκολη. Μια ακτίνα ραδιενέργειας του pulsar μπορεί να είναι πολύ ισχυρή, αλλά αν δεν σαρώνει τη Γη (και να εισέρχεται στο οπτικό πεδίο του τηλεσκοπίου), οι αστρονόμοι μπορεί να μην την παρατηρήσουν. Μια εκπομπή ακτίνων-γ από έναν pulsar μπορεί να σαρώνει μια ευρύτερη περιοχή του ουρανού, αλλά μπορεί επίσης να είναι πιο χαμηλή και πιο δύσκολη στην ανίχνευση.

Σήμερα οι επιστήμονες γνωρίζουν περίπου 2.300 pulsars για τους οποίους έχουν εντοπιστεί μόνο ραδιοκύματα και περίπου 160 pulsars που ακτινοβολούν ακτίνες γάμμα. Οι επιστήμονες σήμερα γνωρίζουν 240 millisecond pulsars, εκ των οποίων 60 ακτινοβολούν ακτίνες γάμμα, δήλωσε ο Ransom. Αυτοί οι αριθμοί αλλάζουν συχνά καθώς ανακαλύπτονται νέοι pulsar.

1.6 Χρησιμότητα των pulsar

Τα pulsars είναι εξαιρετικά κοσμικά εργαλεία για τη μελέτη ενός ευρέος φάσματος φαινομένων.

Το φως που εκπέμπεται από έναν pulsar φέρει πληροφορίες για αυτά τα αντικείμενα και τι συμβαίνει στο εσωτερικό τους (Εικόνα 1.6). Αυτό σημαίνει ότι οι pulsars δίνουν στους επιστήμονες πληροφορίες για τη φυσική των αστέρων νετρονίων, που είναι το πυκνότερο υλικό στο σύμπαν (με εξαίρεση ό, τι συμβαίνει στην ύλη μέσα σε μια μαύρη τρύπα). Υπό τέτοια απίστευτη πίεση η ύλη συμπεριφέρεται με τρόπους που δεν παρατηρείται ποτέ σε κανένα άλλο περιβάλλον του σύμπαντος. Η περίεργη κατάσταση της ύλης μέσα στα αστέρια νετρονίων είναι αυτή που οι επιστήμονες ονομάζουν «nuclear pasta»: Μερικές φορές, τα άτομα κανονίζονται σε επίπεδα φύλλα, όπως τα lasagna ή σπείρες σαν fusilli, ή μικρά κομματάκια όπως τα gnocchi.

Μερικοί pulsars αποδεικνύονται εξαιρετικά χρήσιμοι λόγω της ακρίβειας των παλμών τους. Υπάρχουν πολλοί γνωστοί pulsars που αναβοσβήνουν με τέτοια ακρίβεια που θεωρούνται τα πιο ακριβή φυσικά ρολόγια στο σύμπαν. Ως αποτέλεσμα, οι επιστήμονες μπορούν να παρακολουθήσουν αλλαγές στο αναβοσβήσιμο ενός pulsar που θα μπορούσε να υποδεικνύει κάτι που συμβαίνει στο κοντινό χώρο.

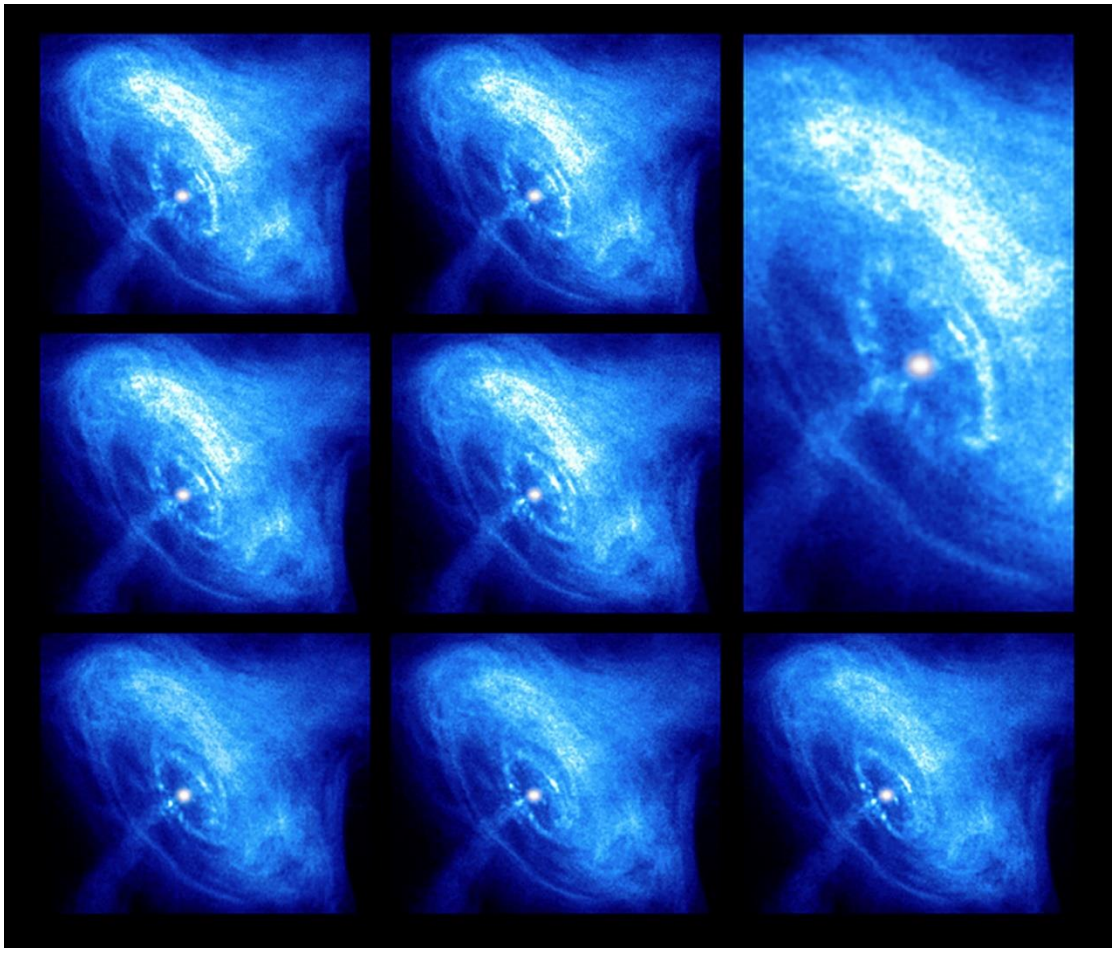
Με αυτή τη μέθοδο οι επιστήμονες άρχισαν να αναγνωρίζουν την παρουσία ξένων πλανητών που βρίσκονται σε τροχιά γύρω από αυτά τα πυκνά αντικείμενα. Ο πρώτος πλανήτης έξω από το ηλιακό σύστημα της Γης βρέθηκε να περιστρέφεται γύρω από έναν pulsar.

Επειδή οι pulsars κινούνται στο διάστημα ενώ αναβοσβήνουν τακτικά ανά δευτερόλεπτο, οι επιστήμονες μπορούν να χρησιμοποιήσουν πολλούς pulsars για να υπολογίσουν τις κοσμικές αποστάσεις. Η μεταβαλλόμενη θέση του pulsar σημαίνει ότι το φως που εκπέμπει χρειάζεται περισσότερο ή λιγότερο χρόνο για να φτάσει στη Γη. Χάρη στην εξαιρετική χρονική ακρίβεια των pulsars, οι επιστήμονες έχουν κάνει μερικές από τις πιο ακριβείς μετρήσεις απόστασης κοσμικών αντικειμένων.

Οι pulsars έχουν χρησιμοποιηθεί για να ελέγξουν πτυχές της θεωρίας της γενικής σχετικότητας του Albert Einstein, όπως τη θεωρία παγκόσμιας έλξης.

Ο κανονικός συγχρονισμός των pulsars μπορεί επίσης να διαταραχθεί από τα βαρυτικά κύματα - οι κυματισμοί στο διάστημα που προβλέπονται από τον Αϊνστάιν και εντοπίστηκαν άμεσα για πρώτη φορά τον Φεβρουάριο του 2016. Υπάρχουν πολλά πειράματα σήμερα που αναζητούν βαρυτικά κύματα μέσω της μεθόδου με pulsar.

Η χρήση pulsar για τέτοιου τύπου εφαρμογές εξαρτάται από το πόσο σταθεροί είναι στην περιστροφή τους (παρέχοντας έτσι πολύ τακτές αναλαμπές). Όλοι οι pulsars επιβραδύνονται σταδιακά καθώς περιστρέφονται, αλλά εκείνοι που χρησιμοποιούνται για μετρήσεις ακριβείας επιβραδύνουν σε ένα εξαιρετικά αργό ρυθμό, έτσι ώστε οι επιστήμονες μπορούν να τους χρησιμοποιούν ως έμπιστες συσκευές μέτρησης χρόνου.



Εικόνα 0.6 Αυτές οι εικόνες του Crab Pulsar που καταγράφηκαν από το Παρατηρητήριο ακτίνων X Chandra στη διάρκεια αρκετών μηνών δείχνουν τον έντονο λευκό pulsar στο κέντρο και πίδακες εκπεμπόμενης ύλης. Πηγή: NASA/CXC/ASU/J.Hester et al.

1.7 Νεκροταφεία pulsar

Όλοι οι pulsars επιβραδύνονται σταδιακά με την ηλικία τους. Η ακτινοβολία που εκπέμπεται από έναν pulsar τροφοδοτείται από το μαγνητικό πεδίο και την περιστροφικότητα του. Ως αποτέλεσμα, ένας pulsar που επιβραδύνει επίσης χάνει ισχύ και σταματά σταδιακά να εκπέμπει ακτινοβολία (ή τουλάχιστον σταματά να εκπέμπει αρκετή ακτινοβολία για να είναι δυνατός ο εντοπισμός τους από τηλεσκόπια). Παρατηρήσεις μέχρι τώρα δείχνουν ότι οι pulsars ακτίνων γάμμα πέφτουν κάτω από

το όριο ανίχνευσης πριν από τους pulsar ραδιοκυμάτων. Το στάδιο αυτό της ζωής των pulsars είναι γνωστό και ως νεκροταφείο pulsar. (Οι pulsars που έχουν σταματήσει να εκπέμπουν μπορούν να θεωρηθούν συνηθισμένα αστέρια από τους αστρονόμους).

Όταν ένα pulsar σχηματίζεται από τα συντρίμμια μιας έκρηξης υπερκαινοφανούς τύπου γυρίζει γρήγορα και ακτινοβολεί πολλή ενέργεια. Το καλά μελετημένο Crab Pulsar είναι ένα παράδειγμα ενός τέτοιου νεαρού pulsar. Αυτή η φάση μπορεί να διαρκέσει μερικές εκατοντάδες χιλιάδες χρόνια, μετά την οποία ο pulsar αρχίζει να επιβραδύνεται και εκπέμπει μόνο ραδιοκύματα. Αυτοί οι pulsars "μέσης ηλικίας" αποτελούν πιθανώς το μεγαλύτερο μέρος του πληθυσμού των pulsars που εκπέμπουν μόνο ραδιοκύματα. Αυτοί οι pulsars ζουν για δεκάδες εκατομμύρια χρόνια, πριν τελικά επιβραδύνουν τόσο πολύ που «σβήνουν» και εισέρχονται στο αστρικό νεκροταφείο pulsar.

Αν όμως ο pulsar βρίσκεται κοντά σε ένα γειτονικό αστέρι μπορεί να αναγεννηθεί. Απορροφώντας ύλη και ενέργεια από το γειτονικό αστέρι αυξάνεται η συχνότητα του σε εκατοντάδες φορές ανά δευτερόλεπτο - δημιουργώντας έτσι έναν millisecond pulsar. Αυτή η αλλαγή μπορεί να συμβεί σε οποιοδήποτε στάδιο της ζωής του pulsar που σημαίνει ότι ο ρυθμός περιστροφής ενός pulsar που πεθαίνει μπορεί να αυξηθεί σε χρονικό διάστημα εκατοντάδων έως και εκατομμυρίων χρόνων. Ο pulsar αρχίζει να εκπέμπει ακτινοβολία X και το ζευγάρι των αντικειμένων είναι γνωστό ως "δυαδικό σύστημα ακτίνων X χαμηλής μάζας". (Αυτοί οι κανιβαλιστικοί pulsars έχουν ονομαστεί "black widow" ή "redback" pulsars από δύο είδη αράχνης που είναι γνωστό ότι σκοτώνουν τους συντρόφους τους). Οι millisecond pulsars είναι οι γηραιότεροι γνωστοί pulsars - μερικοί είναι δισεκατομμυρίων ετών και θα συνεχίσουν να περιστρέφονται σε τέτοιες συχνότητες για δισεκατομμύρια χρόνια.

2. Κεφάλαιο 2: Αναζήτηση pulsar

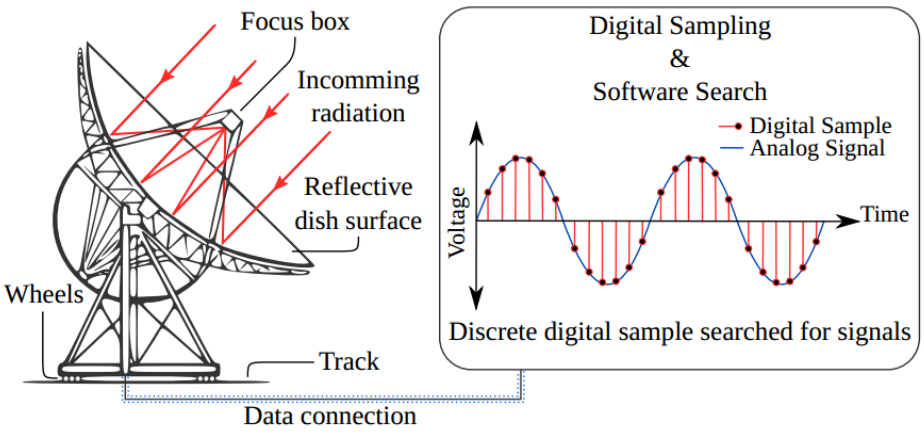
2.1 Ραδιοτηλεσκόπια

Η αναζήτηση για pulsars πραγματοποιείται μέσω διερευνήσεων αφιερωμένων για το συγκεκριμένο σκοπό. Υπάρχουν δύο ειδών γενικών στρατηγικών διερεύνησης: οι στοχευμένες και οι τυφλές διερευνήσεις. Οι στοχευμένες παρατηρήσεις διερευνούν μια συγκεκριμένη τοποθεσία ή μέρος του ουρανού για σήματα από pulsars. Αυτός ο τύπος παρατήρησης επικεντρώνεται σε περιοχές του ουρανού όπου είναι πιο πιθανό να βρεθούν pulsars, όπως πυκνές αστρικές περιοχές σαν GC, είναι σχετικά μικρότερης κλίμακας και απαιτούν λιγότερους πόρους εν σύγκριση με τις “τυφλές” παρατηρήσεις. Οι “τυφλές” παρατηρήσεις είναι οι πιο συνηθισμένες και διερευνούν πολλαπλές περιοχές του ουρανού με γρήγορη διαδοχή με την ελπίδα εύρεσης εκπομπών pulsars. Αυτού του είδους οι παρατηρήσεις είναι απαραίτητες δεδομένου του γεγονότος ότι τα pulsars μπορούν να βρεθούν σε ολόκληρο το εύρος του ουρανού και ότι θέσεις τους αδύνατο να προβλεφθούν. Οι “τυφλές” παρατηρήσεις διεξάγονται από το 1968 [Large

M.I., 1968] και έχουν ανακαλύψει τα περισσότερα από τα 2,524 γνωστά pulsars [Manchester R.N. H. G., The Australia TelescopeNational Facility Pulsar Catalogue, 2005a] [Manchester R.N. H. G., The Australia TelescopeNational Facility Pulsar Catalogue, 2005b]. Παρά ταύτα η παρατήρηση σημάτων από pulsars με αυτή τη μέθοδο είναι εκ φύσεως δύσκολη καθώς η εκπομπή τους είναι πολύ ασθενής όταν φτάνει στη Γη [Lorimer, 2008]. Περιβαλλοντικοί παράγοντες περιπλέκουν περισσότερο τη διερεύνηση αφού ελαττώνουν περαιτέρω την ευαισθησία μας στην παρατήρηση των εκπομπών τους με διάφορους τρόπους, όπως σπινθηροβολία και διασπορά. (Lorimer D.R. and Kramer M., 2005) Ταυτοχρόνως RFI που παράγεται από τη μοντέρνα τεχνολογία καμουφλαρει τα σήματα από pulsar εξαιτίας της κοντινής εγγύτητας και της σχετικής ενέργειας. (Lorimer D.R. and Kramer M., 2005)

Η εύρεση pulsar απαιτεί επομένως δύο βασικά εργαλεία αναζήτησης. Ένα μεγάλο ραδιοτηλεσκόπιο που να είναι ευαίσθητο σε εξαιρετικά αδύναμες εκπομπές και έναν αγωγό αναζήτησης σήματος ικανό να απομονώνει αδύναμα σήματα όπως και όποτε εμφανίζονται εν μέσω RFI και θορύβου.

Τα περισσότερα ραδιοτηλεσκόπια είναι μεγάλες παραβολικές κεραίες που μοιάζουν με τεράστια "πιάτα", όπως και οι κεραίες που χρησιμοποιούνται για τη λήψη των ραδιοκυμάτων της δορυφορικής τηλεόρασης σε πολύ μεγαλύτερη κλίμακα. Το "πιάτο" ενός ραδιοτηλεσκοπίου έχει συχνά μήκος πολλών μέτρων. Με χρήση τέτοιας μεγάλης επιφάνειας μπορεί να ανιχνευθεί ασθενής ακτινοβολία που φτάνει στην επιφάνεια της Γης. Όσο μεγαλύτερη είναι η κεραία τόσο πιο πολύ μεγαλώνει η ευαισθησία στην ανίχνευση πιο αδύναμων σημάτων. Το παραβολικό σχήμα του πιάτου βοηθά στην εστίαση ασθενών σημάτων, αντανακλώντας τα σε ένα στοιχείο γνωστό ως δέκτη. Ο δέκτης στεγάζεται σε μια προστατευτική θήκη όπου η ανακλώμενη ακτινοβολία εστιάζεται από το πιάτο. Η θήκη αυτή ονομάζεται και κουτί ή καμπίνα εστίασης, και απεικονίζεται σαφέστερα στην εικόνα 2.1.



Εικόνα 0.1 Διάγραμμα ενός ραδιοτηλεσκοπίου με κινητό ανακλαστήρα. Αναλογικά σήματα που λαμβάνονται από το πιάτο αντανακλώνται προς το κουτί εστίασης, όπου αρχίζει η επεξεργασία σήματος. Τα αναλογικά σήματα μετατρέπονται σε ψηφιακά σήματα

πριν οι αγωγοί λογισμικού ξεκινήσουν την αναζήτηση παρουσίας περιοδικότητας. Εμπνευσμένο από μια εικόνα που παρέχεται από το Max-Planck-Institut για την Αστρονομία (1998, no author credited).

Μέσα στο κουτί εστίασης ο δέκτης συχνά κρυογονικά ψύχεται στους περίπου -260 ◦ C (Lorimer D.R. and Kramer M., 2005). Αυτό βοηθά στην αποφυγή κάλυψης επιθυμητών σημάτων από το θερμικό θόρυβο που παράγεται από τα ηλεκτρονικά κυκλώματα των ίδιων των αποδεκτών. Το σύστημα του αποδέκτη αποτελείται από διάφορα στοιχεία. Ένα "κέρας τροφοδοσίας" αρχικά διοχετεύει την ανακλώμενη εκπομπή στον αποδέκτη. Η κεραία τροφοδοσίας βοηθά στην καθοδήγηση της ληφθείσας εκπομπής, καθιστώντας το ικανό να έχει ευαισθησία εστίασης σε ευρύτερο φάσμα συχνοτήτων. Η κατευθυνόμενη εκπομπή φθάνει τελικά στον δέκτη, όπου το ηλεκτρομαγνητικό κύμα που το διαπερνά προκαλεί ρεύματα στο στοιχείο λήψης. Στα σύγχρονα τηλεσκόπια πολλαπλά κέρατα τροφοδοσίας συνήθως συγκεντρώνονται μαζί, έτσι ώστε κάθε μεμονωμένη τροφοδοσία συλλέγει την ακτινοβολία από ένα παρακείμενο τμήμα του ουρανού χωρίς αλληλοεπικάλυψη (Lorimer D.R. and Kramer M., 2005). Αυτό επιτρέπει σε ένα μοναδικό σύστημα αποδέκτη να συλλέγει ακτινοβολία από μία μεγαλύτερη περιοχή του ουρανού, επιτρέποντας ταχύτερες αναζητήσεις σε σχέση με την κάλυψη του ουρανού. Τα τηλεσκόπια με πολλαπλά κέρατα τροφοδοσίας περιγράφονται ως αποδέκτες πολλαπλών ακτίνων, όπου κάθε δέσμη είναι ανάλογη με ένα ξεχωριστό 'αυτί' στον ουρανό.

Τα ρεύματα που προκαλούνται στον δέκτη παράγουν τάσεις που ενισχύονται με έναν ενισχυτή χαμηλού θορύβου. Εκεί η εκπομπή ενισχύεται εκατομμύρια φορές λόγω της αδυναμίας των αστρονομικών σημάτων. Το ενισχυμένο σήμα περνά από ένα φίλτρο διέλευσης ζώνης. Το φίλτρο επιτρέπει να διαπεράσουν σήματα συχνοτήτων που μας ενδιαφέρουν και φιλτράρει τα υπόλοιπα. Αυτό επιτρέπει τις περισσότερες παρεμβολές που είναι γνωστό ότι συμβαίνουν σε συγκεκριμένες συχνότητες (λόγω γραμμών ηλεκτρικού ρεύματος, δικτύων κινητής τηλεφωνίας) να αφαιρεθούν. Τα παραμένοντα σήματα μετατρέπονται σε σήματα χαμηλότερης συχνότητας ενδιάμεσου σήματος (Intermediate Frequency, IF) χρησιμοποιώντας έναν ανάμικτη. Αυτό το IF επιτρέπει την πιο αποτελεσματική μετάδοση του σήματος μέσω της καλωδίωσης εντός του κουτιού εστίασης με μικρή απώλεια σήματος. Το IF αναγνωρίζεται από ψηφιακές τράπεζες φίλτρων, δημιουργώντας τελικά ένα ψηφιακό δείγμα ενός αναλογικού σήματος (Lorimer D.R. and Kramer M., 2005).

2.2 Αγωγοί Λογισμικού

Μια ψηφιακή τράπεζα φίλτρου είναι μια σειρά από φίλτρα ζώνης-διέλευσης, που χρησιμοποιείται για να χωρίσει το προ-επεξεργασμένο ψηφιακό σήμα σε έναν αριθμό καναλιών συχνότητας για επεξεργασία λογισμικού. Χρησιμοποιώντας τράπεζες φίλτρων, ένα λαμβανόμενο σήμα χωρίζεται σε κανάλια συχνότητας, το καθένα πλάτους

Δν kHz. Καθένα μεμονωμένο κανάλι περιέχει s_{tot} δείγματα σήματος που λαμβάνονται στο διάστημα t_{samp}(μs), σε μια περίοδο παρατήρησης διάρκειας t_{obs} sec, έτσι ώστε,

$$s_{tot} = \frac{t_{obs}}{t_{samp}}. \qquad (0.1)$$

Έτσι κάθε ξεχωριστή παρατήρηση που γίνεται από το τηλεσκόπιο απεικονίζεται ως ένας πίνακας n_{chans} x s_{tot} όπως φαίνεται από την εξίσωση (2.2)

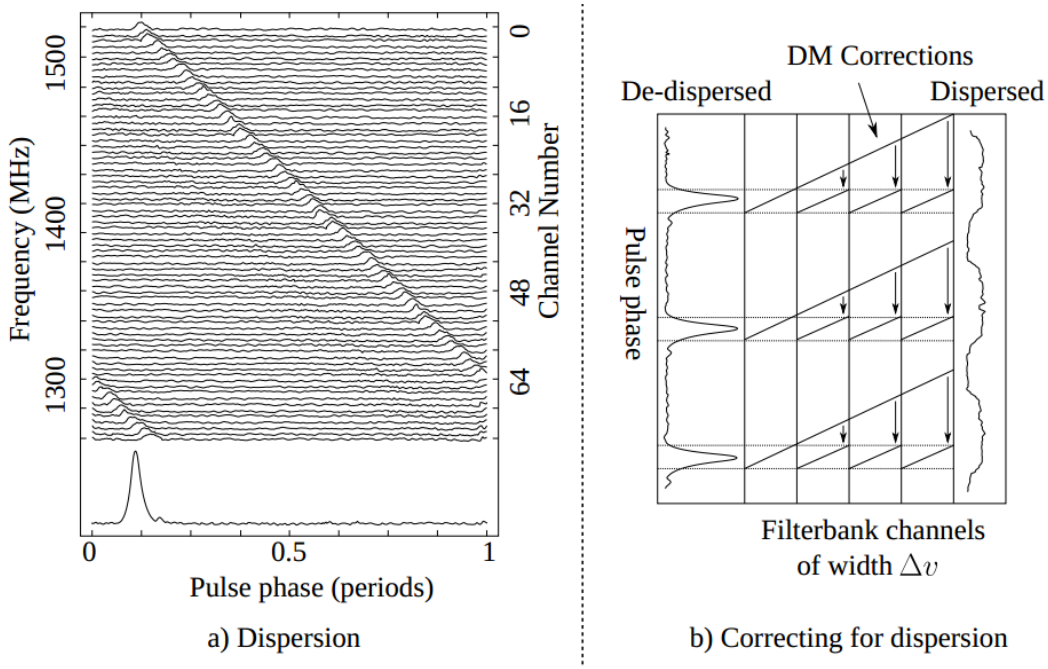
$$M = \begin{bmatrix} a_{1,1} & \cdots & a_{1,s_{tot}} \\ a_{2,1} & \cdots & a_{2,s_{tot}} \\ \vdots & \ddots & \vdots \\ a_{n_{chans},1} & \cdots & a_{n_{chans},s_{tot}} \end{bmatrix}. \qquad (0.2)$$

Μια λογισμική αναζήτηση για σήματα pulsar περιλαμβάνει μια σειρά διαδικαστικών βημάτων που εφαρμόζονται στα δεδομένα του M. Το πρώτο από αυτά περιλαμβάνει αφαίρεση RFI, μέσω της αφαίρεσης των καναλιών (σειρές του πίνακα) που αντιστοιχούν σε γνωστές συχνότητες παρεμβολών που δεν έχουν προηγουμένως αφαιρεθεί [Keith M.J., 2010]. Μόλις οι συχνότητες παρεμβολών έχουν αφαιρεθεί μπορεί να εφαρμοστεί 'Clipping' [Hogden J., 2012] στα δεδομένα, που στοχεύει στη μείωση του αντίκτυπου της παρορμητικής παρεμβολής. Αυτό επιτυγχάνεται ορίζοντας ως μηδέν (ή στον τοπικό μέσο όρο) τα δείγματα του M, τα οποία εμφανίζουν εντάσεις υψηλότερες από κάποιο προκαθορισμένο κατώφλι σε μια δεδομένη στήλη (π.χ.ένταση 2σ πάνω από τον μέσο όρο). Μόλις ολοκληρωθούν αυτά τα αρχικά βήματα, ο αρχικώς πίνακας M μετατρέπεται σε τροποποιημένο πίνακα M_0 έτοιμο για περαιτέρω ανάλυση. Έπειτα η επεξεργασία εισέρχεται σε μια υπολογιστικά δαπανηρή φάση γνωστή ως de-dispersion. Όπως αναφέρθηκε προηγουμένως, τα σήματα που ταξιδεύουν μέσω του ISM επηρεάζονται από αυτό με διάφορους τρόπους [D.R., Radio Pulsar Statisitics.", 2009]. Μέχρι στιγμής η πιο γνωστή επιρροή είναι το φαινόμενο γνωστό και ως διασπορά (Lorimer D.R. and Kramer M., 2005) [D.R., Radio Pulsar Statisitics.", 2009].

Καθώς τα σήματα pulsar μεταδίδονται μέσω του ISM προς τη Γη, αλληλοεπιδρούν με φορτισμένα σωματίδια (ελεύθερα ηλεκτρόνια) στη διαδρομή. Αυτές οι αλληλεπιδράσεις καθυστερούν την άφιξη του σήματος στη Γη (Lorimer D.R. and Kramer M., 2005) με χαρακτηριστικό τρόπο. Τα χαμηλής συχνότητας στοιχεία του σήματος καθυστερούν περισσότερο από τα αντίστοιχα υψηλής συχνότητας. Έτσι, τα στοιχεία χαμηλής συχνότητας φθάνουν στη Γη ελαφρώς πιο μετά από αυτά υψηλότερων συχνοτήτων. Αυτό προκαλεί ένα φαινόμενο διασποράς που κάνει τα σήματα πιο διασκορπισμένα με το χρόνο. Αυτό καθιστά δύσκολη την ανίχνευση παλμών, αφού οι παλμοί τους καθίστανται λιγότερο έντονοι, όπως φαίνεται στην Εικόνα 2.2. Αυτό είναι παρατηρείται ως μείωση του λόγου σήματος-θορύβου (S / N) ενός ανιχνευόμενου παλμού, ένα μέτρο που χρησιμοποιείται για να περιγράψει τη δύναμη ενός σήματος,

$$\mathrm{S/N} = \frac{P_{signal}}{P_{noise}}, \qquad (0.3)$$

όπου P_{signal} είναι η μέση ισχύς του σήματος, και P_{noise} τη μέση ισχύς του θορύβου του υποβάθρου. Εάν ο λόγος επιτύχει την τιμή ένα, τότε το σήμα δεν διακρίνεται από τον θόρυβο. Ισχυρότερα σήματα επιτυγχάνουν λόγους μεγαλύτερους από ένα, έτσι ένα υψηλό S / N δείχνει, αν και δεν αποδεικνύει, την παρουσία ενός νόμιμου σήμα.

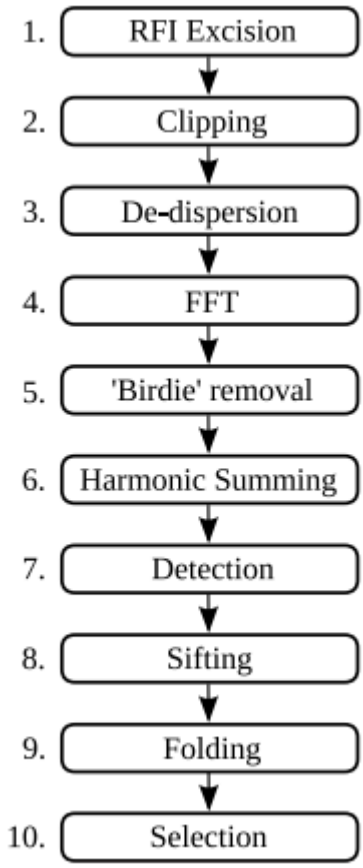


Εικόνα 0.2 Ένα παράδειγμα διασποράς σήματος. Βασίζεται σε διαγράμματα που παρουσιάστηκαν αρχικά από τους Lorimer και Kramer (2005). Plot α) δείχνει πώς διασκορπίζεται ένα σήμα με το χρόνο. Η διασπορά κρύβει το πραγματικό σχήμα παλμού και προκαλεί μείωση του ανιχνευόμενου S / N. Plot b) δείχνει την εφαρμογή διορθώσεων

DM σε διασκορπισμένο σήμα. Η διόρθωση DM είναι διαφορετική σε κάθε κανάλι συχνότητας, καθώς η διασπορά είναι ανάλογη της συχνότητα.

Το ποσοστό της διασποράς ενός σήματος που λαμβάνεται είναι ανάλογο με μια ποσότητα που ονομάζεται μέτρο διασποράς (Dispersion Measure/DM) (Lorimer D.R. and Kramer M., 2005). Το DM είναι η ολοκληρωμένη πυκνότητα των ελεύθερων ηλεκτρονίων μεταξύ ενός παρατηρητή και ενός pulsar [Lorimer, 2008]. Η πραγματική πυκνότητα στήλης, και επομένως ο ακριβής βαθμός στον οποίο ατο σήμα είναι διασκορπισμένο, δεν μπορεί να είναι γνωστό a priori [Keith M.J., 2010] [L, 2012]. Πρέπει να διεξαχθούν δοκιμές μέτρησης διασποράς ή δοκιμές DM για να καθοριστεί αυτή η τιμή με όσο το δυνατόν μεγαλύτερη ακρίβεια. Μπορεί να χρησιμοποιηθεί ένα ακριβές DM για την αναίρεση της διασποράς, επιτρέποντας τη μεγιστοποίηση του S / N του ανιχνευμένου σήματος (Lorimer D.R. and Kramer M., 2005). Για κάθε δοκιμή διασποράς, κάθε κανάλι συχνότητας (σειρά στο M_0) μετατοπίζεται με κατάλληλη καθυστέρηση. Οι επακόλουθες δοκιμές αυξάνουν την καθυστέρηση έως ότου επιτευχθεί μέγιστη τιμή DM. Αυτό το μέγιστο θα διαφέρει ανάλογα με την περιοχή του ουρανού που ερευνήθηκε, τη συχνότητα παρατηρήσεως και το εύρος ζώνης. Η διαδικασία παράγει μία χρονοσειρά ανά κανάλι συχνότητας χωρίς διασπορά. Αυτές αθροίζονται στη συνέχεια για να παράγουν μια ενιαία χρονοσειρά ανά δοκιμή δίχως διασπορά (όπως φαίνεται στην εικόνα 2.2 α). Με ολική απομάκρυνση διασποράς παράγεται ένας αριθμός χρονοσειρών ίσος με το συνολικό αριθμό δοκιμών DM.

Περιοδικά σήματα σε δεδομένα χρονοσειρών χωρίς διασπορά, μπορούν να βρεθούν χρησιμοποιώντας ανάλυση Fourier. Αυτή η διαδικασία είναι γνωστή και ως αναζήτηση περιοδικότητας (Lorimer D.R. and Kramer M., 2005). Το πρώτο βήμα της αναζήτησης περιοδικότητας συνήθως περιλαμβάνει το φιλτράρισμα των δεδομένων για την αφαίρεση ισχυρά φασματικών γνωρισμάτων γνωστών και ως «birdies» [Manchester R.N. L. A., 2001] [Hessels J.W.T., 2007]. Αυτά μπορεί να οφείλονται σε περιοδικές ή σχεδόν περιοδικές παρεμβολές. Στη συνέχεια εφαρμόζονται αθροιστικές τεχνικές, οι οποίες προσθέτουν τα εύρη συχνοτήτων που σχετίζονται αρμονικά με τα αντίστοιχα θεμελιώδη. Αυτό το βήμα είναι απαραίτητο, καθώς στο πεδίο Fourier, η ισχύς από ένα στενό παλμό κατανέμεται μεταξύ των θεμελιωδών συχνοτήτων και των αρμονικών του (Lorimer D.R. and Kramer M., 2005). Έτσι η προσπάθεια εύρεσης pulsar με χρήση μόνο των θεμελιωδών συχνοτήτων θα είχε μικρότερη πιθανότητα επιτυχίας. Περιοδικές ανιχνεύσεις με μεγάλα πλάτη Fourier μετά το άθροισμα(πάνω από το φόντο του θορύβου ή από ένα επίπεδο κατωφλίου), θεωρούνται τότε ως πιθανές περίοδοι. Μια περαιτέρω διαδικασία γνωστή ως shifting [Stovall K., 2013] εφαρμόζεται στις πιθανές περιόδους, η οποία αφαιρεί τις διπλές ανιχνεύσεις του ίδιου σήματος σε ελαφρώς διαφορετικά DM, μαζί με τις συ σχετιζόμενες αρμονικές τους. Ένας μεγάλος αριθμός από πιθανών περιόδων επιβιώνουν τη διαδικασία shifting.

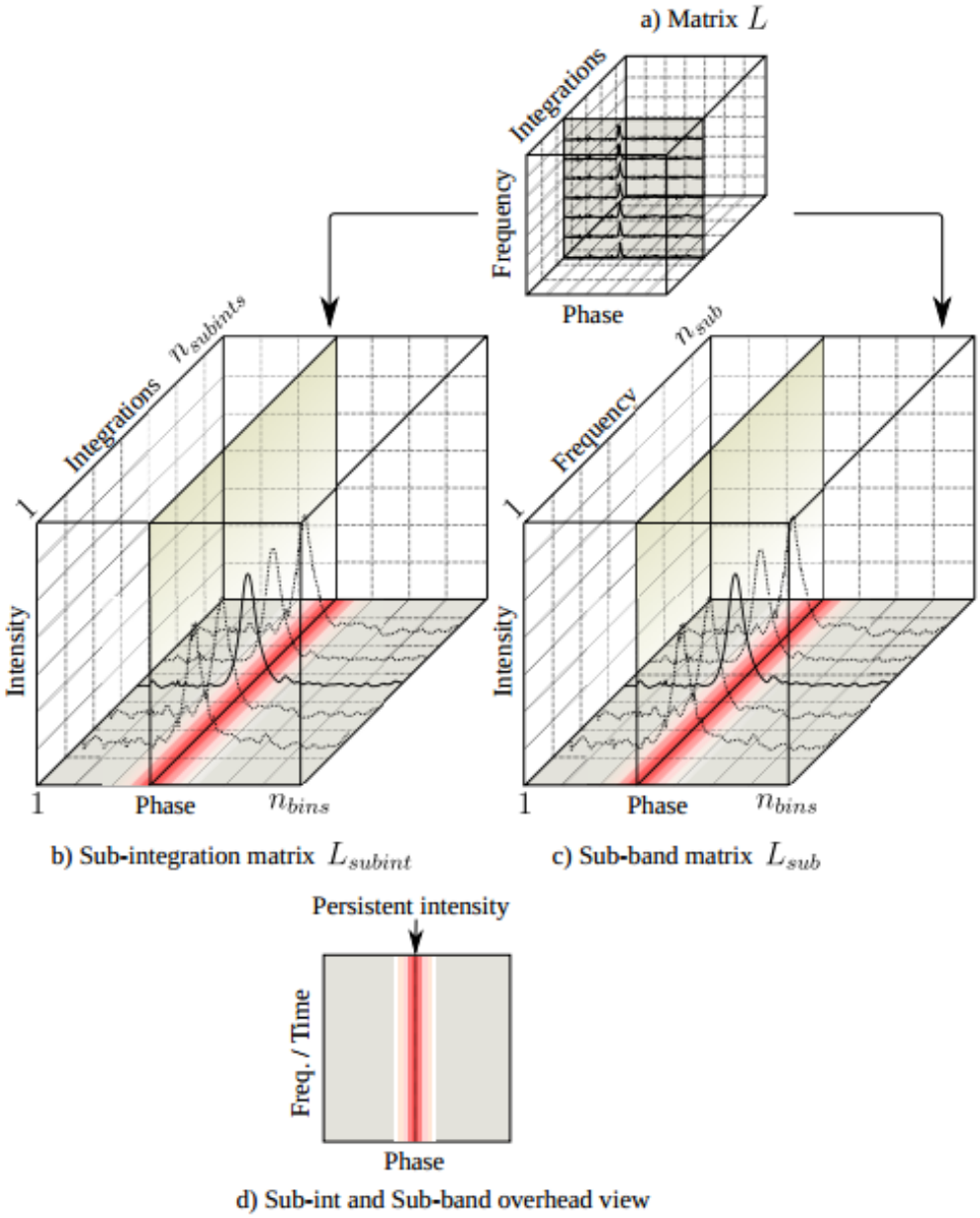


Εικόνα 0.3 Γενικά βήματα της διαδικασίας δημιουργίας υποψηφίων. [Lyon R. J., 2016]

2.3 Αναδιπλωμένα προφίλ pulsar

Τα αναδιπλωμένα υποψήφια δεδομένα ποικίλλουν ανάλογα με τον αγωγό αναζήτησης έρευνας που χρησιμοποιήθηκε. Ωστόσο, γενικά για κάθε πιθανή περίοδο, τα δεδομένα στο M_0 αναδιπλώνονται χρησιμοποιώντας την καλύτερη περίοδο και DM που βρέθηκαν, παράγοντας ένα νέο πίνακα L. Αυτός είναι ένας 3D πίνακας που περιγράφει το υποψήφιο σε φάση, συχνότητα και χρόνο. Το μέγεθος του πίνακα L είναι $n_{bins} \times n_{sub} \times n_{subint}$, όπου n_{bins} αντιστοιχεί στον αριθμό των bins σε συγκεκριμένη φάση, n_{sub} στα bins συγκεκριμένης συχνότητας και n_{subint} στις χρονικές ολοκληρώσεις. Ο πίνακας L είναι χρήσιμος, καθώς τα δεδομένα που περιέχει επιτρέπουν την απεικόνιση ενός υποψηφίου σήματος με διάφορους τρόπους, οι οποίοι μπορούν να βοηθήσουν στην αναγνώριση της πηγής τους. Στην Εικόνα 2.4 φαίνονται δύο από τις βασικές επεικονήσεις που προκύπτουν μέσω του μετασχηματισμού του L σε δύο περαιτέρω πίνακες. Το μέρος της Εικόνας 2.4 b) δείχνει τον παλμό που εντοπίστηκε στο χρόνο. Καθώς τα σήματα pulsar είναι περιοδικά, θα πρέπει να επαναλαμβάνονται σε ολόκληρη την παρατήρηση, όπως φαίνεται σε αυτό το παράδειγμα. Η Εικόνα 2.4 c) δείχνει τον ανιχνευόμενο παλμό σε σχέση με τη συχνότητα.

Καθώς η εκπομπή pulsar είναι ευρυζωνική, ένα σήμα pulsar θα πρέπει γενικά να υπάρχει σε ολόκληρο το εύρος συχνότητας όπως φαίνεται εδώ. Αυτοί οι δύο πίνακες εμφανίζουν τις βασικές ιδιότητες που θα αναμένεται από ένα πραγματικό σήμα pulsar. Αυτές οι πληροφορίες μπορούν να συνοψιστούν εύκολα στο σχήμα 2.4 d) που παρέχει μια γενική προβολή αυτών των πινάκων σε 2 διαστάσεις.



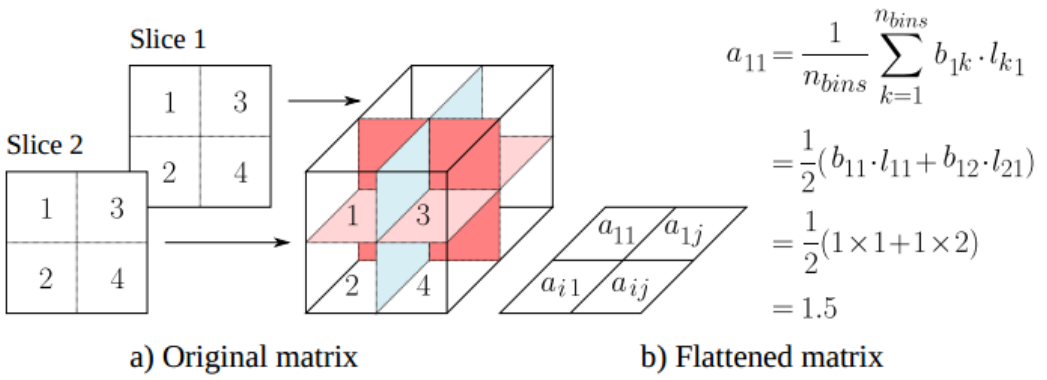
Εικόνα 0.4 Απεικόνιση των δεδομένων που είναι αποθηκευμένα σε έναν υποψήφιο pulsar. Στο a) είναι ο πίνακας που λαμβάνεται όταν τα δεδομένα παρατήρησης αναδιπλώνονται με τη βέλτιστη περίοδο και το βέλτιστο DM ενός υποψηφίου. Στο b) ο πίνακας L_{subint} περιγράφει τον ανιχνευόμενο παλμό στο χρόνο. Στο c) ο πίνακας L_{subint} περιγράφει τον παλμό που έχει ανιχνευθεί μέσω του πεδίου συχνότητας. Τέλος στο d) φαίνεται η οριζόντια προβολή των πινάκων L_{subint} και L_{sub} όταν υπάρχει ένα ισχυρό σήμα τύπου pulsar. Αυτά τα διαγράμματα μπορούν να χρησιμοποιηθούν για να δώσουν μια οπτική ένδειξη της αυθεντικότητας ενός υποψηφίου.

Η απόκτηση των επιπεδοποιημένων πινάκων F_{subint} και F_{sub} είναι απλή. Ο επιπεδοποιημένος πίνακας που προκύπτει από τον L_{subint} δίνεται από,

$$F_{subint} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,j} & \cdots & a_{1,n_{bins}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i,1} & \cdots & a_{i,j} & \cdots & a_{i,n_{bins}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n_{subint},1} & \cdots & a_{n_{subint},j} & \cdots & a_{n_{subint},n_{bins}} \end{bmatrix}, \quad (0.4)$$

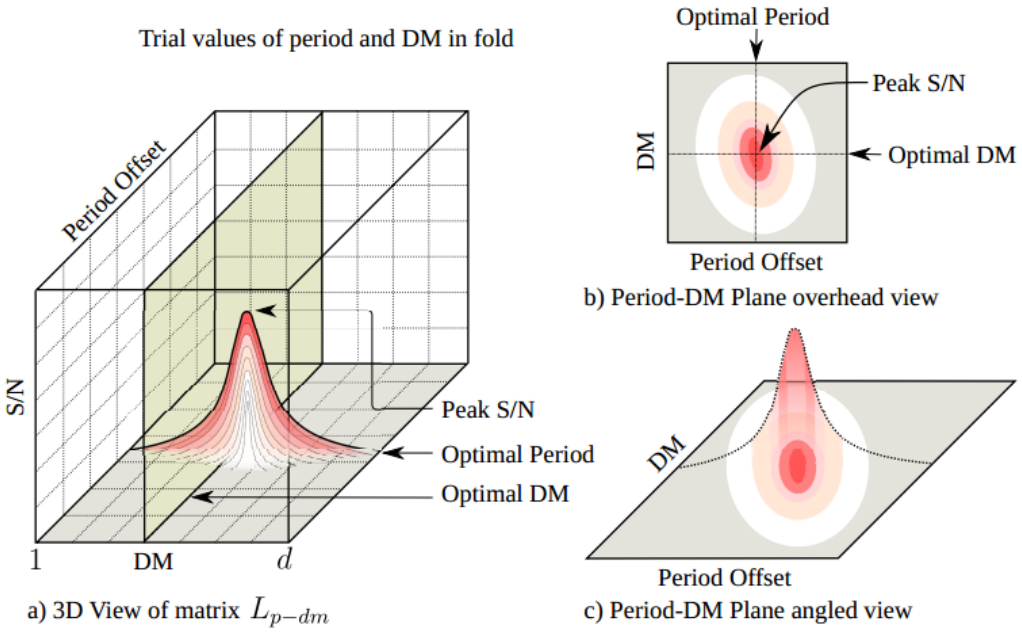
$$a_{ij} = \frac{1}{n_{bins}} \sum_{k=1}^{n_{bins}} b_{ik} \cdot l_{kj}, \quad (0.5)$$

έτσι ώστε το b_{ik} να είναι ένα στοιχείο από ένα μοναδιαίο διάνυσμα γραμμών 1 × nbins, και l_{kj} ένα στοιχείο του $n_{ints} \times n_{bins}$ πίνακα L^i_{subint}, που κόβει καθέτως τον L_{subint} στο i για i = 1,.., nsubint. Ένα παράδειγμα του τρόπου με τον οποίο λειτουργεί στην πράξη δίνεται στο σχήμα 2.5. Σημειώστε ότι οι n_{int} αναφέρονται στον αριθμό των γραμμών του πίνακα L^i_{subint} και καθορίζεται από τον αγωγό της έρευνας. Είναι επίσης δυνατό να ισοπεδωθεί ο πίνακας οριζοντίως παράγοντας ένα από τα πιο σημαντικά συνοπτικά περιγραφικά διαγράμματα για έναν υποψήφιο, το ολοκληρωμένο παλμικό προφίλ του [Ghosh, 2007] (Lorimer D.R. and Kramer M., 2005) [Lorimer, 2008]. Αυτή είναι μια σειρά συνεχόμενων τιμών που περιγράφουν την ένταση του παλμού για όλα τα n_{bins} φάσης, που έχουν κανονικοποιηθεί σε κάποιο εύρος τιμών [a,b].



Εικόνα 0.5 Απεικόνιση του τρόπου μείωσης διαστάσεων των δεδομένων χρησιμοποιώντας την Εξίσωση 2.5. Ο πίνακας 2 × 2 × 2 στο α) περιέχει τα αρχικά δεδομένα. Οι τομές στο α) είναι επίσης 2 × 2 πίνακες. Στο β) υπάρχει ένας πίνακας μειωμένων διαστάσεων και ένα παράδειγμα του πώς οι τιμές του υπολογίζονται.

Τα σήματα που προέρχονται από τη Γη δεν μεταδίδονται μέσω του ISM και έτσι δεν διασκορπίζονται. Εάν διορθώσεις δοκιμών DM εφαρμοστούν σε μη διασκορπισμένο σήμα, ο λόγος S / N θα μειωθεί καθώς το σήμα καταστρέφεται από περιττές διορθώσεις. Ένα τέτοιο σήμα θα επιτύχει το μέγιστο S / N του σε τιμή DM μηδενική. Ένας υποψήφιος που επιτυγχάνει μέγιστη τιμή S / N σε μηδενική τιμή DM, είναι πιθανό να οφείλεται σε τοπικές παρεμβολές. Ένα αυθεντικό σήμα pulsar από την άλλη πλευρά θα είναι διασκορπισμένο, επομένως ο λόγος S / N του θα πρέπει να μεγιστοποιηθεί σε ένα DM μεγαλύτερο από το μηδέν. Επομένως, το ποσοστό διασποράς ενός σήματος είναι ένας χρήσιμος δείκτης της πραγματικής του προέλευσης. Γι' αυτό το λόγο, τα βήματα βελτιστοποίησης που προσπαθούν να βρουν την καλύτερη περίοδο και DM για έναν αναδιπλωμένο υποψήφιο διατηρούν την περίοδο και τις πληροφορίες δοκιμής DM για ανάλυση. Αυτά τα δεδομένα αποθηκεύονται σε έναν περαιτέρω πίνακα δεδομένων $L_{p\text{-}dm}$. Αυτός ο πίνακας περιγράφει την επίδραση της δοκιμής DM και των τιμών μετατόπισης περιόδου, στον λόγο S / N του αναδιπλωμένου υποψηφίου. Αυτός ο πίνακας φαίνεται σαφέστερα στην Εικόνα 2.6.



Εικόνα 0.6 Απεικόνιση των δεδομένων περιόδου-DM αποθηκευμένων σε ένα υποψήφιο σήμα pulsar. Στο a) υπάρχει μια κορυφή που αντιστοιχεί στον συνδυασμό DM και χρονικής περιόδου που έδωσε την υψηλότερη τιμή S / N για τον υποψήφιο. Στο b) η κάτοψη αυτού του πίνακα μετά τη μείωση διαστάσεων. Αυτός ο πίνακας είναι χρήσιμος για την αξιολόγηση των υποψηφίων, καθώς τα αυθεντικά σήματα pulsar θα πρέπει να έχουν κυκλικές περιοχές με αυξανόμενες τιμές S / N όσο η περίοδος της δοκιμής και οι τιμές DM γίνονται βέλτιστες. Στο c) παρέχεται μια πλευρά της όψης του b).

Στην Εικόνα 2.6 a) βλέπουμε μια διαμόρφωση με κορυφή, όπου η κορυφή αντιστοιχεί στο συνδυασμό DM και περιόδου που έδωσε τον υψηλότερο λόγο S / N για τον υποψήφιο. Στην Εικόνα 2.6 b) φαίνεται η κάτοψη αυτού του πίνακα με μείωση διαστάσεων. Το βέλτιστο S / N υποδεικνύεται από την πιο σκούρα σκιασμένη περιοχή. Τα αυθεντικά σήματα pulsar διαθέτουν κυκλικές περιοχές αυξανόμενων τιμών S / N, καθώς η μετατόπιση δοκιμαστικής περιόδου και οι DM τιμές γίνονται βέλτιστες. Σε αυτό το παράδειγμα, ο S / N μειώνεται καθώς χρησιμοποιείται η λανθασμένη περίοδος μετατόπισης και λανθασμένο DM, όπως θα γινόταν στην περίπτωση ενός σήματος από pulsar. Αυτό το διάγραμμα είναι γνωστό και ως επίπεδο περιόδου-DM. Είναι επίσης δυνατό να ισοπεδωθεί ο πίνακας a) οριζόντια, σχηματίζοντας μία καμπύλη DM-SNR που περιγράφει τη σχέση μεταξύ S / N και DM. Μία καμπύλη με κορυφή για μηδενική τιμή DM είναι πιθανόν να προέρχεται από RFI σήματα.

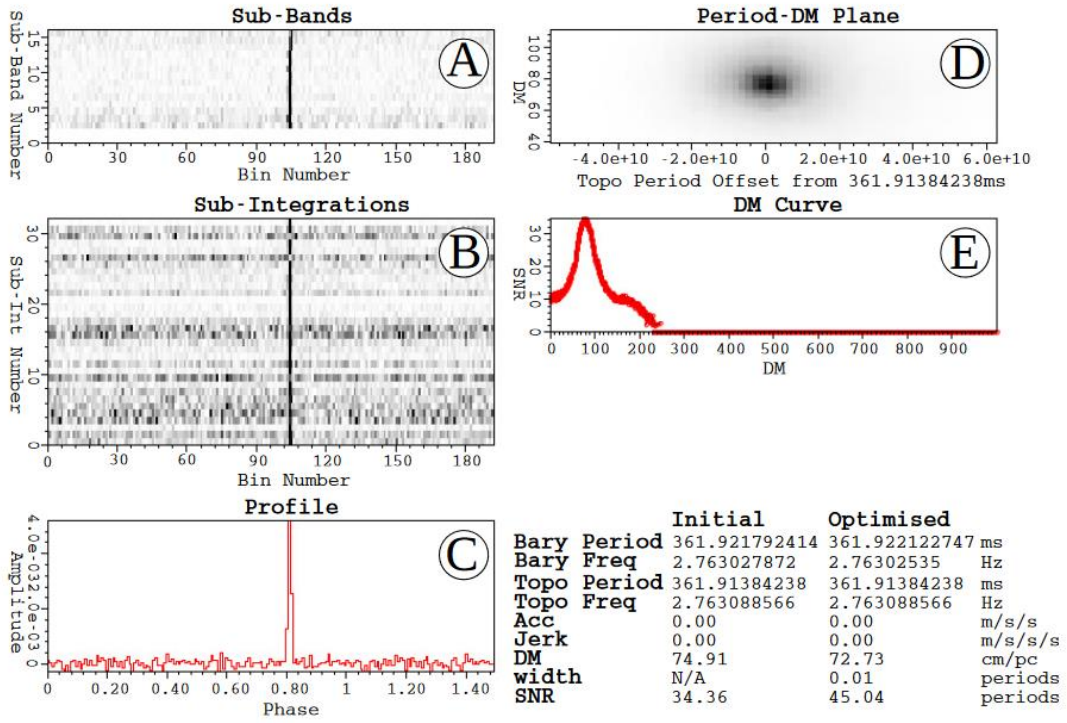
Συγκεντρωτικά τα στοιχεία που περιγράφονται μέχρι σήμερα αποτελούν το μεγαλύτερο μέρος απαιτήσεων αποθήκευσης δεδομένων ενός υποψήφιου. Το κατά προσέγγιση αποτύπωμα ενός υποψηφίου σε bits δίνεται από,

$$c_{bits} = \left(2 \cdot (n_{sub} \cdot n_{bins} \cdot n_{subint}) + n_{bins} + DM_{trials}\right) \cdot n_{bits}, \qquad (0.6)$$

όπου n_{bits} είναι ο αριθμός των bits που χρησιμοποιούνται για την αποθήκευση ενός μεμονωμένου δείγματος, και DM trials ο συνολικός αριθμός των δοκιμών DM που πραγματοποιήθηκαν. Ενώ είναι επιθυμητό να διατηρηθεί όσο το δυνατόν περισσότερη πληροφορία από έναν υποψήφιο, οι περιορισμοί στην αποθήκευση επιβάλλουν πρακτικούς περιορισμούς. Έτσι, κάθε υποψήφιος περιγράφει μόνο μια πολύ μειωμένη έκδοση του ανιχνευμένου σήματος. Ο μέσος υποψήφιος είναι συνήθως περίπου 60kB σε μέγεθος, όπου $n_{bits} = 32$, Δοκιμές DM = 1000, $n_{sub} = 32$ και τέλος $n_{bins} = 128$.

2.4 Υποψήφιοι pulsar

Ένας υποψήφιος pulsar είναι μια ανίχνευση σήματος που γίνεται από έναν αγωγό αναζήτησης pulsar, η οποία παρουσιάζει χαρακτηριστικά «τύπου pulsar» που την καθιστούν άξια περαιτέρω (Lorimer D.R. and Kramer M., 2005) [Eatough R.P., Selection of radio pulsar candidates using artificial neu-ral networks, 2001] [Morello V., 2014]. Κάθε υποψήφιος συνοψίζει μια τέτοια ανίχνευση χρησιμοποιώντας γραφικές παραστάσεις και στατιστικές που προέρχονται από πίνακες που περιεγράφηκαν προηγουμένως, δίνοντας κάποια ένδειξη της προέλευσής του. Αυτά τα γραφήματα και οι στατιστικές πρέπει να επιθεωρούνται είτε με αυτοματοποιημένη μέθοδο είτε από κάποιον εμπειρογνώμονα, προκειμένου να προσδιοριστεί η πιθανή προέλευση του υποψηφίου. Η συντριπτική πλειοψηφία των υποψηφίων θα προκληθεί από τις διακυμάνσεις του θορύβου του Γαλαξιακού υποβάθρου, του θορύβου των οργάνων και επίγειων RFI. Επομένως, θα πρέπει να εξεταστούν περαιτέρω μόνο αυτοί οι υποψήφιοι που είναι πιθανά σήματα pulsar, καθώς επίσης και οι χρόνοι του τηλεσκοπίου για μετέπειτα παρατηρήσεις επιβεβαίωσης. Η διαδικασία απόφασης για το ποιοι υποψήφιοι αξίζουν περαιτέρω ανάλυση ονομάζεται 'επιλογή' υποψηφίων. Η επιλογή υποψηφίων είναι ένα σημαντικό βήμα στην αναζήτηση pulsar. Οι σωστές αποφάσεις επιλογής επιτρέπουν τον πολύτιμο χρόνο του τηλεσκοπίου να δίνεται με προτεραιότητα σε αυτές τις ανιχνεύσεις που είναι πιθανό να οδηγήσουν σε μια νέα ανακάλυψη και να αποτραπεί η απώλεια μη ανακάλυψης pulsar. Λανθασμένες αποφάσεις αποτελούν χάσιμο χρόνου και προσπάθειας, και μπορούν ακόμη και να προκαλέσουν πραγματικά σήματα από pulsars να αγνοηθούν.



Εικόνα 0.7 _Ένας επεξηγηματικός, παραδειγματικός υποψήφιος pulsar, που συνοψίζει την ανίχνευση του PSR J1706-6118. Ο υποψήφιος πάρθηκε κατά τη διάρκεια της επεξεργασίας δεδομένων του High Time Resolution Survey από το Thornton (2013).

Ένας τυπικός υποψήφιος περιγράφεται με έναν μικρό αριθμό χαρακτηριστικών μεταβλητών και διαγραμμάτων. Οι συνοπτικές στατιστικές περιλαμβάνουν γενικά την περίοδο του σήματος, τον λόγο S / N, το μέτρο διασποράς DM και το εύρος του παλμού που ανιχνεύθηκε. Οι περισσότεροι υποψήφιοι περιλαμβάνουν επίσης ένα αντίγραφο του ολοκληρωμένου παλμικού προφίλ [Lorimer, 2008] (Lorimer D.R. and Kramer M., 2005) [Ghosh, 2007], το οποίο απεικονίζει οπτικά μια μέση εκδοχή του σήματος σε όλες τις παρατηρούμενες συχνότητες και χρόνο (βλ. (C) στο Εικόνα 2.7).

Πρόκειται για ένα πίνακα συνεχών τιμών που περιγράφει την ένταση του παλμού για nbins bins φάσης κανονικοποιημένο σε ένα διάστημα [a, b]. Ωστόσο, διαφορετικοί αγωγοί αναζήτησης pulsar παράγουν διαφορετικούς τύπους υποψηφίων, οι οποίοι συχνά περιέχουν πρόσθετες πληροφορίες όπως φαίνεται στο σχήμα 2.7. Αυτός ο υποψήφιος παρέχει πληροφορίες που περιγράφουν τον τρόπο με τον οποίο το σήμα συμπεριφέρεται σε όλο το φάσμα χρόνου και συχνότητας [Eatough R.P., Selection of radio pulsar candidates using artificial neural networks, 2010], χρησιμοποιώντας πίνακα μειωμένων διαστάσεων Fsub στην Εικόνα 2.7 (A) και τον πίνακα Fsubint στην Εικόνα 2.7 (B). Διαγράμματα που περιγράφουν τη σχέση μεταξύ των δοκιμαστικών τιμών DM και του S / N, που προκύπτουν όταν χρησιμοποιείται αυτό το DM για να διορθωθεί η διασπορά, χρησιμοποιούνται επίσης (επίπεδο περίοδος-DM,και καμπύλη DM-SNR).

(Στις μέρες μας η έρευνα για pulsars γίνεται με συνδυασμό χρήσης μεγάλων ραδιοτηλεσκοπίων, αλγορίθμων για την ανάλυση σήματος και ανθρώπινης εφευρετικότητας. Αν και τα τηλεσκόπια είναι αρκετά ικανά στην ανίχνευση των δικών τους σημάτων και οι άνθρωποι ικανοί να αναγνωρίσουν τις δικές τους εκπομπές όταν αυτές προκύπτουν, αυτά δεν επαρκούν για την άμεση και ακριβή εξακρίβωση pulsar. Αυτό συμβαίνει διότι ένα ραδιοτηλεσκόπιο λαμβάνει έναν μεγάλο αριθμό σημάτων που μοιάζουν με το σήμα του pulsar κατά τη διάρκεια μιας τυπικής παρατήρησης. Αυτά τα σήματα εμποδίζουν την παρατήρηση του πραγματικού pulsar και κάποιος θα μπορούσε να το παρομοιάσει την εύρεση pulsar με εύρεση βελόνας στα άχυρα (Lee, 2013). Τα άχυρα στην προκειμένη περίπτωση είναι τα υποψήφια σήματα pulsar.

3. Κεφάλαιο 3: UCI Repository

Το UCI Machine Learning Repository είναι μια συλλογή από βάσεις δεδομένων, θεωρίες πεδίων και γεννήτριες δεδομένων που χρησιμοποιούνται από την κοινότητα μηχανικής μάθησης για την εμπειρική ανάλυση των αλγορίθμων μηχανικής μάθησης. [Dua, 2018]. Το αρχείο δημιουργήθηκε ως αρχείο ftp το 1987 από τον David Aha και συναδέλφους μεταπτυχιακούς φοιτητές στο UC Irvine. Από τότε, έχει χρησιμοποιηθεί ευρέως από φοιτητές, εκπαιδευτικούς και ερευνητές σε όλο τον κόσμο ως πρωταρχική πηγή των συνόλων δεδομένων μηχανικής μάθησης. Ως ένδειξη του αντίκτυπου του αρχείου, έχει αναφερθεί πάνω από 1000 φορές, καθιστώντας το ένα από τα 100 πιο αναφερόμενα "έγγραφα" σε όλο τον τομέα της επιστήμης υπολογιστών. Η τρέχουσα έκδοση του ιστότοπου σχεδιάστηκε το 2007 από τους Arthur Asuncion και David Newman και το έργο αυτό είναι σε συνεργασία με το Rexa.info στο Πανεπιστήμιο της Μασαχουσέτης Amherst και χρηματοδοτείται από το National Science Foundation.

3.1 Dataset

Το HTRU2 είναι ένα σύνολο δεδομένων που περιγράφει ένα δείγμα υποψηφίων pulsar που συλλέχθηκαν κατά τη διάρκεια του Hight Time Resolution Universe Survey (South) [Keith, 2010]. Οι Pulsars είναι ένας σπάνιος τύπος αστέρα νετρονίων που παράγει εκπομπές ραδιοσημάτων που εντοπίζονται στη Γη. Έχουν σημαντικό επιστημονικό ενδιαφέρον για τη διερεύνηση του χωροχρόνου, του διαστρικού μέσου (ISM) και των καταστάσεων της ύλης. (Lorimer D.R. and Kramer M., 2005). Καθώς οι pulsars περιστρέφονται, η δέσμη εκπομπών τους σαρώνει τον ουρανό, και όταν αυτή διασχίσει το οπτικό μας πεδίο, παράγεται ένα ανιχνεύσιμο μοτίβο ραδιοεκπομπών ευρείας ζώνης. Καθώς οι pulsars περιστρέφονται γρήγορα αυτό το μοτίβο επαναλαμβάνεται περιοδικά. Έτσι, η αναζήτηση pulsar περιλαμβάνει την αναζήτηση περιοδικών ραδιοφωνικών σημάτων με μεγάλα ραδιοτηλεσκόπια. Κάθε pulsar παράγει ένα ελαφρώς διαφορετικό μοτίβο εκπομπής, το οποίο ποικίλλει ελαφρώς με κάθε περιστροφή (Lorimer D.R. and Kramer M., 2005). Έτσι, μια πιθανή ανίχνευση σήματος

που είναι γνωστή ως «υποψήφιος», υπολογίζεται κατά μέσο όρο σε πολλές περιστροφές του pulsar, όπως καθορίζεται από τη διάρκεια μιας παρατήρησης. Ελλείψει πρόσθετων πληροφοριών, κάθε υποψήφιος θα μπορούσε ενδεχομένως να περιγράφει έναν πραγματικό pulsar. Ωστόσο, στην πράξη, σχεδόν όλες οι ανιχνεύσεις προκαλούνται από παρεμβολές ραδιοσυχνοτήτων (RFI) και θόρυβο, καθιστώντας δύσκολη την ανίχνευση πραγματικών σημάτων.

Εργαλεία μηχανικής μάθησης χρησιμοποιούνται τώρα για την αυτόματη κατηγοριοποίηση υποψηφίων pulsar για να διευκολύνουν την ταχεία ανάλυση. Συγκεκριμένα, συστήματα ταξινόμησης υιοθετούνται ευρέως, [Lyon R. J., Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach, 2016; Eatough R.P., Selection of radio pulsar candidates using artificial neural networks, 2010; Bates, 2012; Thornton, 2013; Lee, 2013; Morello, 2014] τα οποία αντιμετωπίζουν τα υποψήφια σύνολα δεδομένων ως δυαδικά προβλήματα κατάταξης. Εδώ τα πραγματικά παραδείγματα pulsar είναι μια θετική τάξη μειοψηφίας, και ψευδή παραδείγματα η πλειοψηφική αρνητική τάξη. Επί του παρόντος, οι ετικέτες πολλαπλών κατηγοριών δεν είναι διαθέσιμες, δεδομένων των δαπανών που σχετίζονται με την παρατήρηση των δεδομένων.

Το σύνολο δεδομένων περιέχει 16.259 παραπλανητικά παραδείγματα που προκαλούνται από RFI / θόρυβο και 1.639 πραγματικά παραδείγματα pulsar. Όλα αυτά τα παραδείγματα έχουν ελεγχθεί από τους ανθρώπινους παρατηρητές [R. J. Lyon, 2016].

Τα δεδομένα παρουσιάζονται σε δύο μορφές: CSV και ARFF (χρησιμοποιούνται από το εργαλείο εξόρυξης δεδομένων WEKA). Οι υποψήφιοι αποθηκεύονται και στα δύο αρχεία σε ξεχωριστές σειρές. Κάθε σειρά παραθέτει πρώτα τις μεταβλητές και η ετικέτα κλάσης είναι η τελική καταχώρηση. Οι ετικέτες κλάσης που χρησιμοποιούνται είναι 0 (αρνητικές) και 1 (θετικές).

Τα δεδομένα δεν περιέχουν πληροφορίες θέσης ή άλλες αστρονομικές λεπτομέρειες. Είναι απλά δεδομένα χαρακτηριστικών που εξάγονται από τα υποψήφια αρχεία χρησιμοποιώντας το εργαλείο PulsarFeatureLab [Lyon R. J., PulsarFeatureLab, 2015].

3.2 Επισκόπηση HTRU2 Dataset

3.2.1 Χρήσιμες έννοιες/ορισμοί:

- **Παλμίτης αστέρας (Pulsar)**: Οι πάλσαρ είναι αστέρες νετρονίων με ισχυρό μαγνητικό πεδίο που περιστρέφονται ταχύτατα γύρω από άξονα και καθώς τους παρατηρούμε από την Γη καταγράφουμε αλληλουχία σχεδόν περιοδικών παλμών. Είναι μία κατηγορία παλλόμενων ουράνιων ραδιοπηγών, ουράνια δηλαδή σώματα που εκπέμπουν ανιχνεύσιμη ηλεκτρομαγνητική ακτινοβολία με τη μορφή ραδιοφωνικών κυμάτων. Οι πάλσαρ ξεχωρίζουν από

όλες τις άλλες ουράνιες πηγές επειδή παρατηρούμε από αυτούς ταχύτατους περιοδικούς παλμούς σε διάφορες περιοχές ηλεκτρομαγνητικής ακτινοβολίας, με περιόδους από χιλιοστά του sec (msec), μέχρι μερικά δευτερόλεπτα, σε αντιδιαστολή προς όλα τα άλλα ουράνια σώματα που εμφανίζουν περιόδους κάθε είδους μεταβολών (περιστροφής κλπ.) της τάξεως των ωρών και άνω. Από τη λέξη **pulse** (= παλμός) προέρχεται και η ονομασία τους: **pulsar = PULSating stAR** (παλλόμενος αστέρας), ενώ καταγράφονται με το σύμβολο **PSR** ακολουθούμενο με την ορθή αναφορά τους εκφρασμένη σε χρόνο δευτερολέπτων. Οι πάλσαρ είναι αστέρες νετρονίων που έχουν δημιουργηθεί μετά από τη κατάρρευση ή εκφυλισμό προϋπάρχοντος κανονικού αστέρα που ναι μεν η μάζα τους είναι παραπλήσια του Ηλίου, πλην όμως η διάμετρός τους είναι πολύ μικρή, λίγες δεκάδες χλμ.

- **Μέτρο Διασποράς/DM (Dispersion Measure):** Το "άπλωμα" της διασποράς του λαμβανόμενου σήματος είναι ανάλογο με το λεγόμενο μέτρο διασποράς (Dispersion Measure), το οποίο αντιπροσωπεύει την πυκνότητα της στήλης των ελευθέρων ηλεκτρονίων που παρατηρείται μεταξύ του παρατηρητή και του πάλσαρ σε όλο το μήκος της γραμμής παρατήρησης.

- **High Time Resolution Universe (HTRU)**: Είναι μία έρευνα σε όλο τον ουρανό για pulsar και παροδικά ραδιοσήματα σε συχνότητα 1400 MHz.

- **"Υποψήφιος" Πάλσαρ (Pulsar Candidate)**: Για την εύρεση πιθανών ραδιοσημάτων πάλσαρ χρησιμοποιούνται τεχνικές που απομονώνουν τα περιοδικά ευρείας ζώνης σήματα που παρουσιάζουν διασπορά μετά από τη διάδοσης της στον διαστρικό χώρο (η ύλη και η ακτινοβολία που υπάρχει μεταξύ των συστημάτων αστέρων στον γαλαξία). Σήματα που πληρούν αυτές τις προϋποθέσεις καταγράφονται ως μία συλλογή από διαγνωστικά γραφήματα και περιγραφικές στατιστικές, αναφερόμενα και ως "υποψήφια" πάλσαρ, δηλαδή υποψήφια ανίχνευση πάλσαρ.

- **Επιλογή υποψηφίων (Candidate Selection)**: Η διαδικασία επιλογής "υποψήφιων" πάλσαρ, τα οποία αξίζουν περαιτέρω μελέτη.

- **Ενοποιημένο Προφίλ Πάλσαρ (Integrated Pulsar Profile):** Είναι ένας πίνακας των συνεχών μεταβλητών, οι οποίες περιγράφουν το μέσο όρο σε χρόνο και συχνότητα του σήματος στο γεωγραφικό μήκος.

3.2.2 Περιγραφή μεταβλητών/Στατιστικοί ορισμοί:

Οι πρώτες τέσσερις μεταβλητές προέρχονται από απλή στατιστική του ενοποιημένου προφίλ και οι άλλες τέσσερις από την καμπύλη DM-SNR της εικόνας 1. Συγκεκριμένα έχουμε τη μέση τιμή, την τυπική απόκλιση, την ασυμμετρία και την κύρτωση για το

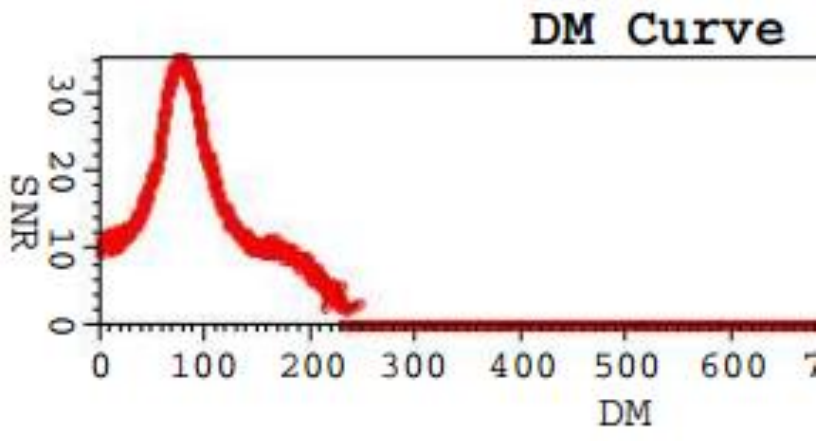
ενοποιημένο προφίλ και τις αντίστοιχες τιμές που λαμβάνουμε από την καμπύλη DM/SNR (Εικόνα 3.1-2).

- **Μέση τιμή (Mean**): Το αλγεβρικό άθροισμα των τιμών μίας μεταβλητής διαιρούμενο δια του πλήθους.

- **Τυπική Απόκλιση (Standard Deviation):** Μετράει πόσο αποκλίνουν κατά μέσο όρο οι τιμές από τη μέση τιμή της κατανομής

- **Ασυμμετρία (Skewness):** Η ασυμμετρία μίας κατανομής εκτιμάται από το συντελεστή ασυμμετρίας, ο οποίος είναι καθαρός αριθμός και παίρνει τιμές σε όλο το φάσμα των πραγματικών αριθμών. Τιμές του συντελεστή ασυμμετρίας στην περιοχή του μηδενός ορίζουν συμμετρικές κατανομές. Όσο μεγαλύτερος είναι κατά απόλυτη τιμή ο συντελεστής ασυμμετρίας τόσο μεγαλύτερη είναι και η ασυμμετρία της κατανομής (προς την αντίστοιχη πλευρά που ορίζει το πρόσημο του συντελεστή) .

- **Κύρτωση (Kurtosis)**: Εκτιμάται από τον συντελεστή κύρτωσης και εκφράζει το βαθμό συγκέντρωσης των τιμών της κατανομής περί το μέσον της. Σε σχέση με την κανονική κατανομή, αν το ποσοστό των παρατηρήσεων της κατανομής που βρίσκονται στο κέντρο της, είναι μεγαλύτερο του αντίστοιχου της κανονικής κατανομής, η κύρτωση της κατανομής είναι θετική και η κατανομή χαρακτηρίζεται ως λεπτόκυρτη. Σε διαφορετική περίπτωση η κύρτωση της κατανομής είναι αρνητική και η κατανομή χαρακτηρίζεται ως πλατύκυρτη.



Εικόνα 3.1 DM-SNR καμπύλη

Feature	Description	Definition
$Prof_{\mu}$	Mean of the integrated profile P.	$\frac{1}{n}\sum_{i=1}^{n} p_i$
$Prof_{\sigma}$	Standard deviation of the integrated profile P.	$\sqrt{\frac{\sum_{i=1}^{n}(p_i - \bar{P})^2}{n-1}}$
$Prof_{k}$	Excess kurtosis of the integrated profile P.	$\frac{\frac{1}{n}(\sum_{i=1}^{n}(p_i - \bar{P})^4)}{(\frac{1}{n}(\sum_{i=1}^{n}(p_i - \bar{P})^2))^2} - 3$
$Prof_{s}$	Skewness of the integrated profile P.	$\frac{\frac{1}{n}\sum_{i=1}^{n}(p_i - \bar{P})^3}{(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - \bar{P})^2})^3}$
DM_{μ}	Mean of the DM-SNR curve D.	$\frac{1}{n}\sum_{i=1}^{n} d_i$
DM_{σ}	Standard deviation of the DM-SNR curve D.	$\sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{D})^2}{n-1}}$
DM_{k}	Excess kurtosis of the DM-SNR curve D.	$\frac{\frac{1}{n}(\sum_{i=1}^{n}(d_i - \bar{D})^4)}{(\frac{1}{n}(\sum_{i=1}^{n}(d_i - \bar{D})^2))^2} - 3$
DM_{s}	Skewness of the DM-SNR curve D.	$\frac{\frac{1}{n}\sum_{i=1}^{n}(d_i - \bar{D})^3}{(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \bar{D})^2})^3}$

Εικόνα 3.2 Τα οχτώ χαρακτηριστικά που προέρχονται από το ενοποιημένο προφίλ πάλσαρ P = {p1, ..., pn} και την καμπύλη DM-SNR D = {d1, ..., dn}.

3.3 TMVA εργαλειοθήκη

Η εργαλειοθήκη για την ανάλυση πολλών μεταβλητών (TMVA) [A. Hoecker, 2007] παρέχει ένα περιβάλλον [Brun, 1997] ενσωματωμένο στο ROOT για την επεξεργασία, την παράλληλη αξιολόγηση και την εφαρμογή της ταξινόμησης πολλαπλών μεταβλητών και - από το TMVA version 4 - τεχνικές παλινδρόμησης πολλαπλών μεταβλητών. Όλες μέθοδοι για την ανάλυση πολλών μεταβλητών στο TMVA ανήκουν στην οικογένεια αλγορίθμων μάθησης υπό επίβλεψη. Χρησιμοποιούν γεγονότα εκπαίδευσης, για τα οποία είναι γνωστό το επιθυμητό αποτέλεσμα, για να προσδιορίσουν συνάρτηση που είτε περιγράφει ένα όριο απόφασης (ταξινόμηση) είτε μια προσέγγιση της υποκείμενης συμπεριφοράς της συνάρτησης που καθορίζει την τιμή στόχου (παλινδρόμηση). Η συνάρτηση μπορεί να περιέχει διάφορους βαθμούς προσεγγίσεων και μπορεί να είναι μια ενιαία καθολική συνάρτηση ή ένα σύνολο τοπικών μοντέλων. Το TMVA έχει σχεδιαστεί ειδικά για εφαρμογές φυσικής υψηλής ενέργειας (HEP), αλλά δεν πρέπει να περιορίζεται σε αυτές.

4. Κεφάλαιο 4: Μηχανική Μάθηση

4.1 Εισαγωγή

Η **Μηχανική Μάθηση (Machine Learning)** είναι ένας κλάδος της Τεχνητής Νοημοσύνης που μπορεί να οριστεί ως οι υπολογιστικές μέθοδοι που χρησιμοποιούν εμπειρία για να βελτιώσουν την απόδοση ενός συστήματος ή να πραγματοποιήσουν ακριβείς προβλέψεις [Mehryar Mohri, 2012]. Η έννοια της **εμπειρίας (experience)** αναφέρεται στην πληροφορία του παρελθόντος η οποία είναι διαθέσιμη στο σύστημα μάθησης, η οποία συνήθως παίρνει την μορφή ηλεκτρονικών δεδομένων που έχουν συλλεχθεί και είναι διαθέσιμα για ανάλυση. Τα δεδομένα αυτά μπορεί να έχουν την μορφή συνόλων εκπαίδευσης τα οποία έχουν χαρακτηριστεί και ταξινομηθεί κατάλληλα από ανθρώπους ή να προέρχονται από την αλληλεπίδραση με το περιβάλλον. Σε κάθε περίπτωση, η ποιότητα και το μέγεθος του συνόλου δεδομένων είναι κρίσιμος παράγοντας για την επιτυχία των προβλέψεων του συστήματος μάθησης.

Η Μηχανική Μάθηση περιλαμβάνει τη σχεδίαση αποδοτικών αλγορίθμων οι οποίοι παράγουν ακριβείς προβλέψεις. Όπως και σε άλλα πεδία της επιστήμης των υπολογιστών, δυο κρίσιμα μέτρα της ποιότητας των αλγορίθμων αυτών είναι η χρονική και χωρική πολυπλοκότητά τους. Στη Μηχανική Μάθηση απαιτείται επιπλέον η έννοια της **πολυπλοκότητας δείγματος (sample complexity),** για την αξιολόγηση του μεγέθους του δείγματος που χρειάζεται ο αλγόριθμος για να μάθει μια οικογένεια εννοιών. Οι θεωρητικές εγγυήσεις για την αποτελεσματικότητα ενός αλγορίθμου μάθησης εξαρτώνται από την πολυπλοκότητα των εννοιών – κλάσεων και από το μέγεθος του συνόλου των δειγμάτων εκπαίδευσης.

Από τη στιγμή που η επιτυχία ενός αλγορίθμου μάθησης εξαρτάται από τα δεδομένα που χρησιμοποιούνται, η Μηχανική Μάθηση σχετίζεται άμεσα με την ανάλυση δεδομένων και την στατιστική. Οι τεχνικές εκμάθησης βασίζονται στα δεδομένα και συνδυάζουν βασικές έννοιες της επιστήμης υπολογιστών με ιδέες από στατιστική, πιθανότητες και από τεχνικές βελτιστοποίησης [Mehryar Mohri, 2012].

4.1.1 Ορισμοί και Ορολογία

Θεωρώντας το πρόβλημα της αυτόματης ταξινόμησης γίνεται μια πρώτη αναφορά σε βασικές έννοιες και στη χρήση των αλγορίθμων μάθησης στην πράξη [Mehryar Mohri, 2012].

- **Δείγματα (Samples).** Αντικείμενα ή στιγμιότυπα δεδομένων που χρησιμοποιούνται για εκμάθηση ή αξιολόγηση. Στο πρόβλημα ταξινόμησης σήματος pulsar ή θορύβου τα δείγματα αυτά αντιστοιχούν σε ένα σύνολο «υποψηφίων» σημάτων pulsar, που θα χρησιμοποιηθούν για την εκμάθηση και για τον έλεγχο αποτελεσματικότητας της αυτόματης ταξινόμησης.

- **Χαρακτηριστικά (Features).** Το σύνολο των γνωρισμάτων (attributes) που σχετίζονται με ένα δείγμα, σχηματίζοντας συνήθως ένα διάνυσμα. Στο συγκεκριμένο παράδειγμα ταξινόμησης τα χαρακτηριστικά είναι η οχτώ στο σύνολο: η μέση τιμή, η τυπική απόκλιση, η κύρτωση, η ασυμμετρία για το ολοκληρωμένο προφίλ pulsar και για την καμπύλη DM-SNR.

- **Κατηγορίες ή Κλάσεις (Categories, Classes).** Οι κατηγορίες ή οι κλάσεις που έχουν ανατεθεί στα δείγματα. Στα προβλήματα ταξινόμησης, κάθε δείγμα κατηγοριοποιείται σε συγκεκριμένες κλάσεις, όπως είναι οι κλάσεις «pulsar» και «θόρυβος».

- **Σύνολο Εκπαίδευσης (Training Set).** Δείγματα που χρησιμοποιούνται για την εκπαίδευση ενός αλγορίθμου εκμάθησης. Για τη συγκεκριμένη ταξινόμηση, το σύνολο εκπαίδευσης αποτελείται από ένα σύνολο υποψηφίων μαζί με τις κατηγορίες στις οποίες έχουν ταξινομηθεί (pulsarς ή μη pulsar). Το σύνολο εκπαίδευσης διαφέρει, ανάλογα με το σενάριο μάθησης.

- **Σύνολο Επικύρωσης (Validation Set).** Δείγματα με γνωστές τις κατηγορίες στις οποίες ανήκουν, για τη ρύθμιση των παραμέτρων του αλγορίθμου εκμάθησης. Οι αλγόριθμοι εκμάθησης διαθέτουν συνήθως μια ή περισσότερες ελεύθερες παραμέτρους και το σύνολο επικύρωσης χρησιμοποιείται για την επιλογή κατάλληλων τιμών για αυτές.

- **Σύνολο Ελέγχου (Test Set).** Δείγματα για την αξιολόγηση της απόδοσης του αλγόριθμου εκμάθησης. Το σύνολο ελέγχου είναι διαφορετικό από το σύνολο εκπαίδευσης και το σύνολο επικύρωσης, και δεν είναι διαθέσιμο κατά την φάση της εκπαίδευσης. Για το παρόν πρόβλημα ταξινόμησης, το σύνολο ελέγχου αποτελείται από ένα σύνολο υποψηφίων, για τους οποίους ο αλγόριθμος εκμάθησης θα πρέπει να προβλέψει τις κλάσεις στις οποίες ανήκουν (pulsar ή μη pulsar), με βάση τα διαθέσιμα χαρακτηριστικά τους. Στη συνέχεια οι προβλέψεις αυτές συγκρίνονται με τις πραγματικές κλάσεις του συνόλου ελέγχου για τη μέτρηση της απόδοσης του αλγορίθμου.

- **Συνάρτηση Σφάλματος (Error Function).** Μια συνάρτηση που μετράει τη διαφορά, ή το σφάλμα, μεταξύ της προβλεπόμενης κλάσης και της πραγματικής κλάσης για κάποιο δείγμα. Θεωρώντας το σύνολο όλων των κλάσεων ως το Y

και το σύνολο όλων των πιθανών προβλέψεων ως το Y′, η συνάρτηση σφάλματος είναι η αντιστοίχιση : L : Y x Y'→ℝ_+ . Στις περισσότερες περιπτώσεις ισχύει ότι Y = Y ′ και η συνάρτηση σφάλματος είναι φραγμένη.

- **Σύνολο Υποθέσεων (Hypothesis set).** Ένα σύνολο συναρτήσεων, οι οποίες αντιστοιχίζουν διανύσματα χαρακτηριστικών (feature vectors) στο σύνολο των κλάσεων Y. Στο πρόβλημα οι συναρτήσεις αυτές αντιστοιχίζουν τα χαρακτηριστικά του συνόλου του σήματος στο σύνολο Y = {pulsar, όχι pulsar}. Σε γενικές γραμμές οι υποθέσεις μπορεί να είναι συναρτήσεις που αντιστοιχίζουν τα χαρακτηριστικά σε ένα διαφορετικό σύνολο Y'. Μπορεί να είναι γραμμικές συναρτήσεις που αντιστοιχίζουν τα διανύσματα χαρακτηριστικών του συνόλου των υποψηφίων σε πραγματικές τιμές, οι οποίες ερμηνεύονται ως βαθμολογίες (scores) (Y' = ℝ), με τις υψηλότερες βαθμολογίες να είναι περισσότερο αντιπροσωπευτικές των υποψηφίων είναι όντως pulsar.

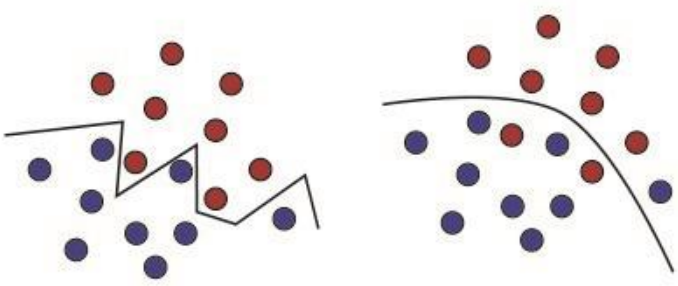
Στη συνέχεια περιγράφονται εν συντομία οι διαφορετικές φάσεις εκμάθησης για το πρόβλημα της ταξινόμησης υποψηφίων ως pulsar ή θόρυβο. Η διαδικασία ξεκινάει με το διαθέσιμο σύνολο δειγμάτων με γνωστές κλάσεις, το οποίο διαχωρίζεται τυχαία σε ένα σύνολο εκπαίδευσης, ένα σύνολο επικύρωσης και ένα σύνολο ελέγχου. Το μέγεθος των συνόλων αυτών εξαρτάται από ένα σύνολο παραγόντων. Για παράδειγμα, το πλήθος των δεδομένων επικύρωσης εξαρτάται από τον αριθμό των ελεύθερων παραμέτρων του αλγορίθμου. Επίσης, όταν το πλήθος των δεδομένων με γνωστές κλάσεις είναι μικρό, τα δεδομένα εκπαίδευσης θα πρέπει να είναι περισσότερα από τα δεδομένα ελέγχου, για να επιτευχθεί καλύτερη απόδοση.

Έπειτα, γίνεται επιλογή των κατάλληλων χαρακτηριστικών που θα χρησιμεύσουν στη διαδικασία εκπαίδευσης. Το βήμα αυτό είναι ιδιαίτερα κρίσιμο, καθώς χρήσιμα χαρακτηριστικά οδηγούν σωστά τον αλγόριθμο εκμάθησης ενώ μια λανθασμένη επιλογή χαρακτηριστικών με άσχετη πληροφορία θα έχει ως αποτέλεσμα πολύ χαμηλή απόδοση. Η επιλογή των χαρακτηριστικών αντικατοπτρίζει την εκ των προτέρων γνώση (prior knowledge) που υπάρχει σχετικά με το πρόβλημα μάθησης και επηρεάζει ιδιαίτερα την απόδοση του συστήματος μάθησης.

Τα χαρακτηριστικά που έχουν επιλεγεί χρησιμοποιούνται για την εκπαίδευση του συστήματος μάθησης. Για κάθε τιμή των παραμέτρων του συστήματος, ο αλγόριθμος επιλέγει μια διαφορετική υπόθεση από το σύνολο των υποθέσεων. Από τις υποθέσεις αυτές επιλέγεται εκείνη που δίνει μεγαλύτερη απόδοση στο σύνολο επικύρωσης. Η απόδοση του αλγορίθμου υπολογίζεται χρησιμοποιώντας την συνάρτηση σφάλματος που σχετίζεται με το πρόβλημα. Για παράδειγμα, για το πρόβλημα της ταξινόμησης σήματος ως θόρυβο ή σήμα, χρησιμοποιείται η συνάρτηση σφάλματος μηδέν-ένα για τη σύγκριση των προβλεπόμενων κλάσεων με τις πραγματικές κλάσεις.

Η απόδοση ενός αλγορίθμου υπολογίζεται με βάση τα λάθη του στο σύνολο ελέγχου και όχι στο σύνολο εκπαίδευσης. Ένας αλγόριθμος μπορεί να είναι συνεπής ως προς το σύνολο εκπαίδευσης, αλλά να μην αποδίδει καλά στο σύνολο ελέγχου. Αυτό το

φαινόμενο παρουσιάζεται σε συστήματα μάθησης με πολύπλοκες επιφάνειες απόφασης, όπως αυτή στην Εικόνα 4.1, η οποία τείνει να απομνημονεύει ένα σχετικά μικρό δείγμα δεδομένων αντί να έχει καλή απόδοση πρόβλεψης σε νέα δεδομένα. Το παράδειγμα αυτό δείχνει τον **διαχωρισμό μεταξύ της απομνημόνευσης και της ικανότητας πρόβλεψης**, η οποία είναι η βασική ιδιότητα που θα πρέπει να έχει ένα σύστημα μάθησης [Mehryar Mohri, 2012].



Εικόνα 0.1 Η τεθλασμένη γραμμή στα αριστερά είναι συνεπής ως προς το σύνολο εκπαίδευσης, αλλά είναι μια πολύπλοκη επιφάνεια απόφασης η οποία δεν παράγει καλές προβλέψεις για νέα δεδομένα. Αντιθέτως, η επιφάνεια απόφασης στο σχήμα δεξιά είναι πιο απλή και παρέχει καλύτερες προβλέψεις, παρά την λανθασμένη ταξινόμηση λίγων δειγμάτων του συνόλου εκπαίδευσης.

4.1.2 Σενάρια Μάθησης

Στη συνέχεια παρουσιάζονται τα περισσότερο διαδεδομένα σενάρια μάθησης. Τα σενάρια αυτά διαφέρουν ως προς τον τύπο των διαθέσιμων δεδομένων, τη μέθοδο με την οποία λαμβάνονται τα δεδομένα εκπαίδευσης και τα δεδομένα ελέγχου για την αξιολόγηση του συστήματος [Mehryar Mohri, 2012].

- **Μάθηση υπό επίβλεψη (supervised learning).** Το σύστημα μάθησης λαμβάνει ένα σύνολο από δείγματα κατηγοριοποιημένα σε κλάσεις ως δεδομένα εκπαίδευσης και πραγματοποιεί προβλέψεις για νέα δεδομένα. Αυτό είναι το πιο κοινό σενάριο που σχετίζεται με προβλήματα όπως ταξινόμηση, πρόβλεψη τιμής συνάρτησης και ταξινόμηση με βάση κάποιο κριτήριο.

- **Μάθηση χωρίς επίβλεψη (unsupervised learning).** Το σύστημα μάθησης λαμβάνει δεδομένα εκπαίδευσης για τα οποία δεν είναι γνωστές οι κλάσεις τους και παράγει προβλέψεις για νέα δεδομένα. Επειδή δεν υπάρχουν δεδομένα με γνωστές κλάσεις είναι δύσκολο να γίνει ποσοτική αξιολόγηση της απόδοσης του συστήματος. Δυο παραδείγματα προβλημάτων μάθησης χωρίς επίβλεψη είναι η ομαδοποίηση (clustering) και ελάττωση διαστάσεων (dimensionality reduction).

- **Μάθηση με ενίσχυση (Reinforcement learning).** Οι φάσεις εκπαίδευσης και ελέγχου εναλλάσσονται σε αυτό το σενάριο μάθησης. Για τη συλλογή πληροφορίας, το σύστημα μάθησης αλληλοεπιδρά ενεργά με το περιβάλλον και σε κάποιες περιπτώσεις το επηρεάζει, λαμβάνοντας άμεση επιβράβευση για κάθε ενέργεια που πραγματοποιεί. Ο στόχος του συστήματος μάθησης είναι να μεγιστοποιήσει τις επιβραβεύσεις για τις ενέργειες που πραγματοποιεί κατά την αλληλεπίδρασή του με το περιβάλλον. Παρόλα αυτά, δεν παρέχεται μακροπρόθεσμη ανάδραση ως προς τις επιβραβεύσεις, οπότε το σύστημα αντιμετωπίζει το δίλλημα της εξερεύνησης ή εκμετάλλευσης (exploration versus exploitation) διότι θα πρέπει να επιλέξει μεταξύ της εξερεύνησης άγνωστων ενεργειών για να λάβει περισσότερη πληροφορία και της εκμετάλλευσης της ήδη υπάρχουσας πληροφορίας.

Το συγκεκριμένο πρόβλημα ανήκει στην κατηγορία προβλημάτων ταξινόμησης:

Μια υποκατηγορία προβλημάτων με τα οποία ασχολείται η Μηχανική Μάθηση είναι τα προβλήματα Ταξινόμησης (Classification Problems). Στα προβλήματα αυτά το σχήμα μάθησης τροφοδοτείται με ένα σύνολο από δείγματα (δεδομένα εκπαίδευσης) τα οποία έχουν εκ των προτέρων κατανεμηθεί σε συγκεκριμένες κατηγορίες. Από την πληροφορία αυτή, το σύστημα μάθησης θα πρέπει να εξάγει τη δυνατότητα ταξινόμησης νέων δεδομένων που αφορούν το πρόβλημα σε κατηγορίες. Τα νέα δεδομένα δεν ανήκουν στο σύνολο των δεδομένων εκπαίδευσης (Ian H.Witten, 2005).

Η μάθηση Ταξινόμησης αρκετές φορές αποτελεί μάθηση υπό επίβλεψη, διότι για να λειτουργήσει προϋποθέτει ένα σύνολο ταξινομημένων δειγμάτων, αλλά υπάρχουν και περιπτώσεις με μάθηση χωρίς επίβλεψη. Οι **κλάσεις** του προβλήματος είναι οι διαφορετικές κατηγορίες στις οποίες μπορούν ανήκουν τα δεδομένα. Τα συστήματα μάθησης που έχουν εκπαιδευτεί ώστε να επιλύουν κάποιο πρόβλημα ταξινόμησης ονομάζονται **Ταξινομητές (Classifiers)**. Διακρίνονται δυο κατηγορίες προβλημάτων Ταξινόμησης, ανάλογα με το πλήθος των κλάσεων στις οποίες θα πρέπει να κατανεμηθούν τα δεδομένα:

- **Δυαδικά προβλήματα Ταξινόμησης (Binary Classification)**. Τα δεδομένα θα πρέπει να κατανεμηθούν σε δυο κλάσεις, συνήθως με βάση το αν έχουν η όχι κάποια συγκεκριμένη ιδιότητα.

- **Προβλήματα Ταξινόμησης Πολλών Κλάσεων (Multi-class Classification)**. Στην περίπτωση αυτή τα δεδομένα κατηγοριοποιούνται σε περισσότερες από δυο κλάσεις.

Αρκετοί αλγόριθμοι που χρησιμοποιούνται για την κατασκευή ταξινομητών έχουν σχεδιαστεί ώστε να επιλύουν αποκλειστικά δυαδικά προβλήματα ταξινόμησης.

4.1.3 Προ επεξεργασία Δεδομένων

Η προ επεξεργασία των δεδομένων αποσκοπεί στην αντιμετώπιση διάφορων προβλημάτων που προκύπτουν κατά τη διαδικασία συλλογής τους, όπως οι ακραίες τιμές και τιμές που λείπουν για κάποια δείγματα. Επίσης, τα δεδομένα συνήθως μετασχηματίζονται σε κατάλληλη μορφή, ώστε να μπορέσουν να αποτελέσουν κατάλληλη είσοδο για τους αλγορίθμους εκμάθησης. Υπάρχει πλήθος τεχνικών οι οποίες μπορούν να μετασχηματίσουν τα δεδομένα σε κατάλληλη μορφή, όπως μαθηματικοί ή λογικοί μετασχηματισμοί ή μετασχηματισμοί βασισμένοι στην γνώση του εκάστοτε πεδίου.

Κανονικοποίηση

Σε πολλές πρακτικές εφαρμογές θα πρέπει να αντιμετωπιστούν χαρακτηριστικά τα οποία έχουν διαφορετικό εύρος τιμών. Τα χαρακτηριστικά τα οποία παίρνουν μεγαλύτερες τιμές επηρεάζουν περισσότερο την συνάρτηση κόστους σε σχέση με τα χαρακτηριστικά με μικρότερες τιμές, χωρίς όμως αυτό να σημαίνει ότι είναι όντως πιο σημαντικά για το πρόβλημα. Αυτή η δυσκολία μπορεί να ξεπεραστεί με την κανονικοποίηση των χαρακτηριστικών, έτσι ώστε οι τιμές τους να βρίσκονται στο ίδιο εύρος.

Μια άμεση τεχνική για να επιτευχθεί αυτό είναι η **κανονικοποίηση με χρήση εκτιμήσεων για τη μέση τιμή και την διακύμανση**. Για διαθέσιμα δεδομένα, η κανονικοποίηση του χαρακτηριστικού γίνεται ως εξής [Sergios Theodoridis, 2008]:

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik} \ , \quad k = 1, 2, \dots, l \tag{0.1}$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{\iota=1}^{N} (x_{ik} - \bar{x}_k)^2 \tag{0.2}$$

$$\hat{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k} \tag{0.3}$$

Μετά από αυτόν το μετασχηματισμό, όλα τα κανονικοποιημένα χαρακτηριστικά θα έχουν μηδενική μέση τιμή και μοναδιαία διακύμανση. Η παραπάνω μέθοδος είναι γραμμική. Άλλες γραμμικές τεχνικές περιορίζουν τις τιμές των χαρακτηριστικών στο εύρος [0,1] ή στο [-1,1] με κατάλληλη κλιμάκωση (scaling).

Εκτός από τις γραμμικές μεθόδους, μπορούν να εφαρμοστούν και μη γραμμικές μέθοδοι σε περιπτώσεις στις οποίες τα δεδομένα δεν είναι εξίσου κατανεμημένα γύρω από τη μέση τιμή.

Επιλογή χαρακτηριστικών

Στις περισσότερες περιπτώσεις ο σχεδιαστής ενός συστήματος Μηχανικής Μάθησης έχει στην διάθεσή του πάρα πολλά χαρακτηριστικά. Αν και εκ πρώτης όψεως αυτό μπορεί να φαίνεται επιθυμητό, δημιουργεί προβλήματα διότι εισάγει θόρυβο και μειώνει την απόδοση του συστήματος. Κάποια από τα χαρακτηριστικά ενδεχομένως να συσχετίζονται μεταξύ τους, ενώ άλλα χαρακτηριστικά μπορεί να παρέχουν πληροφορία άσχετη προς το συγκεκριμένο πρόβλημα. Επίσης, εάν ο διανυσματικός χώρος των χαρακτηριστικών έχει πολλές διαστάσεις (δηλαδή πολλά χαρακτηριστικά), ο όγκος του χώρου αυτού αυξάνει ιδιαίτερα γρήγορα, οπότε τα δεδομένα για το πρόβλημα θα είναι αραιά κατανεμημένα (sparse), δημιουργώντας προβλήματα στις μεθόδους που προσπαθούν να επιτύχουν στατιστική σημαντικότητα. Το πλήθος των δεδομένων που χρειάζονται ώστε αυτά να θεωρούνται πυκνά αυξάνει εκθετικά σε σχέση με την διάσταση του χώρου χαρακτηριστικών. Το φαινόμενο αυτό είναι γνωστό ως **«η κατάρα της διαστασιμότητας» (curse of dimensionality)**. Θα πρέπει ακόμη να αναφερθεί ότι ένας μεγάλος αριθμός χαρακτηριστικών αυξάνει τον αριθμό των παραμέτρων του συστήματος μάθησης, επομένως και την πολυπλοκότητά του, χωρίς αυτό να σημαίνει ότι θα έχει καλύτερη απόδοση [Sergios Theodoridis, 2008].

Το πρόβλημα που θα πρέπει να επιλυθεί περιγράφεται ως εξής: δοθέντος ενός αριθμού από χαρακτηριστικά, πως μπορεί κανείς να επιλέξει τα πιο σημαντικά από αυτά ώστε να μειώσει τον αριθμό τους και παράλληλα να διατηρήσει όσο το δυνατό περισσότερη χρήσιμη πληροφορία. Η διαδικασία αυτή ονομάζεται **επιλογή χαρακτηριστικών (feature selection)** ή αλλιώς **ελάττωση χαρακτηριστικών (feature reduction)**. Το βήμα αυτό είναι κρίσιμο διότι αν επιλεχθούν χαρακτηριστικά με μικρή διαχωριστική ικανότητα, το σύστημα Μάθησης που θα προκύψει δεν θα έχει ικανοποιητική απόδοση. Αν όμως επιλεχθούν χαρακτηριστικά που παρέχουν χρήσιμη πληροφορία, το σύστημα που θα σχεδιαστεί θα είναι απλό και αποτελεσματικό.

Μια από τις στρατηγικές που μπορεί να ακολουθηθεί είναι να εξεταστούν τα χαρακτηριστικά ένα προς ένα μέσω ενός μέτρου διαχωρισιμότητας των κλάσεων και να απορριφθούν εκείνα που διαθέτουν μικρή διαχωριστική ικανότητα. Ο στόχος είναι η επιλογή των χαρακτηριστικών αυτών που οδηγούν σε μεγάλες αποστάσεις μεταξύ των ομάδων των δειγμάτων και μικρή διακύμανση μεταξύ της ίδιας ομάδας. Αυτό σημαίνει ότι τα χαρακτηριστικά θα πρέπει να παίρνουν μακρινές τιμές για διαφορετικές κλάσεις και κοντινές τιμές για την ίδια κλάση. Η στρατηγική αυτή είναι γνωστή ως «**Βαθμωτή επιλογή χαρακτηριστικών**» ή «**φιλτράρισμα**» **(Filtering)** και ακολουθήθηκε στην παρούσα εργασία με μέτρο την καμπύλη ROC, όπου θα περιγραφεί στη συνέχεια.

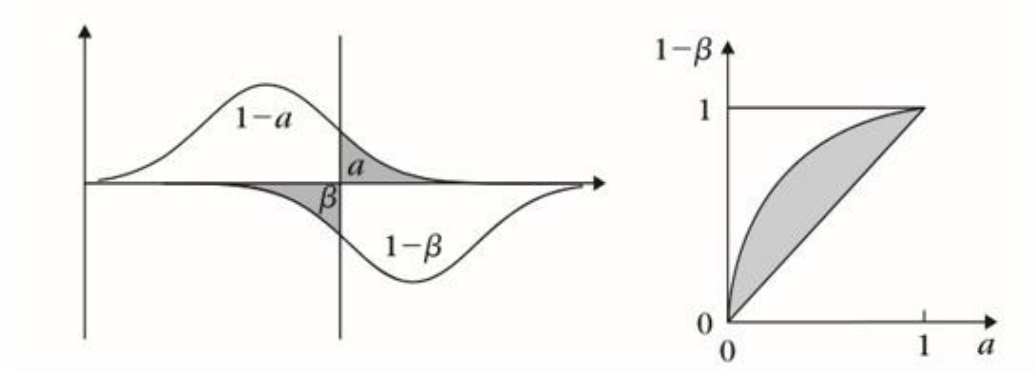
Στην προσέγγιση αυτή τα χαρακτηριστικά αντιμετωπίζονται χωριστά. Μπορεί να χρησιμοποιηθεί οποιοδήποτε από τα κριτήρια που μετρούν τη διαχωριστική ικανότητα

των χαρακτηριστικών όπως οι καμπύλες ROC, το κριτήριο FDR, η μονοδιάστατη απόκλιση (Divergence) και άλλα. Η τιμή C(k) για το κάθε κριτήριο υπολογίζεται για κάθε χαρακτηριστικό k, με k=1, 2, … ,m και στη συνέχεια τα κριτήρια ταξινομούνται σε φθίνουσα σειρά ως προς C(k). Τα l χαρακτηριστικά που αντιστοιχούν στις l καλύτερες τιμές C(k) του επιλέγονται ώστε να σχηματιστεί το υποσύνολο χαρακτηριστικών [8]. Η τεχνική αυτή είναι γνωστή ως φιλτράρισμα (filtering). Σύμφωνα με αυτή τη προσέγγιση, η επιλογή των χαρακτηριστικών είναι ανεξάρτητη από το είδος του συστήματος Μάθησης. Το μεγαλύτερο πλεονέκτημα που παρέχει η βαθμωτή επιλογή χαρακτηριστικών είναι οι απλοί υπολογισμοί, χωρίς όμως να λαμβάνει υπόψη τις συσχετίσεις που υπάρχουν μεταξύ των χαρακτηριστικών. Υπάρχουν τεχνικές που μπορούν να εισάγουν πληροφορίες συσχέτισης για τη βαθμωτή επιλογή χαρακτηριστικών (Sergios Theodoridis, 2008).

Καμπύλες ROC

Πολλοί έλεγχοι υποθέσεων παρέχουν στατιστικά στοιχεία σχετικά με τις διαφορές των μέσων τιμών ενός μόνο χαρακτηριστικού στις διάφορες κλάσεις. Αν και αυτή η πληροφορία είναι χρήσιμη, σε περίπτωση που οι αντίστοιχες μέσες τιμές είναι κοντά, δεν επαρκεί για να εγγυηθεί καλή διαχωριστική ικανότητα για ένα χαρακτηριστικό που δεν θα απορριφθεί από τον έλεγχο. Επίσης, υπάρχει η περίπτωση όπου οι μέσες τιμές μπορεί να διαφέρουν σημαντικά αλλά η διασπορά γύρω από αυτές να είναι μεγάλη, οπότε δεν μπορεί να γίνει εύκολη διάκριση των κλάσεων.

Στην Εικόνα 4.2 (Sergios Theodoridis, 2008), στη γραφική παράσταση αριστερά παρουσιάζεται ένα παράδειγμα δυο επικαλυπτόμενων συναρτήσεων πυκνότητας πιθανότητας που περιγράφουν την κατανομή ενός χαρακτηριστικού σε δύο κλάσεις, μαζί με ένα κατώφλι (η μια συνάρτηση πυκνότητας πιθανότητας σχεδιάστηκε ανάστροφα για να είναι περισσότερο ευδιάκριτη). Οι τιμές αριστερά του κατωφλίου ανήκουν στην κλάση ω_1 και οι τιμές δεξιά στην κλάση ω_2. Αυτή η απόφαση μπορεί να είναι **λανθασμένη με πιθανότητα α**, παράγοντας λανθασμένο συμπέρασμα για την κλάση ω_1 (η πιθανότητα μιας σωστής απόφασης θα είναι 1-α). Η πιθανότητα της λανθασμένης απόφασης είναι γραμμοσκιασμένη κάτω από την αντίστοιχη καμπύλη. Ομοίως, έστω η πιθανότητα λανθασμένης απόφασης β σχετικά με την κλάση ω_1 και 1-β **η πιθανότητα σωστής απόφασης** σχετικά με την κλάση ω_2.



Εικόνα 0.2 Αριστερά: Επικαλυπτόμενες Κατανομές πυκνότητας πιθανότητας δύο κλάσεων. Δεξιά: Αντίστοιχη καμπύλη ROC

Αν οι δύο κατανομές επικαλύπτονται πλήρως, τότε για οποιαδήποτε θέση του κατωφλίου θα ισχύει ότι α = 1 - β. Αυτή η περίπτωση αντιστοιχεί στην ευθεία γραμμή, στο δεξιό σχήμα της Εικόνας 4.2, όπου οι δυο άξονες είναι το α και το 1 - β. Καθώς οι δυο κατανομές απομακρύνονται, η αντίστοιχη καμπύλη στο δεξί σχήμα διαφοροποιείται από την ευθεία. Όσο λιγότερο επικαλύπτονται οι κλάσεις τόσο μεγαλύτερο είναι το εμβαδόν της περιοχής μεταξύ της καμπύλης και της ευθείας γραμμής. Στην άλλη ακραία περίπτωση όπου οι δυο κατανομές είναι πλήρως διαχωρισμένες, αν το κατώφλι μετακινείται ώστε να καλύψει όλο το εύρος των τιμών του α στο [0,1], το 1 - β παραμένει ίσο με τη μονάδα. Επομένως, το εμβαδόν της περιοχής της καμπύλης του δεύτερου σχήματος θα κυμαίνεται από μηδέν για κατανομές που επικαλύπτονται πλήρως έως 1⁄2 για πλήρη διαχωρισμό (το εμβαδό του τριγώνου που σχηματίζεται), και είναι ένα μέτρο της διαχωριστικής ικανότητας του συγκεκριμένου χαρακτηριστικού που εξετάζεται. Η καμπύλη αυτή ονομάζεται **καμπύλη ROC (Receiver Operating Characteristic)** και σχεδιάζεται εύκολα στην πράξη μετακινώντας το κατώφλι για τις πιθανότητες των σωστών και λανθασμένων αποφάσεων για όλα τα χαρακτηριστικά του συνόλου εκπαίδευσης (Sergios Theodoridis, 2008).

4.2 Γραμμική Διαχωριστική Ανάλυση

Η Γραμμική Διαχωριστική Ανάλυση (Linear Discriminant Analysis – LDA) είναι μια βασική μέθοδος ταξινόμησης η οποία προτάθηκε από τον R. Fisher (Petros Xanthopoulos, 2013). Η βασική ιδέα της τεχνικής αυτής είναι ο προσδιορισμός ενός διανυσματικού υποχώρου μικρότερης διάστασης από τον αρχικό διανυσματικό χώρο των δειγμάτων, στον οποίο τα δεδομένα του προβλήματος είναι γραμμικώς διαχωρίσιμα. Η Διαχωρισιμότητα ορίζεται και παρατηρείται με τη βοήθεια στατιστικών μέτρων όπως η μέση τιμή και η διακύμανση.

Ένα από τα πλεονεκτήματα της μεθόδου είναι ότι η λύση προκύπτει επιλύοντας ένα γενικευμένο πρόβλημα ιδιοτιμών, επιτρέποντας τη γρήγορη και μαζική επεξεργασία των δειγμάτων του προβλήματος. Επίσης, υπάρχει η δυνατότητα επέκτασης σε μη Γραμμική Διαχωριστική Ανάλυση, χρησιμοποιώντας συναρτήσεις Πυρήνα. Ο αρχικός αλγόριθμος της μεθόδου είχε σχεδιαστεί για δυαδικά προβλήματα ταξινόμησης αλλά στη συνέχεια προτάθηκαν διάφορες γενικεύσεις που αντιμετωπίζουν προβλήματα με πολλές κλάσεις (Petros Xanthopoulos, 2013).

Έστω $x_1, \dots, x_p \in \mathbb{R}^m$ ένα σύνολο από p δείγματα, τα οποία μπορούν να ανήκουν σε δυο διαφορετικές κλάσεις, A και B. Για κάθε κλάση ορίζεται η μέση τιμή των δειγμάτων:

$$\bar{x}_A = \frac{1}{N_A}\sum_{x\in A} x \qquad \bar{x}_B = \frac{1}{N_B}\sum_{x\in B} x \tag{0.4}$$

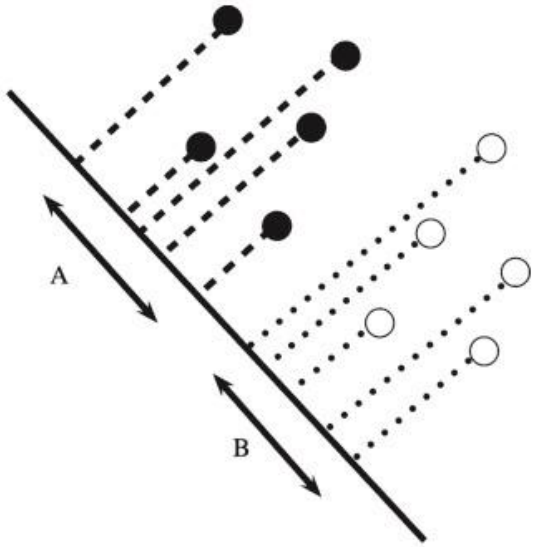
όπου , N_A, N_B το πλήθος των δειγμάτων στην κλάση A και B αντίστοιχα. Για κάθε κλάση μπορούν να οριστούν οι θετικά ημι-ορισμένοι Πίνακες Διασποράς (Scatter Matrices), οι οποίοι περιγράφονται από τις εξισώσεις:

$$S_A = \sum_{x\in A}(x-\bar{x}_A)(x-\bar{x}_A)^T \tag{0.5}$$

$$S_B = \sum_{x\in B}(x-\bar{x}_B)(x-\bar{x}_B)^T \tag{0.6}$$

Οι Πίνακες S_A και S_B εκφράζουν την μεταβλητότητα στην αντίστοιχη κλάση. Ιδανικά, θα πρέπει να βρεθεί **ένα υπερεπίπεδο,** το οποίο ορίζεται από το διάνυσμα , **στο οποίο οι προβολές των δειγμάτων έχουν την ελάχιστη διακύμανση.** Αυτό μπορεί να εκφραστεί μαθηματικά ως εξής (Petros Xanthopoulos, 2013):

$$\min_{\varphi}(\varphi^T S_A \varphi + \varphi^T S_B \varphi) = \min_{\varphi} \varphi^T (S_A + S_B)\,\varphi = \min_{\varphi} \varphi^T S \varphi \tag{0.7}$$



Εικόνα 0.3 Γραμμική Διαχωριστική Ανάλυση: Τα δείγματα δυο διαστάσεων προβάλλονται σε ένα χώρο χαμηλότερων διαστάσεων (ευθεία γραμμή). Η ευθεία θα πρέπει να επιλεγεί έτσι ώστε η προβολή να μεγιστοποιεί τη διαχωρισιμότητα των προβαλλόμενων δειγμάτων *όπου* $S = S_A + S_B$.

Ο Πίνακας Διασποράς μεταξύ των δυο κλάσεων δίνεται από την εξίσωση:

$$S_{AB} = (\bar{x}_A - \bar{x}_B)(\bar{x}_A - \bar{x}_B)^T \tag{0.8}$$

Σύμφωνα με την ιδέα του Fisher, θα πρέπει να βρεθεί ένα υπερεπίπεδο έτσι ώστε να μεγιστοποιηθεί η απόσταση μεταξύ των μέσων των δυο κλάσεων και ταυτόχρονα να ελαχιστοποιηθεί η διακύμανση σε κάθε κλάση. Η μαθηματική έκφραση αυτής της ιδέας είναι το κριτήριο μεγιστοποίησης του Fisher:

$$\max_{\varphi} \mathcal{F}(\varphi) = \max_{\varphi} \frac{\varphi^T S_{AB} \varphi}{\varphi^T S \varphi} \tag{0.9}$$

Αυτό το πρόβλημα βελτιστοποίησης μπορεί να έχει άπειρο πλήθος λύσεων με την ίδια τιμή αντικειμενικής συνάρτησης, αφού για μια λύση φ^* όλα τα διανύσματα $c \cdot \varphi^*$ έχουν την ίδια τιμή. Αν, χωρίς απώλεια της γενικότητας, αντικατασταθεί ο παρανομαστής με ένα περιορισμό ισότητας ώστε να επιλεγεί μια μόνο λύση, το πρόβλημα θα είναι:

$$\max_{\varphi} \varphi^T S_{AB} \varphi \qquad (0.10)$$

$$\text{με τον περιορισμό: } \varphi^T S \varphi = 1 \qquad (0.11)$$

Η συνάρτηση Lagrange για το συγκεκριμένο πρόβλημα βελτιστοποίησης είναι η εξής:

$$\mathcal{L}_{LDA}(x, \lambda) = \varphi^T S_{AB} \varphi - \lambda(\varphi^T S \varphi - 1) \qquad (0.12)$$

όπου λ είναι ο πολλαπλασιαστής Lagrange που σχετίζεται με τον περιορισμό. Αφού ο S_{AB} είναι θετικά ημι-ορισμένος το πρόβλημα βελτιστοποίησης είναι κυρτό, οπότε το ολικό μέγιστο θα είναι στο σημείο όπου ισχύει (Petros Xanthopoulos, 2013):

$$\frac{\partial \mathcal{L}_{LDA}(x, \lambda)}{\partial x} = 0 \Leftrightarrow S_{AB} \varphi - \lambda S \varphi = 0 \qquad (0.13)$$

Το βέλτιστο φ μπορεί να προκύψει ως **το ιδιοδιάνυσμα που αντιστοιχεί στη μικρότερη ιδιοτιμή** του παρακάτω γενικευμένου ιδιοσυστήματος:

$$S_{AB} \varphi = \lambda S \varphi \qquad (0.14)$$

Η Γραμμική Διαχωριστική Ανάλυση **για προβλήματα πολλών κλάσεων** είναι μια επέκταση της περίπτωσης που αναλύθηκε. Θα πρέπει να οριστούν οι Πίνακες Διασποράς για τις n διαθέσιμες κλάσεις. Το άθροισμα με τους Πίνακες Διασποράς των κλάσεων είναι:

$$S = S_1 + S_2 + \cdots + S_n \qquad (0.15)$$

ενώ ο Πίνακας Διασποράς που ορίζεται μεταξύ των κλάσεων θα είναι ο εξής:

$$S_{1,\ldots,n} = \sum_{i=1}^{n} p_i (\bar{x}_i - \bar{x}) (\bar{x}_i - \bar{x})^T \qquad (0.16)$$

όπου p_i είναι ο αριθμός των δειγμάτων στην στη i-στη κλάση, $\bar{x}\_i$ είναι ο μέσος όρος κάθε κλάσης, και $\bar{x}$ ο καθολικός μέσος όρος ο οποίος υπολογίζεται ως εξής:

$$\bar{x} = \frac{1}{p} \sum_{i=1}^{n} p_i \bar{x}_i \qquad (0.17)$$

Ο γραμμικός μετασχηματισμός που πρέπει να βρεθεί προκύπτει λύνοντας το γενικευμένο πρόβλημα ιδιοτιμής:

$$S_{1,\ldots,n} \varphi = \lambda S \varphi \qquad (0.18)$$

Η Γραμμική Διαχωριστική Ανάλυση μπορεί να χρησιμοποιηθεί για ανίχνευση των πιο σημαντικών χαρακτηριστικών, με τη βοήθεια του αντίστοιχου συντελεστή στο υπερεπίπεδο προβολής και να ταξινομήσει νέα δείγματα. Εφόσον βρεθεί ο μετασχηματισμός φ, η ταξινόμηση μπορεί να γίνει στον μετασχηματισμένο χώρο χαμηλότερων διαστάσεων με τη χρήση κάποιου μέτρου απόστασης d. Η κλάση κάποιου νέου σημείου z προσδιορίζεται ως εξής:

$$\kappa\lambda\acute{\alpha}\sigma\eta(z) = arg\min_{n}\{d(z\varphi, \bar{x}_n\varphi)\} \qquad (0.19)$$

όπου $\bar{x}_n$ το κέντρο της n-στης κλάσης. Συνεπώς, πρώτα θα πρέπει να προβληθούν τα κέντρα όλων των κλάσεων και τα άγνωστα σημεία στον υποχώρο που ορίζεται από τον φ και στη συνέχεια τα σημεία αυτά αναθέτονται στην πλησιέστερη κλάση ως προς (Petros Xanthopoulos, 2013).

Η μέθοδος της Γραμμικής Διαχωριστικής Ανάλυσης παράγει το πολύ m-1 προβολές χαρακτηριστικών (όπου m η διάσταση του χώρου εισόδου). Εάν το σφάλμα ταξινόμησης υποδεικνύει ότι θα πρέπει να χρησιμοποιηθούν επιπλέον χαρακτηριστικά, θα πρέπει να εφαρμοστεί κάποια άλλη μέθοδος ώστε να τα παρέχει. Επίσης, η μέθοδος αυτή είναι παραμετρική, καθώς υποθέτει μονοτροπικές, Γκαουσιανές συναρτήσεις Πιθανοφάνειας (unimodal Gaussian likelihoods). Επομένως, εάν οι κατανομές των δειγμάτων είναι σημαντικά μη-Γκαουσιανές, οι προβολές που θα προκύψουν ενδεχομένως να μην διατηρήσουν την περιπλοκή δομή της κατανομής των δειγμάτων, η οποία είναι απαραίτητη για την ταξινόμησή τους. Εάν η πληροφορία για τον διαχωρισμό των κλάσεων βασίζεται στην διακύμανση και όχι στο μέσο όρο, η Γραμμική Διαχωριστική Ανάλυση ενδεχομένως να μην αποδώσει σε αυτό το πρόβλημα (Gutierrez-Osuna).

4.3 Νευρωνικά Δίκτυα

4.3.1 Εισαγωγή

Τα **Νευρωνικά Δίκτυα** είναι συστήματα που έχουν σχεδιαστεί ώστε να μοντελοποιούν τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος πραγματοποιεί μια συγκεκριμένη λειτουργία ή εργασία. Για να επιτύχουν υψηλή απόδοση, τα νευρωνικά δίκτυα εμπεριέχουν διασυνδέσεις απλούστερων υπολογιστικών

κυττάρων, τα οποία ονομάζονται **νευρώνες** (neurons) ή μονάδες επεξεργασίας (processing units) (Haykin S. , 1999). Εξαιτίας της αναλογίας τους με τον ανθρώπινο εγκέφαλο, για την περιγραφή τους χρησιμοποιούνται όροι από το επιστημονικό πεδίο της Νευροβιολογίας.

Μπορούμε να υιοθετήσουμε τον παρακάτω ορισμό για το Νευρωνικό Δίκτυο, θεωρώντας το ως **μια μηχανή που έχει την δυνατότητα να προσαρμόζεται** (adaptive machine) (Haykin S. , 1999):

Ένα Νευρωνικό Δίκτυο είναι ένας επεξεργαστής που λειτουργεί παράλληλα και κατανεμημένα σε ευρεία κλίμακα. Έχει προκύψει από απλούστερες μονάδες επεξεργασίας, και έχει την δυνατότητα να αποθηκεύει εμπειρική γνώση και να την διαθέτει προς χρήση. Φέρει ομοιότητες με τον ανθρώπινο εγκέφαλο στα εξής σημεία:

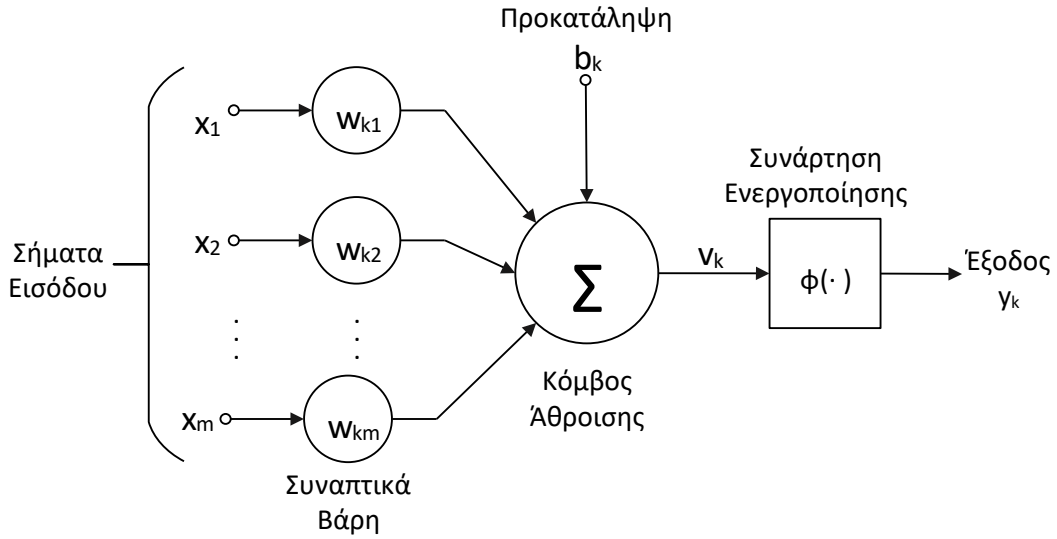
1. Το δίκτυο αποκτά γνώση από το περιβάλλον του μέσω μιας διαδικασίας εκμάθησης.
2. Οι συνδέσεις μεταξύ των νευρώνων χαρακτηρίζονται από το πόσο ισχυρές είναι. Αυτό εκφράζεται αναθέτοντας στις συνδέσεις αυτές βάρη, που ονομάζονται **συναπτικά βάρη** (synaptic weights). Αυτή η παράμετρος των νευρωνικών συνδέσεων χρησιμοποιείται για να καταστεί δυνατή η αποθήκευση της προσλαμβανόμενης γνώσης.

Μια μεγάλη κατηγορία νευρωνικών δικτύων έχει τη δυνατότητα να πραγματοποιεί χρήσιμους υπολογισμούς μέσω διαδικασιών εκμάθησης, οι οποίες υλοποιούνται με τον **αλγόριθμο εκμάθησης** (learning algorithm). Ο αλγόριθμος εκμάθησης τροποποιεί τα συναπτικά βάρη των συνδέσεων του δικτύου με συγκεκριμένο τρόπο για να επιτύχει ένα συγκεκριμένο στόχο σχεδίασης για το δίκτυο.

Η τροποποίηση των συναπτικών βαρών είναι η παραδοσιακή προσέγγιση για την σχεδίαση των νευρωνικών δικτύων. Αυτή η μέθοδος προσεγγίζει την θεωρία των γραμμικών προσαρμοστικών φίλτρων (linear adaptive filter theory), η οποία ήδη έχει εδραιωθεί και εφαρμοστεί επιτυχώς σε διάφορα επιστημονικά πεδία (Haykin S. , 1999; Stearns, 1985). Ωστόσο, ένα νευρωνικό δίκτυο έχει την δυνατότητα να τροποποιεί την τοπολογία του, όπως ακριβώς συμβαίνει και με τον ανθρώπινο εγκέφαλο, όπου κάποιοι νευρώνες πεθαίνουν και νέες συνδέσεις μεταξύ των νευρώνων μπορούν να αναπτυχθούν.

Περιγραφή Νευρώνα

Ο νευρώνας είναι η βασική μονάδα επεξεργασίας και είναι απαραίτητος για τη λειτουργία του νευρωνικού δικτύου. Στο παρακάτω σχήμα (Εικόνα 4.4) βλέπουμε το μοντέλο ενός νευρώνα, το οποίο αποτελεί την βάση για τον σχεδιασμό των νευρωνικών δικτύων (Haykin S. , 1999):



Εικόνα 0.4 Μοντέλο Νευρώνα

Όπως φαίνεται και στο παραπάνω σχήμα, τα βασικά στοιχεία ενός νευρώνα είναι τα εξής (Haykin S. , 1999):

1. Το σύνολο των συνδέσεων των σημάτων εισόδου x_i, i = 1…m με τον νευρώνα k, οι οποίες χαρακτηρίζονται από τα βάρη w_{kj}. Το βάρος w_{kj} αναφέρεται στην σύνδεση του σήματος εισόδου x_j με τον νευρώνα k. Τα βάρη μπορούν να πάρουν θετικές και αρνητικές τιμές.
2. Ο αθροιστής που αθροίζει τα σήματα εισόδου, όπως αυτά προκύπτουν μετά από τον πολλαπλασιασμό τους με τα αντίστοιχα βάρη. Ο αθροιστής αυτός σχηματίζει έναν γραμμικό συνδυαστή (linear combiner).
3. Η συνάρτηση ενεργοποίησης φ(·) (activation function), η οποία καθορίζει την έξοδο του νευρώνα k και περιορίζει το εύρος της εξόδου σε πεπερασμένες τιμές. Το κανονικοποιημένο εύρος τιμών εξόδου είναι το κλειστό διάστημα [0, 1] ή το [-1,1].

Από το παραπάνω μοντέλο προκύπτουν οι παρακάτω εξισώσεις (Haykin S. , 1999):

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \qquad (0.20)$$

$$y_k = \varphi(u_k + b_k) \qquad (0.21)$$

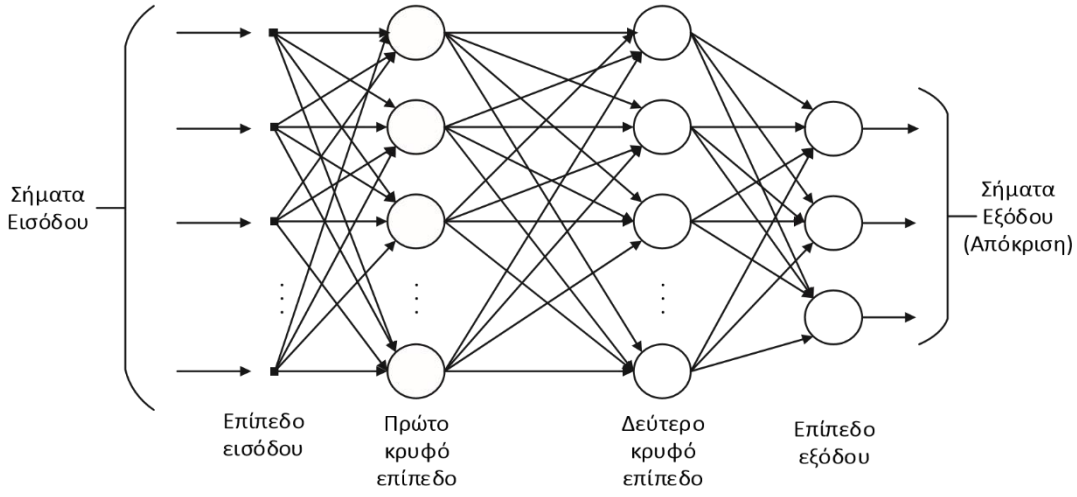
όπου $x_1, x_2, \dots, x_m$ είναι τα σήματα εισόδου, $w_{k1}, w_{k2}, \dots, w_{km}$ τα βάρη του νευρώνα k, u_k είναι η έξοδος του γραμμικού συνδυαστή (γραμμικός συνδυασμός των σημάτων εξόδου), b_k η προκατάληψη (bias) του μοντέλου, $\varphi(\cdot)$ είναι η συνάρτηση ενεργοποίησης και y_kτο σήμα εξόδου του νευρώνα. Η προκατάληψη (bias) του μοντέλου προκαλεί ένα συγγενή μετασχηματισμό (affine transformation) στην έξοδο του γραμμικού συνδυαστή του μοντέλου, ως εξής:

$$v_k = u_k + b_k \qquad (0.22)$$

Πολυεπίπεδα Perceptrons

Οι νευρώνες μπορούν να συνδυαστούν, σε διαφορετικά επίπεδα, και σε διάφορες συνδεσμολογίες, σχηματίζοντας δίκτυα νευρώνων. Τα **πολυ-επίπεδα Perceptrons (Multilayer Perceptrons, MLPs)** είναι δίκτυα που αποτελούνται από ένα σύνολο κόμβων εισόδου, οι οποίοι αποτελούν το **επίπεδο εισόδου (input layer)**, από ένα ή περισσότερα **κρυφά επίπεδα νευρώνων (hidden layers)** που περιέχουν νευρώνες για την ολοκλήρωση των υπολογισμών και ένα **επίπεδο εξόδου (output layer)** που περιέχει τους νευρώνες που δίνουν την έξοδο του δικτύου [Haykin S. , 1999].

Η αρχιτεκτονική του δικτύου φαίνεται στο παρακάτω σχήμα (Εικόνα 4.5) [Haykin S. , 1999]:



Εικόνα 0.5 Αρχιτεκτονική μοντέλου Perceptron πολλαπλών επιπέδων

Σε αυτό το δίκτυο υπάρχουν δυο είδη σημάτων:

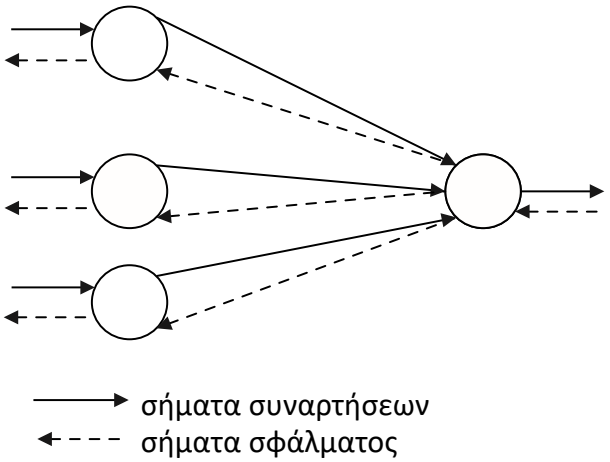
1. Σήματα Συναρτήσεων

Το σήμα συνάρτησης είναι ένα σήμα εισόδου το οποίο εφαρμόζεται στην είσοδο του δικτύου, διαδίδεται προς τα εμπρός στο δίκτυο, από επίπεδο σε επίπεδο μέσω των νευρώνων και εμφανίζεται στην έξοδο του δικτύου ως σήμα εξόδου. Ονομάζεται σήμα συνάρτησης επειδή γίνεται η υπόθεση ότι υλοποιεί μια χρήσιμη συνάρτηση στην έξοδο του δικτύου. Επίσης, σε κάθε νευρώνα από τον οποίο περνάει το σήμα, υπολογίζεται ως μια συνάρτηση των εισόδων και των σχετικών βαρών που εφαρμόζονται στον συγκεκριμένο νευρώνα. Τα σήματα συνάρτησης επίσης αναφέρονται και ως σήματα εισόδου.

2. Σήματα σφάλματος

Το σήμα σφάλματος παράγεται στην έξοδο του δικτύου και διαδίδεται προς τα πίσω από επίπεδο σε επίπεδο στο δίκτυο. Ονομάζεται σήμα σφάλματος διότι ο υπολογισμός του σε κάθε νευρώνα του δικτύου συμπεριλαμβάνει κάποια συνάρτηση που εξαρτάται από τη διαφορά της εξόδου του δικτύου από την επιθυμητή απόκριση (σφάλμα).

Οι ροές των παραπάνω σημάτων στο δίκτυο φαίνονται στο παρακάτω σχήμα (Εικόνα 4.6) (Haykin S. , 1999):



Εικόνα 0.6 Ροές σήματος στο Perceptron πολλαπλών επιπέδων

Κάθε νευρώνας σε ένα perceptron πολλαπλών επιπέδων, που βρίσκεται σε κρυφό επίπεδο ή στο επίπεδο εξόδου, έχει σχεδιαστεί ώστε να πραγματοποιεί δυο ειδών υπολογισμούς:

1. Τον υπολογισμό του σήματος συνάρτησης που εμφανίζεται στην έξοδο του νευρώνα, το οποίο εκφράζεται ως μια συνεχής μη γραμμική συνάρτηση του σήματος εισόδου και των βαρών που σχετίζονταιμε αυτόν τον νευρώνα.

2. Τον υπολογισμό μιας εκτίμησης του διανύσματος κλίσης (πχ την κλίση της επιφάνειας του σφάλματος σε σχέση με τα βάρη που συνδέονται στην είσοδο του νευρώνα), ο οποίος δημιουργεί μια προς τα πίσω ροή σήματος στο δίκτυο.

4.3.2 Μέθοδος οπισθοδιάδοσης σφάλματος (Back-propagation / BP)

Ο πιο κοινός αλγόριθμος προσαρμογής βαρών για τη βελτιστοποίηση της απόδοσης του νευρωνικού δικτύου ονομάζεται backpropagation και ανήκει στις μεθόδους μάθησης υπό επίβλεψη. Το αποτέλεσμα του νευρωνικού δικτύου (για απλοποίηση θεωρείται ένα κρυφό στρώμα με συνάρτηση ενεργοποίησης Tanh) δίνεται από:

$$y_{\mathrm{ANN}} = \sum_{j=1}^{n_{\mathrm{h}}} y_j^{(2)} w_{j1}^{(2)} = \sum_{j=1}^{n_{\mathrm{h}}} \tanh\left(\sum_{i=1}^{n_{\mathrm{var}}} x_i w_{ij}^{(1)}\right) \cdot w_{j1}^{(2)} , \qquad (0.23)$$

όπου n_{var} και n_h είναι ο αριθμός των νευρώνων στο στρώμα εισόδου και στο κρυφό στρώμα αντιστοίχως, $w_{ij}^{(1)}$ είναι τα βάρη μεταξύ του στρώματος εισόδου i και του κρυφού στρώματος j, και $w_{j1}^{(2)}$ είναι τα βάρη μεταξύ του κρυφού στρώματος j και του νευρώνα εξόδου.

Κατά τη διάρκεια της εκπαίδευσης το νευρωνικό δίκτυο τροφοδοτείται με N γεγονότα εκπαίδευσης $x_a = (x_{1,} \ldots, x_{nvar})_{a,}$ a = 1,…,N. Σε κάθε γεγονός εκπαίδευσης α υπολογίζεται το αποτέλεσμα του νευρωνικού δικτύου $y_{ANN,a}$ και συγκρίνεται με το επιθυμητό αποτέλεσμα $\hat{y}_a \in \{1,0\}$ (1 για γεγονός σήματος και 0 για γεγονός θορύβου). Η *συνάρτηση σφάλματος E* που υπολογίζει τη συμφωνία του αποτελέσματος του νευρωνικού δικτύου με το επιθυμητό αποτέλεσμα ορίζεται ως:

$$E(\mathbf{x}_1, \ldots, \mathbf{x}_N | \mathbf{w}) = \sum_{a=1}^{N} E_a(\mathbf{x}_a | \mathbf{w}) = \sum_{a=1}^{N} \frac{1}{2} \left(y_{\mathrm{ANN},a} - \hat{y}_a\right)^2 , \qquad (0.24)$$

με w το σύνολο των μεταβαλλόμενων βαρών στο δίκτυο. Το σύνολο των βαρών που ελαχιστοποιούν τη συνάρτηση σφάλματος υπολογίζεται με τη μέθοδο κυρτής βελτιστοποίησης (steepest or gradient descent). Ξεκινώντας από έναν τυχαίο

συνδυασμό βαρών $w^{(\rho)}$ τα βάρη ενημερώνονται κινούμενα σταδιακά προς την κατεύθυνση $-\nabla_w E$ στον w χώρο, για την οποία το E μειώνεται πιο γρήγορα

$$\mathbf{w}^{(\rho+1)} = \mathbf{w}^{(\rho)} - \eta \boldsymbol{\nabla}_{\mathbf{w}} E\,, \tag{0.25}$$

όπου ο θετικός αριθμός η είναι ο ρυθμός μάθησης.

$$\mathbf{w}^{(\rho+1)} = \mathbf{w}^{(\rho)} - \eta \boldsymbol{\nabla}_{\mathbf{w}} E\,, \tag{0.26}$$

Τα βάρη που συνδέονται με το στρώμα εξόδου ενημερώνονται με τη συνάρτηση

$$\Delta w_{j1}^{(2)} = -\eta \sum_{a=1}^{N} \frac{\partial E_a}{\partial w_{j1}^{(2)}} = -\eta \sum_{a=1}^{N} \left(y_{\mathrm{ANN},a} - \hat{y}_a\right) y_{j,a}^{(2)}\,, \tag{0.27}$$

Και τα βάρη που συνδέονται με τα κρυφά στρώματα ενημερώνονται με τη συνάρτηση

$$\Delta w_{ij}^{(1)} = -\eta \sum_{a=1}^{N} \frac{\partial E_a}{\partial w_{ij}^{(1)}} = -\eta \sum_{a=1}^{N} \left(y_{\mathrm{ANN},a} - \hat{y}_a\right) y_{j,a}^{(2)} (1 - y_{j,a}^{(2)}) w_{j1}^{(2)} x_{i,a}\,, \tag{0.28}$$

με χρήση tanh'x = tanhx(1-tanhx). Αυτή η μέθοδος εκπαίδευσης ονομάζεται bulk learning, καθώς το άθροισμα των σφαλμάτων από όλα τα γεγονότα εκπαίδευσης χρησιμοποιείται για την ενημέρωση των βαρών. Το TMVA χρησιμοποιεί μία εναλλακτική μέθοδο μάθησης, την online μάθηση, στην οποία η ενημέρωση των βαρών πραγματοποιείται σε κάθε γεγονός. Οι ενημερώσεις των βαρών προκύπτουν από τις παραπάνω εξισώσεις αφαιρώντας τα αθροίσματα κάθε γεγονότος. Σε αυτή την περίπτωση είναι σημαντική η χρήση ενός τυχαίου δείγματος μάθησης.

4.3.3 BFGS

Η μέθοδος Broyden-Fletcher-Goldfarb-Shannon (BFGS) [Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shannon, 1970] διαφέρει από τη μέθοδο οπισθοδιάδοσης σφάλματος (back-propagation) στη χρήση παραγώγων δεύτερης τάξης της συνάρτησης σφάλματος για την προσαρμογή των βαρών σύναψης. Για το σκοπό αυτό χρησιμοποιείται ένας αλγόριθμος τεσσάρων βημάτων:

1. Υπολογισμός δύο διανυσμάτων, D και Y. Το διάνυσμα αλλαγής βαρών D αντιπροσωπεύει τη μεταβολή μεταξύ μίας επανάληψης (κ-1) με την επόμενη (κ). Κάθε στοιχείο του διανύσματος αντιστοιχεί σε ένα βάρος σύναψης. Το διάνυσμα Y είναι το διάνυσμα σφαλμάτων κλίσης.

$$D_i^{(k)} = w_i^{(k)} - w_i^{(k-1)}, \qquad (0.29)$$

$$Y_i^{(k)} = g_i^{(k)} - g_i^{(k-1)}, \qquad (0.30)$$

, όπου i είναι ο δείκτης της σύναψης, g_i η i-οστή κλίση*, w_i το βάρος της i-οστης σύναψης, και k μετρητής του αριθμού των επαναλήψεων.

2. Προσέγγιση του ανάστροφου Hessian πίνακα, H^{-1} στην k επανάληψη,

$$H^{-1(k)} = \frac{D \cdot D^T \cdot (1 + Y^T \cdot H^{-1(k-1)} \cdot Y)}{Y^T \cdot D} - D \cdot Y^T \cdot H + H \cdot Y \cdot D^T + H^{-1(k-1)}, \qquad (0.31)$$

,όπου οι εκθέτες (k) υπονοούνται για τα διανύσματα D και Y.

3. Υπολογισμός του διανύσματος αλλαγής βαρών

$$D^{(k)} = -H^{-1(k)} \cdot Y^{(k)} . \qquad (0.32)$$

4. Υπολογισμός ενός νέου διανύσματος βαρών με εφαρμογή του αλγορίθμου διερεύνησης γραμμής (line search), όπου η συνάρτηση σφάλματος προσεγγίζεται τοπικά από μία παραβολή. Ο αλγόριθμος υπολογίζει τις δεύτερες παραγώγους και καθορίζει το αναμενόμενο ελάχιστο σημείο της παραβολής. Το συνολικό σφάλμα υπολογίζεται γι' αυτό το σημείο. Στη συνέχεια ο αλγόριθμος υπολογίζει τα σημεία της γραμμής που καθορίζεται από την κατεύθυνση της κλίσης στο χώρο των βαρών για την εύρεση ολικού ελαχίστου. Τα βάρη του ελαχίστου χρησιμοποιούνται για την επόμενη επανάληψη. Η παράμετρος μάθησης, που καθορίζεται από την μεταβολή των βαρών σε μία εποχή πάνω στη γραμμή του αναμενόμενου ελαχίστου, πολλαπλασιάζεται με τον ρυθμό μάθησης όσο το σφάλμα εκπαίδευσης του νευρωνικού με τα μεταβλημένα βάρη είναι μικρότερο του αμετάβλητου νευρωνικού. Αν το σφάλμα μάθησης του μεταβλημένου νευρωνικού ήταν ήδη μεγαλύτερο για την αρχική παράμετρο μάθησης, τότε διαιρείται με τον ρυθμό μάθησης μέχρι το σφάλμα μάθησης μειωθεί. Η ακρίβεια του επαναλαμβανόμενου και προσεγγιστικού υπολογισμού του $H^{-1(κ)}$ μειώνεται με τον αριθμό των επαναλήψεων. Γι' αυτό το λόγο ο πίνακας αρχικοποιείται στον μοναδιαίο πίνακα κάθε ResetStep βήματα.

*Η κλίση της σύναψης υπολογίζεται με τον ίδιο τρόπο όπως η μέθοδος BP (με αρχικοποίηση κλίσεων και βαρών στο μηδέν)

Το πλεονέκτημα της BFGS μεθόδου εν σύγκριση με την BG είναι ο μικρότερος αριθμός επαναλήψεων. Παρα ταύτα, επειδή ο χρόνος υπολογισμού για μία επανάληψη

είναι ανάλογος του τετραγώνου του αριθμού των συνάψεων, μεγαλύτερα δίκτυα επηρεάζονται αρνητικά σημαντικά.

4.4 Δένδρα αποφάσεων

Τα δέντρα αποφάσεων (decision trees) είναι δενδρικές δομές, οι οποίες χρησιμοποιούνται για να κατηγοριοποιούν δεδομένα, τα οποία ταξινομούν με βάση τις τιμές τους σε διάφορα χαρακτηριστικά (attributes) [Kotsiantis, 2007]. Ένα δέντρο αποφάσεων λαμβάνει ως είσοδο ένα αντικείμενο ή μια κατάσταση που περιγράφεται από ένα σύνολο χαρακτηριστικών και επιστρέφει μια απόφαση, την προβλεπόμενη τιμή εξόδου για την είσοδο [Stuart Russell, 2005]. Τα χαρακτηριστικά εισόδου μπορεί να είναι διακριτά ή συνεχή. Διακρίνονται δυο κατηγορίες δέντρων αποφάσεων, ανάλογα με το αν η τιμή εξόδου είναι συνεχής η διακριτή:

·Δέντρα ταξινόμησης, όπου η έξοδος έχει διακριτές τιμές.

·Δέντρα παλινδρόμησης, όπου η έξοδος έχει συνεχή τιμή.

Τα δέντρα αποφάσεων εξάγουν τις αποφάσεις τους εκτελώντας μια ακολουθία ελέγχων. Κάθε εσωτερικός κόμβος στο δέντρο αντιστοιχεί σε έναν έλεγχο τιμής για ένα από τα χαρακτηριστικά, ενώ οι διακλαδώσεις από κάθε κόμβο λαμβάνουν ως ετικέτα τις δυνατές τιμές του χαρακτηριστικού αυτού. Κάθε κόμβος-φύλλο στο δέντρο προσδιορίζει την τιμή που θα επιστραφεί ως έξοδος όταν η προσπέλαση φτάσει μέχρι το συγκεκριμένο φύλλο. Τα δείγματα κατατάσσονται σε κατηγορίες ξεκινώντας από τον κόμβο-ρίζα και καταλήγοντας σε κάποιο κόμβο-φύλλο [Stuart Russell, 2005; Kotsiantis, 2007].

4.4.1 Boosted Decision Trees (Ενισχυμένα δένδρα αποφάσεως)

Boosting

Η περισσότερο διαδεδομένη μέθοδος επιλογής δειγμάτων καθοδηγούμενης από μοντέλο (Model-guided Instance Selection) είναι η μέθοδος Boosting. Η μέθοδος Boosting , η οποία ονομάζεται και arcing (Adaptive Resampling and Combining), είναι μια γενική μέθοδος βελτίωσης της απόδοσης ενός ασθενούς συστήματος μάθησης (weaklearner). Τα ασθενή συστήματα μάθησης επιτυγχάνουν απόδοση καλύτερη από την τυχαία ταξινόμηση αλλά συνήθως δεν καταφέρουν να προσεγγίσουν σε πολύ μεγάλο βαθμό την αληθινή ταξινόμηση. Η μέθοδος λειτουργεί με επαναληπτική εκτέλεση ενός ασθενούς ταξινομητή σε διάφορα κατανεμημένα δεδομένα εκπαίδευσης. Οι ταξινομητές που παράγονται από τα ασθενή συστήματα μάθησης συνδυάζονται σε ένα σύνθετο ισχυρό ταξινομητή για την επίτευξη μεγαλύτερης ακρίβειας - LiorRokach,Ensemble-basedclassifiers,ArtifIntellRev(2010)33:1–39,SpringerScience&BusinessMedia

Η μέθοδος boosting ενισχύει τη σταθεροποίηση της απόκρισης των δέντρων αποφάσεων σε σχέση με τις διακυμάνσεις στο δείγμα εκπαίδευσης και είναι σε θέση να βελτιώσει σημαντικά την απόδοση. Εφαρμόστηκαν δύο μέθοδοι boosting, Adaptive Boosting και Gradient Boosting.

Adaptive Boost (Ada-Boosting)

Ο πιο δημοφιλής αλγόριθμός boosting ονομάζεται AdaBoost (adaptive boosting) [Schapire, 1997]. Σε προβλήματα κατηγοριοποίησης, γεγονότα τα οποία κατηγοριοποιούνται λανθασμένα κατά την εκπαίδευση ενός δένδρου απόφασης λαμβάνουν μεγαλύτερο βάρος κατά την απόφαση του επόμενου δένδρου. Ξεκινώντας με τα αρχικά βάρη των γεγονότων στην εκπαίδευση του δένδρου απόφασης, το επόμενο δένδρο εκπαιδεύεται χρησιμοποιώντας ένα αλλαγμένο δείγμα γεγονότων, όπου τα βάρη προηγούμενων λανθασμένα κατηγοριοποιημένων γεγονότων πολλαπλασιάζονται με ένα κοινό βάρος ενίσχυσης α. Το ενισχυμένο βάρος υπολογίζεται από το ρυθμό λανθασμένης κατηγοριοποίησης, err, του προηγούμενου δένδρου.

$$\alpha = \frac{1 - \mathrm{err}}{\mathrm{err}} \,. \qquad (0.33)$$

Εν συνεχεία, τα βάρη του συνολικού δείγματος γεγονότων κανονικοποιούνται έτσι ώστε το σύνολο του αθροίσματος των βαρών να παραμένει σταθερό.

Ορίζουμε το αποτέλεσμα κάθε ταξινομητή ως h(x) (x την πλειάδα των μεταβλητών εισαγωγής), με h(x) = +1 και -1 για σήμα και θόρυβο αντιστοίχως. Η ταξινόμηση του ενισχυμένου γεγονότος δίνεται τότε από

$$y_{\mathrm{Boost}}(\mathbf{x}) = \frac{1}{N_{\mathrm{collection}}} \cdot \sum_{i}^{N_{\mathrm{collection}}} \ln(\alpha_i) \cdot h_i(\mathbf{x}) \,, \qquad (0.34)$$

όπου το άθροισμα γίνεται στο σύνολο των ταξινομητών. Μικρές (μεγάλες) τιμές του yboost (x) σημαίνουν γεγονός σήματος (θορύβου). Η εξίσωση (4.37) είναι η εξίσωση του γενικού αλγορίθμου ενίσχυσης.

Ο αλγόριθμος AdaBoost φέρει καλύτερα αποτελέσματα σε αδύναμους ταξινομητές, πολύ μικρής διαχωριστικής ικανότητας, όπως μικρά δένδρα απόφασης με βάθος από 2-3 δένδρα. Τέτοια μικρά δένδρα είναι λιγότερο επιρρεπή στο overtraining εν συγκρίση με τα απλά δένδρα απόφασης και σε γενικές γραμμές αποδίδουν πολύ καλύτερα. Η απόδοση πολλές φορές ενισχύεται περαιτέρω εφαρμόζοντας "αργή εκπαίδευση" με μεγαλύτερο αριθμό βημάτων ενίσχυσης. Ο ρυθμός μάθησης του αλγορίθμου AdaBoost ελέγχεται από έναν παράγοντα β σε εκθετικό $\alpha \rightarrow \alpha^{\beta}$.

Για δένδρα παλινδρόμησης (regression trees) ο αλγόριθμος πρέπει να διαφοροποιηθεί. Η TMVA χρησιμοποιεί τον αλγόριθμο AdaBoost.R2 [Drucker, ICML 1997]. Η ιδέα που ακολουθείται είναι ίδια με αυτή του AdaBoost με τη διαφορά του επαναπροσδιορισμού του σφάλματος κάθε γεγονότος, λαμβάνοντας υπόψιν την απόκλιση της αποτίμησης της επιθυμητής τιμής από την πραγματική. Επιπλέον, αφού δεν υπάρχουν σωστά και λανθασμένα κατηγοριοποιημένα γεγονότα, όλα τα βάρη των γεγονότων πρέπει να με βάση τα σφάλματα τους, τα οποία -για το k γεγονός- δίνονται από τις συναρτήσεις:

$$Linear: \qquad L(k) = \frac{|y(k)-\hat{y}(k)|}{\max\limits_{\text{events } k'}(|y(k')-\hat{y}(k')|)}\ , \qquad (0.35)$$

$$Square: \qquad L(k) = \left[\frac{|y(k)-\hat{y}(k)|}{\max\limits_{\text{events } k'}(|y(k')-\hat{y}(k')|)}\right]^2\ , \qquad (0.36)$$

$$Exponential: \quad L(k) = 1 - \exp\left[-\frac{|y(k)-\hat{y}(k)|}{\max\limits_{\text{events } k'}(|y(k')-\hat{y}(k')|)}\right]\ . \qquad (0.37)$$

Το μέσο σφάλμα για έναν ταξινομητή για το συνολικό δείγμα μάθησης, $\langle L\rangle^{(i)} = \sum_{\text{events } k'} (k')L^{(i)}(k')$, μπορεί να θεωρηθεί ανάλογο μέρους του σφάλματος στην κατηγοριοποίηση. Έχοντας το <L>, μπορεί να υπολογιστεί η ποσότητα $\beta_{(i)} = \langle L\rangle^{(i)}/ (1 - \langle L\rangle^{(i)})$, που χρησιμοποιείται στην ενίσχυση των γεγονότων και για το συνδυασμό μεθόδων παλινδρόμησης με χρήση ενίσχυσης. Το ενισχυμένο βάρος, $w^{(i+1)}(k)$, για ένα γεγονός k και βήμα ενίσχυσης i+1

$$w^{(i+1)}(k) = w^{(i)}(k) \cdot \beta_{(i)}^{1-L^{(i)}(k)}\ . \qquad (0.38)$$

Το άθροισμα των βαρών των γεγονότων κανονικοποιείται ξανά για να δημιουργήσει ξανά το ολικό αριθμό των γεγονότων. Στο τελευταίο βήμα της παλινδρόμησης, y_{Boost} χρησιμοποιείται ο μέσος όρος των βαρών, $\tilde{y}_{(i)}$, όπου (i) είναι ο ελάχιστος αριθμός που ικανοποιεί την εξίσωση

$$\sum_{\substack{t\in \text{sorted collection} \\ t\leq i}} \ln \frac{1}{\beta_{(t)}} \geq \frac{1}{2} \sum_{t}^{N_{\text{collection}}} \ln \frac{1}{\beta_{(t)}} \qquad (0.39)$$

Gradient Boost

Η ιδέα της εκτίμησης της συνάρτησης με ενίσχυση μπορεί να γίνει κατανοητή εξετάζοντας μια απλή προσέγγιση προσθετικού αναπτύγματος. Η υπό εξέταση συνάρτηση F (x) θεωρείται ότι είναι ένα άθροισμα συναρτήσεων βάσης με παραμέτρους και βάρη f (x; am)(«αδύναμοι μαθητές»). Έτσι, κάθε συνάρτηση βάσης στο ανάπτυγμα αντιστοιχεί σε ένα δένδρο απόφασης

$$F(\mathbf{x}; P) = \sum_{m=0}^{M} \beta_m f(x; a_m); \qquad P \in \{\beta_m; a_m\}_0^M \,. \qquad (0.40)$$

Η διαδικασία ενίσχυσης χρησιμοποιείται για την προσαρμογή των παραμέτρων, έτσι ώστε η απόκλιση μεταξύ του μοντέλου απόκρισης F (x) και της πραγματικής τιμής που λαμβάνεται από το δείγμα εκπαίδευσης να ελαχιστοποιείται. Η απόκλιση μετριέται από τη συνάρτηση σφάλματος L (F, y), με δημοφιλή επιλογή την τετραγωνική ρίζα της συνάρτησης σφάλματος L (F, y) = (F (x) −y)^2. Μπορεί να αποδειχθεί ότι η συνάρτησης σφάλματος καθορίζει πλήρως τη διαδικασία ενίσχυσης.

Η πιο δημοφιλής μέθοδος ενίσχυσης, το AdaBoost, βασίζεται σε εκθετική απώλεια, L (F, y) = e $^{-F(x)}$. Η εκθετική απώλεια έχει το μειονέκτημα ότι στερείται αντοχής (robustness) παρουσία ακραίων ή λανθασμένα κατηγοριοποιημένων σημείων. Επομένως, η απόδοση του AdaBoost αναμένεται να μειωθεί σε θορυβώδη δεδομένα. Για να ξεπεραστεί αυτή η αδυναμία χρησιμοποιείται ο αλγόριθμος Gradient Boost, όπου διατηρείται η καλή απόδοση του AdaBoost με εφαρμογή πιθανόν πιο στιβαρών συναρτήσεων σφάλματος. Η τρέχουσα υλοποίηση του TMVA του GradientBoost χρησιμοποιεί την απώλεια δυαδικής λογαριθμικής πιθανότητας για κατηγοριοποίηση

$$L(F, y) = \ln\left(1 + e^{-2F(\mathbf{x})y}\right) , \qquad (0.41)$$

Καθώς ο αλγόριθμος ενίσχυσης που αντιστοιχεί σε αυτή τη συνάρτηση σφάλματος δεν μπορεί να επιτευχθεί με έναν απλό τρόπο, πρέπει να γίνει εφαρμογή της μεθόδου μείωσης της κλίσης για την ελαχιστοποίηση. Αυτό γίνεται με τον υπολογισμό της τρέχουσας κλίσης της συνάρτησης σφάλματος και στη συνέχεια ανάπτυξη ενός δέντρου παλινδρόμησης του οποίου τιμές των φύλλων προσαρμόζονται ώστε να ταιριάζουν με τη μέση τιμή της κλίσης σε κάθε περιοχή που ορίζεται από τη δομή του δέντρου. Η επανάληψη αυτής της διαδικασίας αποδίδει το επιθυμητό σύνολο δέντρων αποφάσεων που ελαχιστοποιούν τη συνάρτηση σφάλματος. Η Gradient Boost μπορεί να προσαρμοστεί σε οποιαδήποτε συνάρτηση σφάλματος εφόσον είναι εφικτός ο υπολογισμός της κλίσης.

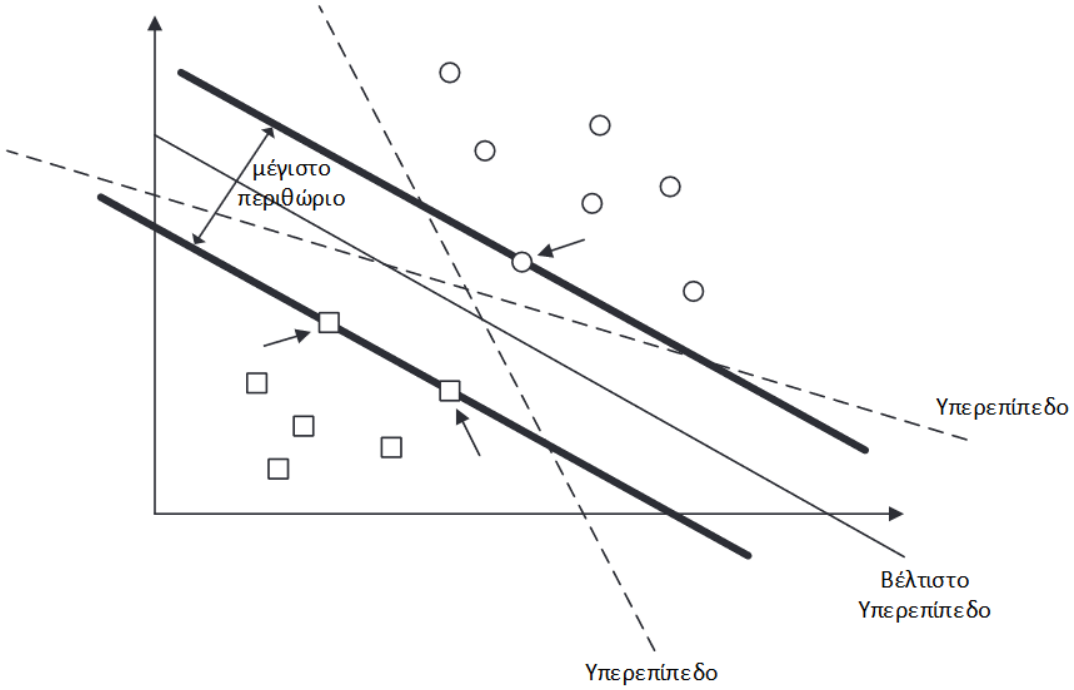
Η μέθοδος Gradient Boost, όπως και η AdaBoost λειτουργεί καλύτερα σε αδύναμους ταξινομητές, δηλαδή μικρά ατομικά δέντρα αποφάσεως με βάθος συχνά μόλις 2 έως 4. Τα μικρά δένδρα είναι λιγότερο επιρρεπή στο φαινόμενο της υπερεκπαίδευσης (overtraining). Η στιβαρότητα (robustness) μπορεί να ενισχυθεί μειώνοντας τον ρυθμό μάθησης του αλγορίθμου μέσω του Shrinkage, ο οποίος ελέγχει το βάρος των μεμονωμένων δέντρων. Μια μικρή συρρίκνωση (0,1-0,3) απαιτεί περισσότερα δένδρα, αλλά μπορεί να βελτιώσει σημαντικά την ακρίβεια της πρόβλεψης σε δύσκολες ρυθμίσεις.

4.5 Support Vector Machine (SVM)

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines - SVMs) είναι από τις νεότερες τεχνικές μάθησης υπό επίβλεψη. Η μεθοδολογία αυτή προτάθηκε από τον Vapnik (Boseret.al.1992,Cortesetal.1995,Vapnik1995,1998) [Haykin S. , 1999] και μπορεί να χρησιμοποιηθεί σε προβλήματα ταξινόμησης και μη γραμμικής παλινδρόμησης. Η βασική ιδέα των μηχανών διανυσμάτων υποστήριξης είναι η κατασκευή ενός υπερ-επιπέδου, το οποίο θα αποτελέσει την επιφάνεια απόφασης, με τέτοιο τρόπο ώστε να μεγιστοποιείται η απόσταση που διαχωρίζει τα πλησιέστερα δείγματα που ανήκουν σε διαφορετικές κλάσεις [Haykin S. , 1999].

Ο στόχος της ταξινόμησης με διανύσματα υποστήριξης είναι η σχεδίαση ενός υπολογιστικά αποδοτικού τρόπου για την εκμάθηση κατάλληλων διαχωριστικών υπερεπιπέδων, τα οποία βρίσκονται σε έναν υψηλών διαστάσεων διανυσματικό χώρο. Τα κατάλληλα υπερ-επίπεδα είναι εκείνα που βελτιστοποιούν τα φράγματα στο σφάλμα γενίκευσης. Η γενίκευση (generalization) είναι η διαδικασία ταξινόμησης νέων δεδομένων, τα οποία δεν υπάρχουν στο σύνολο δεδομένων εκπαίδευσης. Για να είναι υπολογιστικά αποδοτικοί, οι αλγόριθμοι εκπαίδευσης που θα χρησιμοποιηθούν

θα πρέπει να μπορούν να χειριστούν σύνολα δεδομένων της τάξης των 100000 δειγμάτων [Nello Cristianini, 2000]. Η βασική ιδέα της μεθοδολογίας αυτής μπορεί να δειχθεί, αν θεωρήσουμε δυο κλάσεις και δυο χαρακτηριστικά για ένα σύνολο δεδομένων, όπως φαίνεται στην Εικόνα 4.7 [Kotsiantis, 2007]. Οι δυο ομάδες που σχηματίζουν τα δεδομένα είναι εμφανείς. Σε αυτή την περίπτωση, οι μηχανές διανυσμάτων υποστήριξης θα υπολογίσουν την εξίσωση της ευθείας που θα διαχωρίζει τις δυο ομάδες με τη μέγιστη δυνατή απόσταση, το οποίο σημαίνει ότι η απόσταση μεταξύ της ευθείας και των πλησιέστερων δειγμάτων (σημείων) μεγιστοποιείται. Αν το πλήθος των χαρακτηριστικών ήταν μεγαλύτερο, τότε οι διαστάσεις του χώρου θα ήταν περισσότερες και επομένως προκύπτει ένα υπερεπίπεδο αντί της ευθείας. Το υπερεπίπεδο αυτό ορίζεται από ένα υποσύνολο των σημείων των δυο κλάσεων, τα οποία ονομάζονται Διανύσματα Υποστήριξης (SupportVectors). Τα σημεία αυτά έχουν σημειωθεί με βέλη στο συγκεκριμένο παράδειγμα (Εικόνα 4.7) [Kotsiantis, 2007] .



Εικόνα 0.7 Μέγιστο Περιθώριο και Βέλτιστο υπερεπίπεδο

Οι Μηχανές Διανυσμάτων Υποστήριξης μπορούν να χρησιμοποιηθούν για μη γραμμική ταξινόμηση, χρησιμοποιώντας μη γραμμικές συναρτήσεις πυρήνα (non-linear kernel). Η μη γραμμική συνάρτηση πυρήνα είναι μια μαθηματική συνάρτηση η οποία μετασχηματίζει τα δεδομένα από τον γραμμικό διανυσματικό χώρο των χαρακτηριστικών (linear feature space) σε μη γραμμικό διανυσματικό χώρο.

Διαφορετικές συναρτήσεις πυρήνα θα πρέπει να εφαρμοστούν σε κάθε σύνολο δεδομένων προβλήματος, ώστε να βρεθεί πια συνάρτηση αποδίδει καλύτερα για το εκάστοτε πρόβλημα [Kotsiantis, 2007; Choi].

Το σφάλμα λανθασμένης ταξινόμησης στις Μηχανές Διανυσμάτων Υποστήριξης μπορεί να ελεγχθεί από την παράμετρο C. Αν η παράμετρος αυτή πάρει μεγάλες τιμές, η λανθασμένη ταξινόμηση μπορεί να περιοριστεί και αν οι τιμές της είναι μικρές δεδομένα εκπαίδευσης που απέχουν πολύ από τα δεδομένα που έχουν συλλεχθεί επιτρέπεται στο μοντέλο να τα ταξινομήσει εσφαλμένα. Συνεπώς, θέτοντας μια κατάλληλη τιμή στην παράμετρο C περιορίζονται οι ακραίες τιμές των δεδομένων εκπαίδευσης (outliers). Επίσης, με αυτό τον τρόπο μπορεί να περιοριστεί το πρόβλημα της υπερπροσαρμογής.

Γραμμικώς Διαχωρίσιμα Δεδομένα

Το ποιο απλό μοντέλο των Μηχανών Διανυσμάτων Υποστήριξης είναι ο Ταξινομητής Μεγίστου Περιθωρίου (Maximal Margin Classifier), το οποίο μπορεί να λειτουργήσει μόνο σε δεδομένα που είναι γραμμικώς διαχωρίσιμα, οπότε και δεν μπορεί να εφαρμοστεί σε πραγματικές εφαρμογές. Με βάση όμως το μοντέλο αυτό συνθέτονται περισσότερο πολύπλοκες Μηχανές Διανυσμάτων Υποστήριξης. Θεωρούμε πρόβλημα ταξινόμησης με δυο κλάσεις, P και N. Αν τα δεδομένα του προβλήματος είναι διαχωρίσιμα, τότε υπάρχει (w,b) τέτοιο ώστε [Kotsiantis, 2007; Nello Cristianini, 2000][35],[96]:

$$\boldsymbol{w}^T \boldsymbol{x_i}^+ + b \geq 1, \gamma\iota\alpha\ \kappa\acute{\alpha}\theta\varepsilon\ \boldsymbol{x_i}^+ \epsilon P \qquad (0.42)$$

$$\boldsymbol{w}^T \boldsymbol{x_i}^- + b \leq -1, \gamma\iota\alpha\ \kappa\acute{\alpha}\theta\varepsilon\ \boldsymbol{x_i}^- \epsilon N \qquad (0.43)$$

Το διάνυσμα των βαρών w είναι κάθετο στο υπερεπίπεδο. Είναι εύκολο να δειχθεί ότι όταν είναι δυνατό να διαχωριστούν γραμμικά δυο κλάσεις, μπορεί να βρεθεί ένα βέλτιστο υπερεπίπεδο το οποίο διαχωρίζει τις κλάσεις αυτές. Το υπερεπίπεδο προκύπτει από την ελαχιστοποίηση του τετραγώνου της νόρμας του. Το συγκεκριμένο πρόβλημα ελαχιστοποίησης μπορεί να εκφραστεί ως ένα πρόβλημα Κυρτού Τετραγωνικού Προγραμματισμού (Convex Quadratic Programming QP) [Kotsiantis, 2007; Nello Cristianini, 2000].

$$\min_{\boldsymbol{w},b} \Phi(\boldsymbol{w}) = \frac{\mathbf{1}}{\mathbf{2}} \|\boldsymbol{w}\|^2 \qquad (0.44)$$

$$\mu\varepsilon\ \tau o\nu\ \pi\varepsilon\rho\iota o\rho\iota\sigma\mu\acute{o}:\ y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) \geq 1, i = 1, \ldots, l \qquad (0.45)$$

Ορίζουμε το συναρτησιακό περιθώριο (functional margin) του δείγματος (x_i, y_i) σε σχέση με ένα υπερεπίπεδο (**w**,b) ως την ποσότητα:

$$\gamma_i = y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) \qquad (0.46)$$

καθώς επίσης και το γεωμετρικό περιθώριο (geometrical margin) ως το συναρτησιακό περιθώριο της κανονικοποιημένης γραμμικής συνάρτησης

$$\rho_i = y_i\left(\frac{\boldsymbol{w}^T}{\|\boldsymbol{w}\|}\boldsymbol{x_i} + \frac{b}{\|\boldsymbol{w}\|}\right) \qquad (0.47)$$

Αν γ_i>0 τότε το δείγμα (x_i,y_i) έχει ταξινομηθεί σωστά. Η κατανομή του συναρτησιακού περιθωρίου του υπερεπιπέδου (**w**,b) ως προς ένα σύνολο δεδομένων εκπαίδευσης S είναι η κατανομή των περιθωρίων στο S. Σε κάποιες περιπτώσεις γίνεται αναφορά στην ελάχιστη κατανομή περιθωρίου του υπερεπιπέδου (**w**,b) ως προς ένα σύνολο δεδομένων εκπαίδευσης S. Αν αντικατασταθεί το συναρτησιακό περιθώριο με το γεωμετρικό περιθώριο παίρνουμε την γραμμική κανικοποιημένη συνάρτηση, η οποία μετράει τις Ευκλείδειες αποστάσεις των σημείων από το όριο απόφασης στο διανυσματικό χώρο εισόδου. **Το περιθώριο ενός συνόλου εκπαίδευση S** ορίζεται ως το μέγιστο γεωμετρικό περιθώριο, από τα γεωμετρικά περιθώρια όλων των υπερεπιπέδων. Το γεωμετρικό περιθώριο ενός γραμμικά διαχωρίσιμου συνόλου δεδομένων θα είναι θετικό [Kotsiantis, 2007; Nello Cristianini, 2000; Choi].

Το γεωμετρικό περιθώριο του υπερεπιπέδου που προκύπτει στο πρόβλημα ταξινόμησης που έχουμε θεωρήσει θα είναι [Nello Cristianini, 2000]

$$\rho = \frac{1}{2}\left(\frac{\boldsymbol{w}^T}{\|\boldsymbol{w}\|}\boldsymbol{x}^+ - \frac{\boldsymbol{w}^T}{\|\boldsymbol{w}\|}\boldsymbol{x}^-\right) = \frac{1}{2}\frac{1}{\|\boldsymbol{w}\|}\left(\boldsymbol{w}^T\boldsymbol{x}^+ - \boldsymbol{w}^T\boldsymbol{x}^-\right) = \frac{1}{\|\boldsymbol{w}\|} \quad (0.48)$$

Στην περίπτωση των γραμμικά διαχωρίσιμων δεδομένων, κατά την εύρεση του βέλτιστου επιπέδου που διαχωρίζει τις κλάσεις τα σημεία που βρίσκονται πάνω στο περιθώριο του υπερεπιπέδου ονομάζονται Διανύσματα Υποστήριξης (Support Vectors). Η λύση του προβλήματος βελτιστοποίησης εκφράζεται ως γραμμικός συνδυασμός των Διανυσμάτων Υποστήριξης. Όλα τα υπόλοιπα σημεία αγνοούνται, καθώς δεν συνεισφέρουν στον υπολογισμό της λύσης. Επομένως, η πολυπλοκότητα του μοντέλου μιας Μηχανής Διανυσμάτων Υποστήριξης δεν επηρεάζεται από το πλήθος των χαρακτηριστικών του συνόλου δεδομένων εκπαίδευσης (το πλήθος των Διανυσμάτων Υποστήριξης που προκύπτουν συνήθως είναι μικρό). Συνεπώς, η μέθοδος αυτή μπορεί να αντιμετωπίσει προβλήματα στα οποία το πλήθος των χαρακτηριστικών είναι μεγάλο σε σχέση με το πλήθος των δειγμάτων [Kotsiantis, 2007].

Δυϊκό Πρόβλημα Βελτιστοποίησης

Ο μετασχηματισμός του προβλήματος κυρτής βελτιστοποίησης που ορίσαμε παραπάνω, χρησιμοποιώντας τους πολλαπλασιαστές Lagrange οδηγεί σε μια εναλλακτική δυϊκή περιγραφή η οποία μπορεί να επιλυθεί ευκολότερα από το αρχικό πρόβλημα, αφού ο άμεσος χειρισμός των περιορισμών με ανισότητες είναι ιδιαίτερα δύσκολος. Οι πολλαπλασιαστές Lagrange ονομάζονται δυϊκές μεταβλητές. Η δυϊκή μέθοδος βασίζεται στην ιδέα ότι οι πολλαπλασιαστές Lagrange είναι οι βασικοί άγνωστοι του προβλήματος που θα πρέπει να υπολογιστούν.

Υπάρχει η δυνατότητα μετατροπής του αρχικού προβλήματος στο δυϊκό του με το να τεθούν ίσες με το μηδέν όλες οι παράγωγοι της συνάρτησης Lagrange ως προς τις αρχικές μεταβλητές, οπότε και αφαιρείται η επίδρασή τους. Αυτό αντιστοιχεί στον υπολογισμό της συνάρτησης:

$$\theta(\alpha, b) = \inf_{w \in \Omega} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) \quad (0.49)$$

Η συνάρτηση που προκύπτει περιέχει μόνο δυϊκές μεταβλητές και θα πρέπει να μεγιστοποιηθεί, με απλούστερους περιορισμούς [Nello Cristianini, 2000].

Η συνάρτηση Lagrange για το αρχικό πρόβλημα θα είναι η εξής:

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} \alpha_i [\, y_i (\, \boldsymbol{w}^T \boldsymbol{x_i} + b) - 1] \qquad (0.50)$$

όπου α_i είναι οι πολλαπλασιαστές Lagrange.

Το αντίστοιχο δυαδικό πρόβλημα προκύπτει παραγωγίζοντας ως προς **w** και b και θεωρώντας στατικότητα, με τις παραγώγους αυτές να ισούται με μηδέν:

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{l}} y_i \alpha_i \boldsymbol{x_i} = 0 \qquad (0.51)$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i = 0 \qquad (0.52)$$

Με αντικατάσταση των εξισώσεων (4.54) και (4.55) στην εξίσωση (4.53) προκύπτει:

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i \, y_j \, \alpha_i \alpha_j \boldsymbol{x_i}^T \boldsymbol{x_j} \qquad (0.53)$$

οπότε το αρχικό πρόβλημα μετασχηματίζεται στο δυϊκό του, το οποίο είναι το εξής [Nello Cristianini, 2000; Choi]:

$$\max_{a} W(a) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i \, y_j \, \alpha_i \alpha_j \boldsymbol{x_i}^T \boldsymbol{x_j} \qquad (0.54)$$

$$\text{με τον περιορισμό: } \sum_{i=1}^{l} y_i \alpha_i = 0 \,, \alpha_i \geq 0, i = 1, \ldots, l \qquad (0.55)$$

Έστω ότι οι παράμετροι α* επιλύουν το **πρόβλημα τετραγωνικής βελτιστοποίησης** που περιγράφεται από τις εξισώσεις (4.57) και (4.58), δοθέντος ενός γραμμικά διαχωρίσιμου συνόλου δεδομένων εκπαίδευσης $D=\{(x_i,y_i)^i{}_{i=1}\}$. Τότε, το διάνυσμα βάρους w* [Nello Cristianini, 2000; Choi]:

$$w^* = \sum_{i=1}^{l} \alpha_i^* y_i \boldsymbol{x_i} \qquad (0.56)$$

δημιουργεί το υπερεπίπεδο μεγίστου περιθωρίου, με γεωμετρικό περιθώριο $\rho = 1/\|w\|$.

Η τιμή της παραμέτρου b δεν εμφανίζεται στο δυϊκό πρόβλημα, οπότε το b* θα πρέπει να υπολογιστεί χρησιμοποιώντας τους αρχικούς περιορισμούς [Nello Cristianini, 2000]:

$$b^* = -\frac{\max\limits_{y_i=-1}\left(\boldsymbol{w}^{*T}\boldsymbol{x_i}\right) + \min\limits_{y_i=1}\left(\boldsymbol{w}^{*T}\boldsymbol{x_i}\right)}{2} \qquad (0.57)$$

Οι συμπληρωματικές συνθήκες Karush-Kuhn-Tycker από την θεωρία βελτιστοποίησης εφαρμόζονται στο πρόβλημα αυτό και δίνουν χρήσιμες πληροφορίες για την μορφή της λύσης. Για τις βέλτιστες λύσεις α_i, (**w***,b) ισχύει [Nello Cristianini, 2000]:

$$\alpha_i^* \left[y_i \left(\boldsymbol{w}^{*T}\boldsymbol{x_i} + b^* \right) - 1 \right] = 0, i = 1, \ldots, l \qquad (0.58)$$

Από την εξίσωση (4.61) προκύπτουν οι εξής παρατηρήσεις:

•Αν , οπότε το $y_i (w^{*T} x_i + b^*) \neq 1$ δεν είναι Διάνυσμα Υποστήριξης, τότε $\alpha_i = 0$.

•Αν τότε $\alpha_i \neq 0$ τότε $y_i (w^{*T} x_i + b^*) = 1$οπότε **το x_i είναι Διάνυσμα Υποστήριξης.**

Μόνο τα Διανύσματα Υποστήριξης συμμετέχουν στον υπολογισμό του διανύσματος βάρους w* και κατά συνέπεια στο βέλτιστο υπερεπίπεδο [Nello Cristianini, 2000].

Υλοποίηση TMVA

Η υλοποίηση TMVA του Support Vector Machine ακολουθεί στενά την περιγραφή που δίνεται στη βιβλιογραφία. Χρησιμοποιεί μια διαδοχική ελάχιστη βελτιστοποίηση (Sequential Minimum Optimization/SMO) (Platt, 1999) για την επίλυση του τετραγωνικού προβλήματος. Επιτάχυνση της ελαχιστοποίησης επιτυγχάνεται διαιρώντας ένα σύνολο διανυσμάτων σε μικρότερα υποσύνολα (S. Keerthi, 1999). Ο αριθμός των υποσύνολων εκπαίδευσης ελέγχεται από το option NSubSets. Η μέθοδος SMO επιλέγει υποσύνολα δύο διανυσμάτων (Burges, 1998). Τα ζεύγη διανυσμάτων επιλέγονται, χρησιμοποιώντας ευρετικούς κανόνες, για να επιτευχθεί η καλύτερη βελτιστοποίηση (ελαχιστοποίηση) ανά βήμα. Η διαδικασία ελαχιστοποίησης επαναλαμβάνεται αναδρομικά έως ότου βρεθούν τα ελάχιστα. Ο αλγόριθμος SMO έχει αποδειχθεί ότι είναι πολύ ταχύτερος από άλλες μεθόδους και έχει γίνει η πιο κοινή μέθοδος ελαχιστοποίησης που χρησιμοποιείται σε εφαρμογές SVM. Η ακρίβεια της ελαχιστοποίησης ελέγχεται από την παράμετρο ανοχής Tol (βλ. Πίνακα 1). Ο χρόνος εκπαίδευσης SVM μπορεί να μειωθεί αυξάνοντας την ανοχή. Τα περισσότερα προβλήματα ταξινόμησης θα πρέπει να λυθούν χωρίς 1000 επαναλήψεις εκπαίδευσης. Η διακοπή του αλγορίθμου SVM χρησιμοποιώντας την επιλογή MaxIter είναι χρήσιμη κατά τη βελτιστοποίηση των παραμέτρων εκπαίδευσης SVM. Το MaxIter μπορεί να χρησιμοποιηθεί για την τελική εκπαίδευση του ταξινομητή.

Πίνακας 1 Επιλογές διαμόρφωσης για τη μέθοδο SVM του MVA. Αν οι κατηγορίες έχουν προκαθοριστεί, η προεπιλεγμένη κατηγορία επισημαίνεται με το σύμβολο 'α'

Option	Array	Default	Predefined Values	Description
`Gamma`	–	`1`	–	RBF kernel parameter: Gamma (size of the Kernel)
`C`	–	`1`	–	Cost parameter
`Tol`	–	`0.01`	–	Tolerance parameter
`MaxIter`	–	`1000`	–	Maximum number of training loops

5. Κεφάλαιο 4ο: Ανάλυση

5.1 Απεικόνιση δεδομένων

5.1.1 Ιστογράμματα πυκνότητας πιθανότητας μεταβλητών

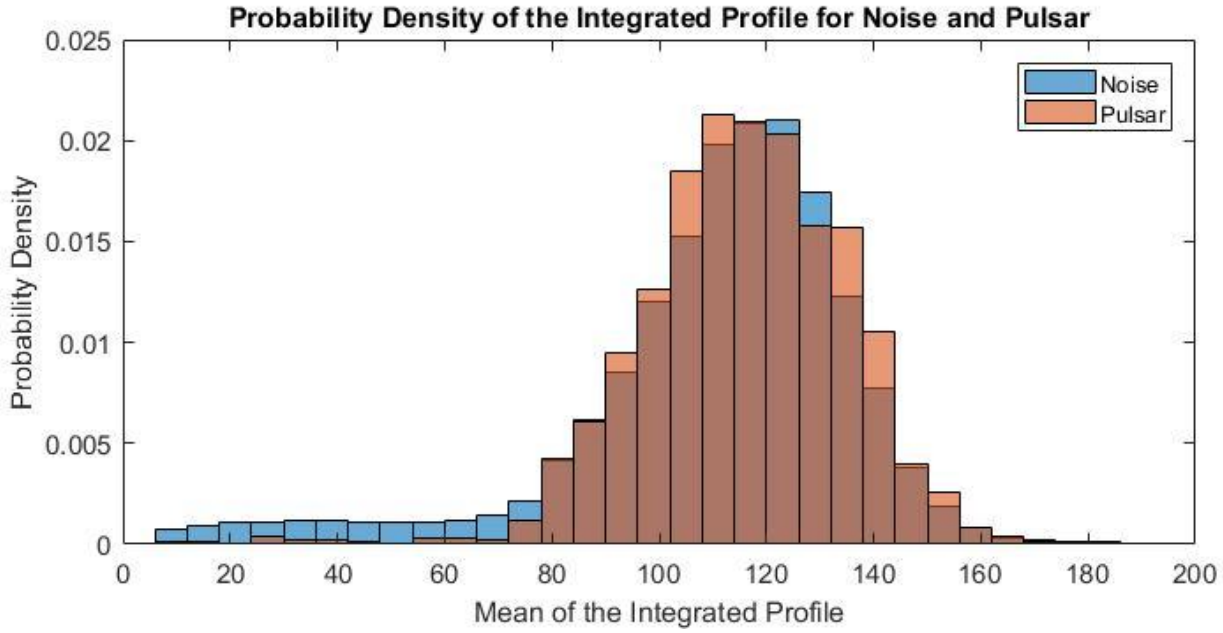
Συνάρτηση πυκνότητας πιθανότητας

Πυκνότητα πιθανότητας είναι η σχέση μεταξύ παρατηρήσεων και των σχετικών τους πιθανοτήτων. Μερικά αποτελέσματα τυχαίων μεταβλητών έχουν χαμηλή πυκνότητα πιθανότητας και μερικά άλλα έχουν υψηλή. Το ολικό σχήμα της πυκνότητας πιθανότητας ονομάζεται συνάρτηση πιθανότητας. Η πιθανότητα ενός αποτελέσματος υπολογίζεται με χρήση της συνάρτησης πυκνότητας πιθανότητας. Η γνώση της συνάρτησης πυκνότητας πιθανότητας είναι χρήσιμη για διάφορους λόγους. Πρώτον, γνωρίζοντας με ποια πιθανότητα εμφανίζονται οι παρατηρήσεις μπορούμε να χαρακτηρίσουμε αυτές με χαμηλή πιθανότητα ως ανώμαλα σημεία ή outliers. Δεύτερον, μπορούμε να επιλέξουμε κατάλληλες μεθόδους μάθησης που απαιτούν τα εισαγόμενα δεδομένα να έχουν συγκεκριμένες κατανομές πιθανότητας.

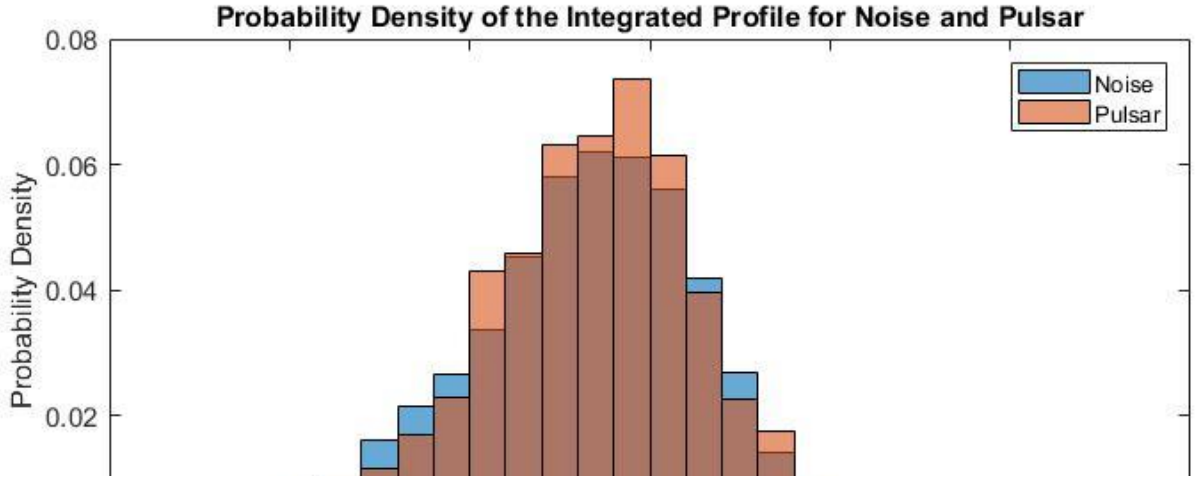
Εκτίμηση πυκνότητας πιθανότητας

Είναι απίθανο να γνωρίζουμε τη συνάρτηση πυκνότητας πιθανότητας εκ των προτέρων και γι' αυτό το λόγω πρέπει να γίνει μία εκτίμηση της. Για την εκτίμηση της συνάρτησης πυκνότητας πιθανότητας αρχικώς δημιουργούμε ιστογράμματα για την κάθε μεταβλητή. Για τη δημιουργία του ιστογράμματος, ομαδοποιούμε τις παρατηρήσεις σε bins μετρώντας τον αριθμό των γεγονότων του κάθε bin. Εν συνεχεία, οι συχνότητες των παρατηρήσεων του κάθε bin σχεδιάζονται σε ένα γράφημα στηλών με τα bins στον άξονα x και τη συχνότητα στον άξονα y. Παρακάτω παραθέτονται τα κανονικοποιημένα ιστογράμματα πυκνότητας πιθανότητας για κάθε μεταβλητή για τα δεδομένα που αντιστοιχούν σε θόρυβο (μπλε) και πάλσαρ (κόκκινο). Από τις διαφορές των ιστογραμμάτων πυκνότητας πιθανότητας μπορεί να γίνει μία πρώτη διερεύνηση για το ποιες μεταβλητές βοηθούν περισσότερο στην κατηγοριοποίηση των δεδομένων. Μπορούμε να συμπεράνουμε ότι η μέση τιμή του ενοποιημένου προφίλ, η μέση τιμή, η κύρτωση και η ασυμμετρία της καμπύλης DM-SNR μπορεί να συνεισφέρουν περισσότερο στην κατηγοριοποίηση.

1) Μεταβλητές ενοποιημένου προφίλ

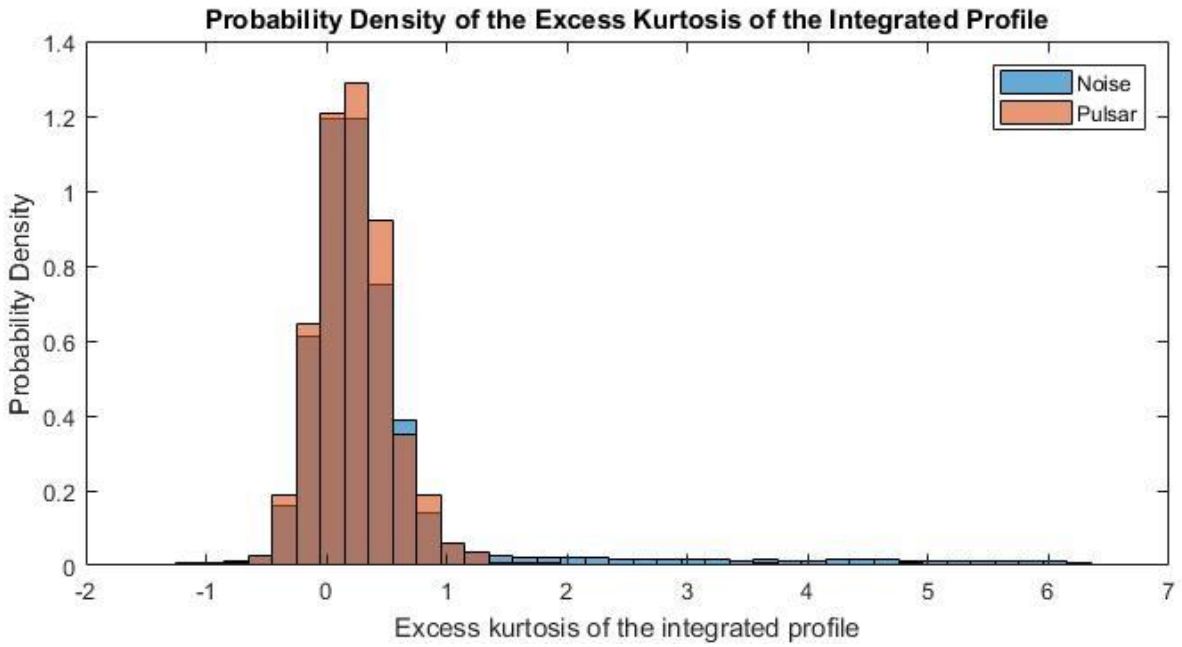


Εικόνα 0.1 Πυκνότητα πιθανότητας για τη μέση τιμή του ενοποιημένου προφίλ για σήμα και θόρυβο

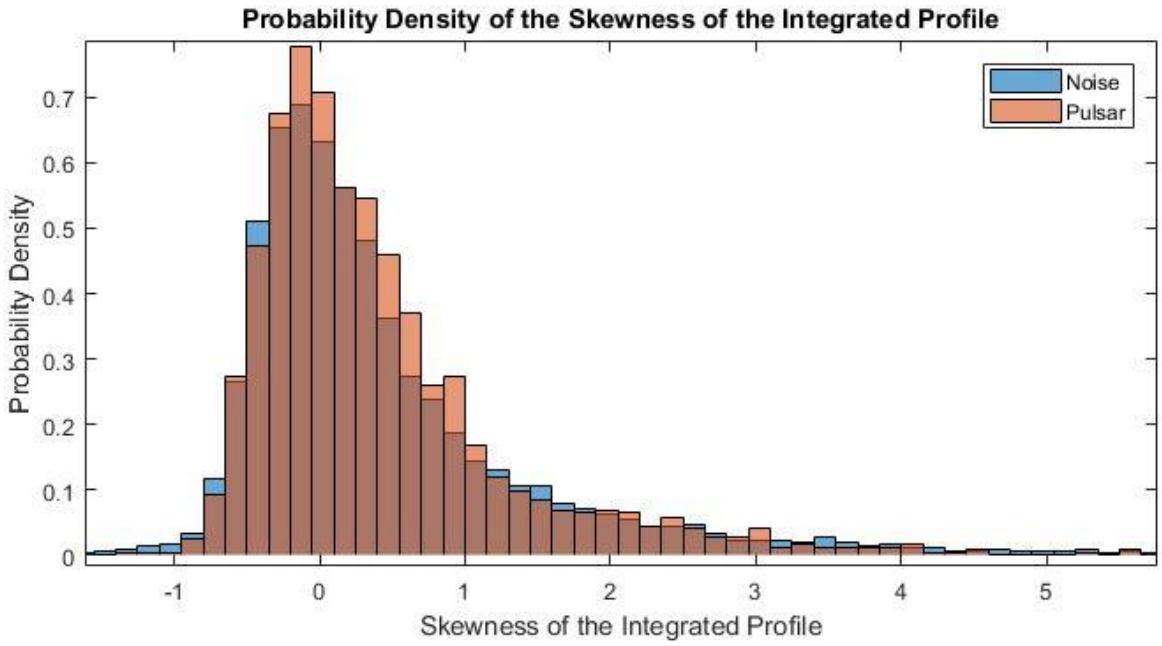


Εικόνα 0.2 Πυκνότητα πιθανότητας για την τυπική απόκλιση του ενοποιημένου προφίλ για σήμα και θόρυβο

Standard Deviation of the Integrated Profile



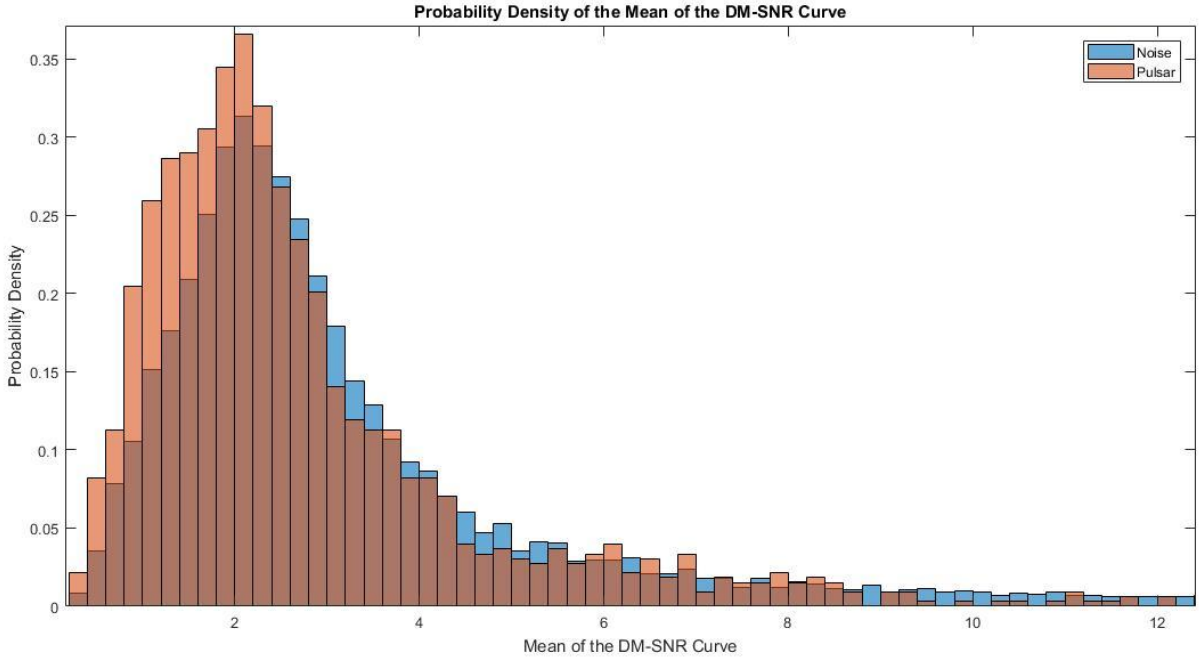
Εικόνα 0.3 Πυκνότητα πιθανότητας για την κύρτωση του ενοποιημένου προφίλ για σήμα και θόρυβο



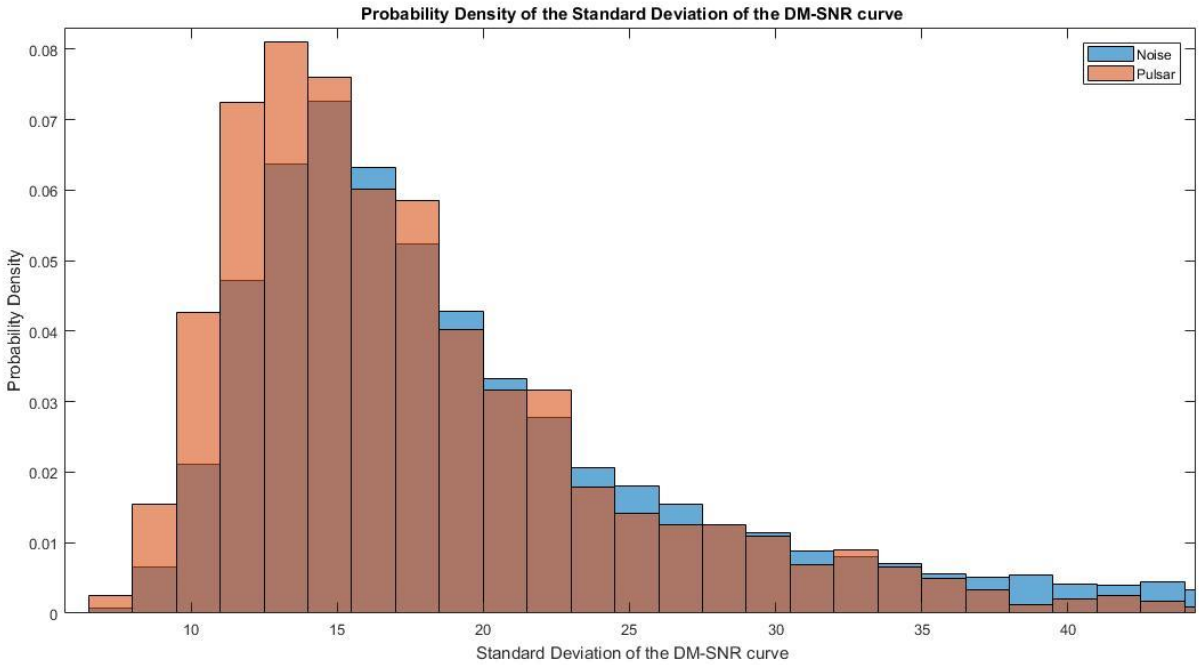
Εικόνα 0.4 Πυκνότητα πιθανότητας για την ασυμμετρία του ενοποιημένου προφίλ για σήμα και θόρυβο

2) Μεταβλητές καμπύλης DM-SNR

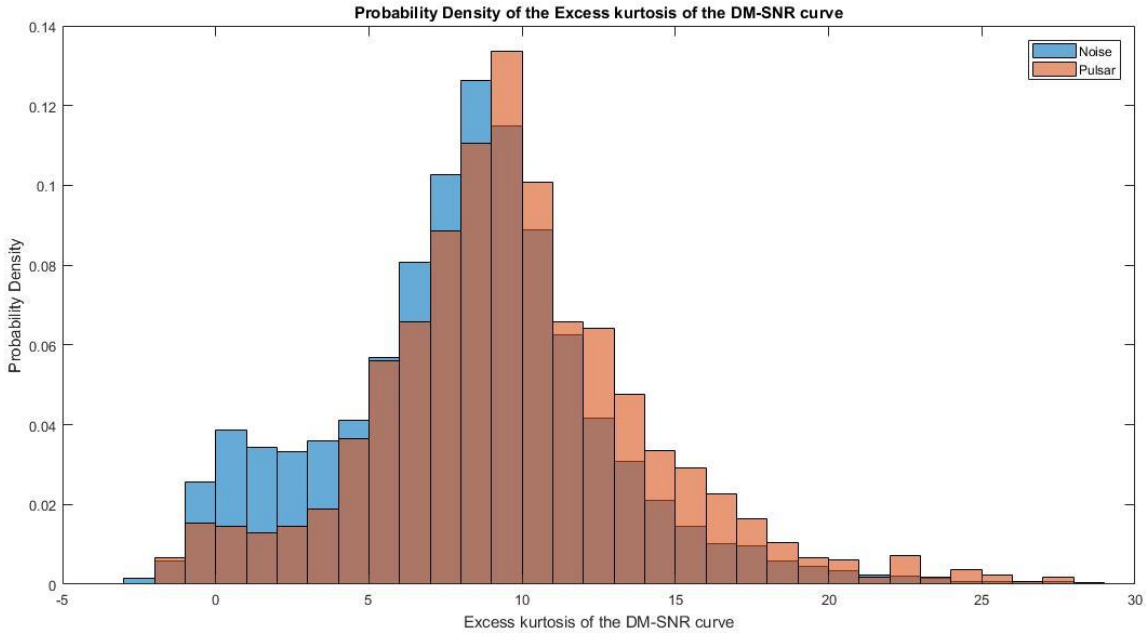
Μέση Τιμή



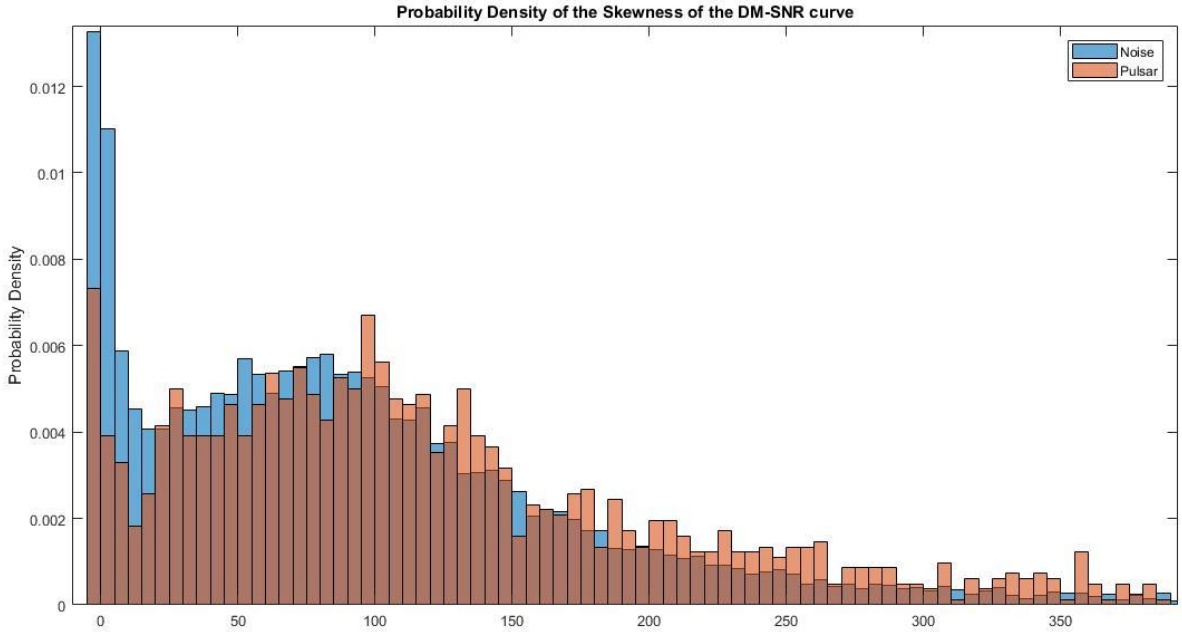
Εικόνα 0.5 Πυκνότητα πιθανότητας για τη μέση τιμή της καμπύλης DM-SNR για σήμα και θόρυβο



Εικόνα 0.6 Πυκνότητα πιθανότητας για την τυπική απόκλιση της καμπύλης DM-SNR για σήμα και θόρυβο

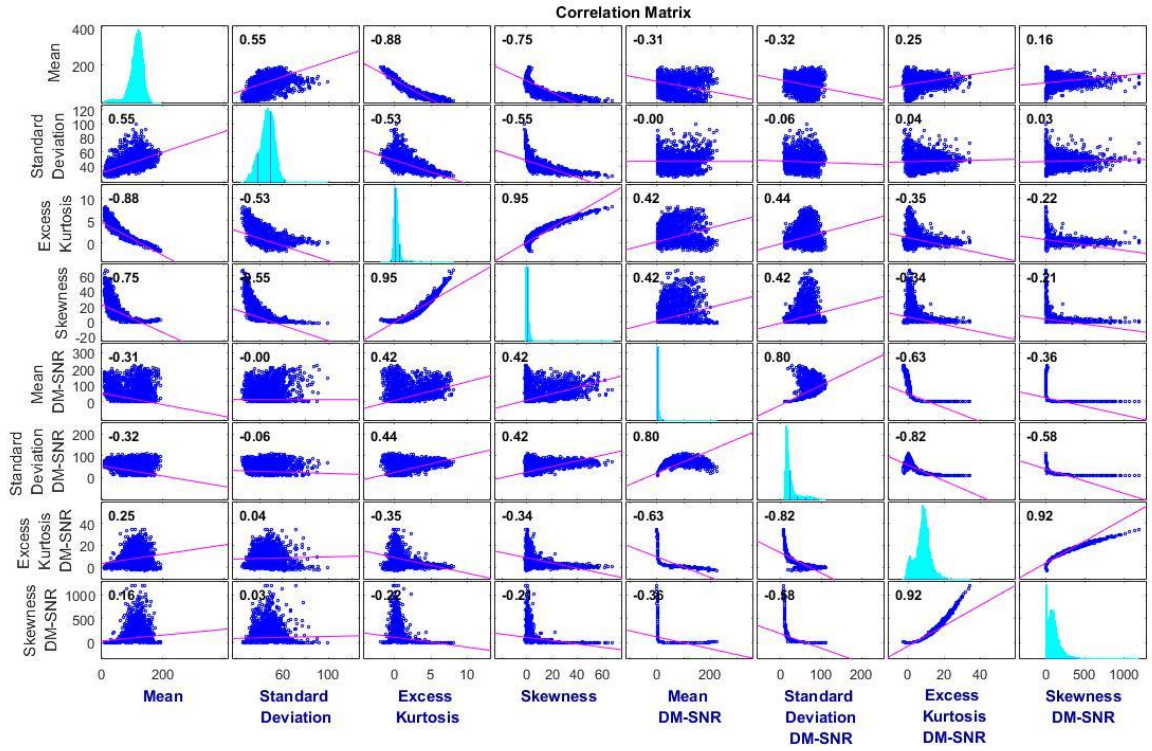


Εικόνα 0.7 Πυκνότητα πιθανότητας για την κύρτωση της καμπύλης DM-SNR για σήμα και θόρυβο

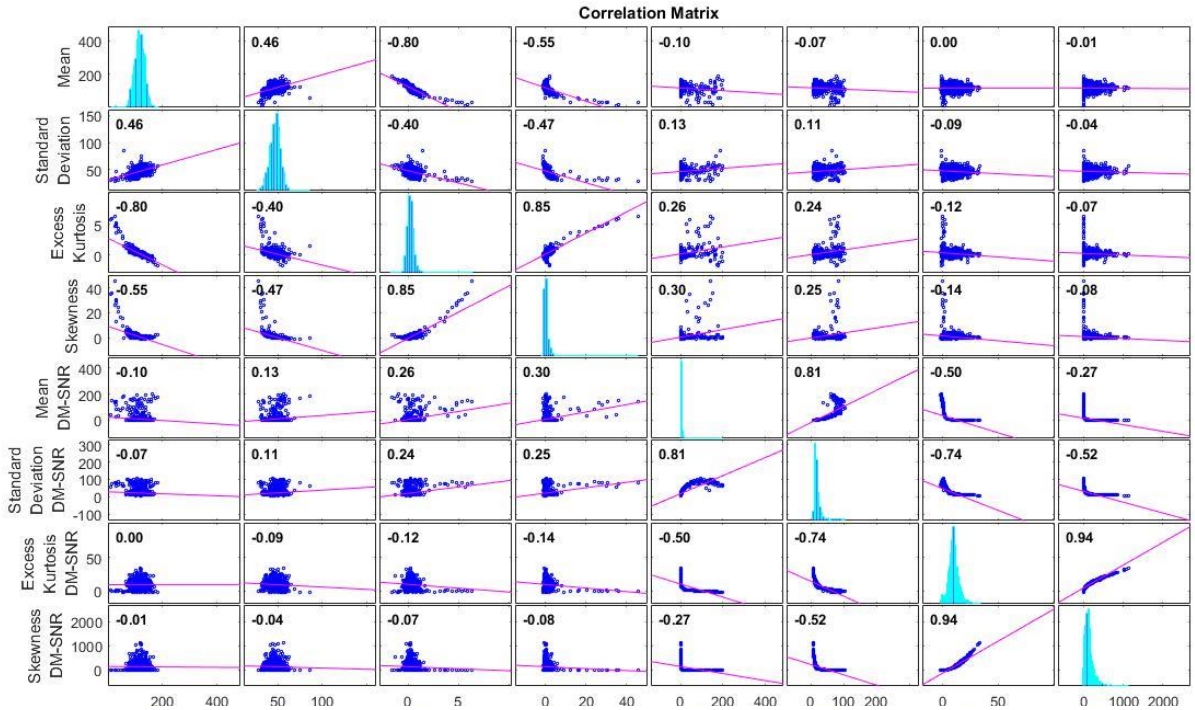


Εικόνα 0.8 Πυκνότητα πιθανότητας για την ασυμμετρία της καμπύλης DM-SNR για σήμα και θόρυβο

5.1.2 Πίνακες Συσχέτισης



Εικόνα 0.9 Πίνακας συσχέτισης για θόρυβο



Εικόνα 0.10 Πίνακας συσχέτισης για σήμα

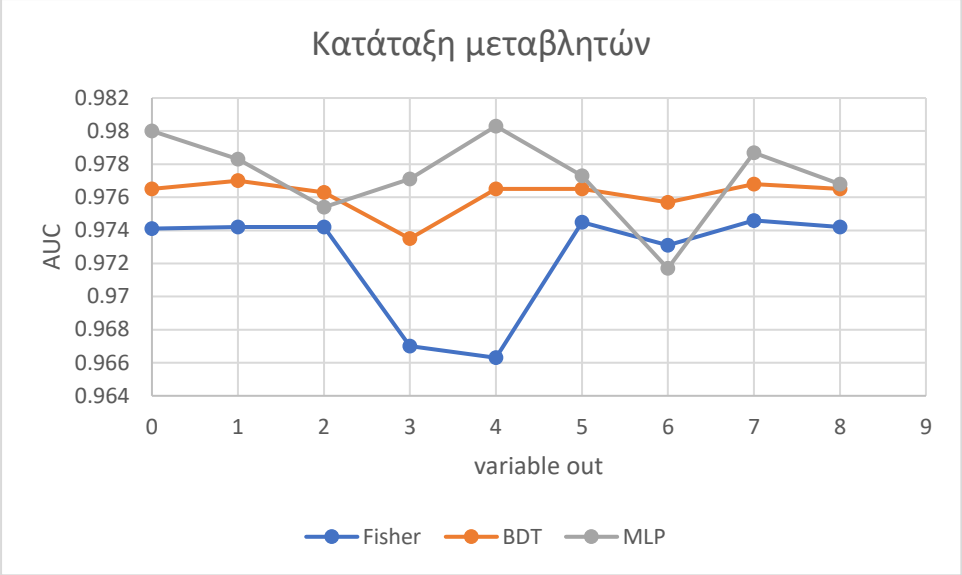
Στις εικόνες παρατηρούμε τους πίνακες συσχέτισης για θόρυβο και σήμα αντιστοίχως. Από αυτά μπορούμε να συμπεράνουμε αν κάποιες μεταβλητές συσχετίζονται αρνητικά ή θετικά μεταξύ τους. Οι ισχυρά συ σχετιζόμενες μεταβλητές επηρεάζουν την ανάλυση λιγότερο και μπορούν να θεωρηθούν ως λιγότερο σημαντικές.

5.1.3 Επιλογή μεταβλητών

Για την επιλογή των χαρακτηριστικών που φέρουν τη μεγαλύτερη διαχωριστική ικανότητα χρησιμοποιήθηκε ως μέτρο απόδοσης το εμβαδόν κάτω από την καμπύλη (AUC/ Area Under Curve) του λειτουργικού χαρακτηριστικού δείκτη (ROC/ Receiver Operating Characteristics. Η κατάταξη των μεταβλητών για όλες τις μεθόδους έγινε αφαιρώντας κάθε φορά και από μία μεταβλητή και υπολογίζοντας ξανά την απόδοση του κάθε ταξινομητή . Για το νευρωνικό δίκτυο συγκεκριμένα προτιμήθηκε η αρχιτεκτονική με έξι κρυφά στρώματα, όπως αναφέρεται στη συνέχεια. Στον πίνακα και στο διάγραμμα παρουσιάζονται οι τιμές της τιμής AUC για τους διάφορους ταξινομητές. Με βάση την έως τώρα ανάλυση οι μετέπειτα υπολογισμοί έγιναν δίχως τη χρήση της μεταβλητής 4, καθώς κρίθηκε ότι φέρει μικρότερη διαχωριστική ικανότητα. Η απόφαση πάρθηκε με βάση το γεγονός ότι ο ταξινομητής Fisher χρησιμοποιείται για μία πρώτη εκτίμηση της σημαντικότητας των μεταβλητών και παρουσίασε μερική πτώση της απόδοσης όταν αφαιρέθηκε η μεταβλητή 4. Επίσης ο ταξινομητής Fisher είναι κατάλληλος για κανονικές κατανομές που φαίνεται ν αείναι η περίπτωση στο συγκεκριμένο πρόβλημα.

Πίνακας 2 Απόδοση (εμβαδόν κάτω από την καμπύλη (AUC/ Area Under Curve) του λειτουργικού χαρακτηριστικού δείκτη (ROC/ Receiver Operating Characteristics) για τους ταξινομητές Fisher, BDT και MLP με διαδοχική αφαίρεση μίας μεταβλητής

Variable Out	Fisher	BDT	MLP
0	0.9741	0.9765	0.98
1	0.9742	0.977	0.9783
2	0.9742	0.9763	0.9754
3	0.967	0.9735	0.9771
4	0.9663	0.9765	0.9803
5	0.9745	0.9765	0.9773
6	0.9731	0.9757	0.9717
7	0.9746	0.9768	0.9787
8	0.9742	0.9765	0.9768



Εικόνα 0.11 Σχεδιαγραμματική απεικόνιση της απόδοσης (εμβαδόν κάτω από την καμπύλη (AUC/ Area Under Curve) του λειτουργικού χαρακτηριστικού δείκτη (ROC/ Receiver Operating Characteristics) για τους ταξινομητές Fisher, BDT και MLP με διαδοχική αφαίρεση μίας μεταβλητής

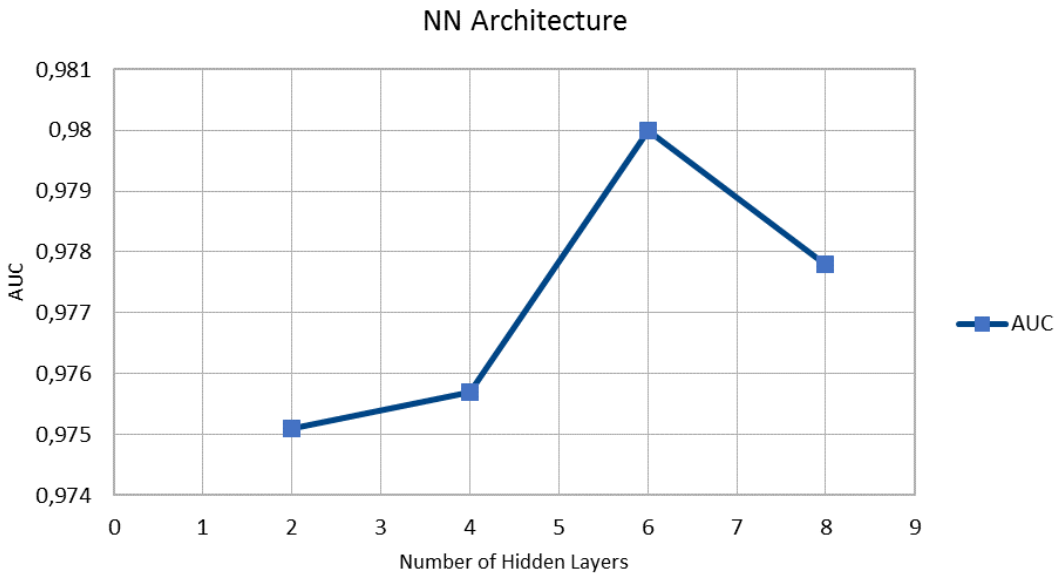
5.1.4 Γραμμικός Ταξινομητής Fisher

Για το γραμμικό ταξινομητή Fisher η επιλογή γεγονότων πραγματοποιείται σε ένα χώρο μετασχηματιζόμενων μεταβλητών με μηδέν γραμμική συσχέτιση διαχωρίζοντας τις μέσες τιμές των κατανομών του σήματος και του θορύβου. Ο γραμμικός ταξινομητής ορίζει έναν άξονα στον Ευκλείδιο χώρο των μεταβλητών εισόδου, στον οποίον αν προβληθούν οι κλάσεις των μεταβλητών εξόδου (σήμα και θόρυβος) θα εμφανίζουν μέγιστο διαχωρισμό, ενώ γεγονότα που ανήκουν στην ίδια κλάση θα βρίσκονται κοντά μεταξύ τους. Η γραμμική ιδιότητα του ταξινομητή καθορίζεται από τους πίνακες συσχέτισης του χώρου διαχωρισμού των μεταβλητών

5.1.5 Νευρωνικό δίκτυο

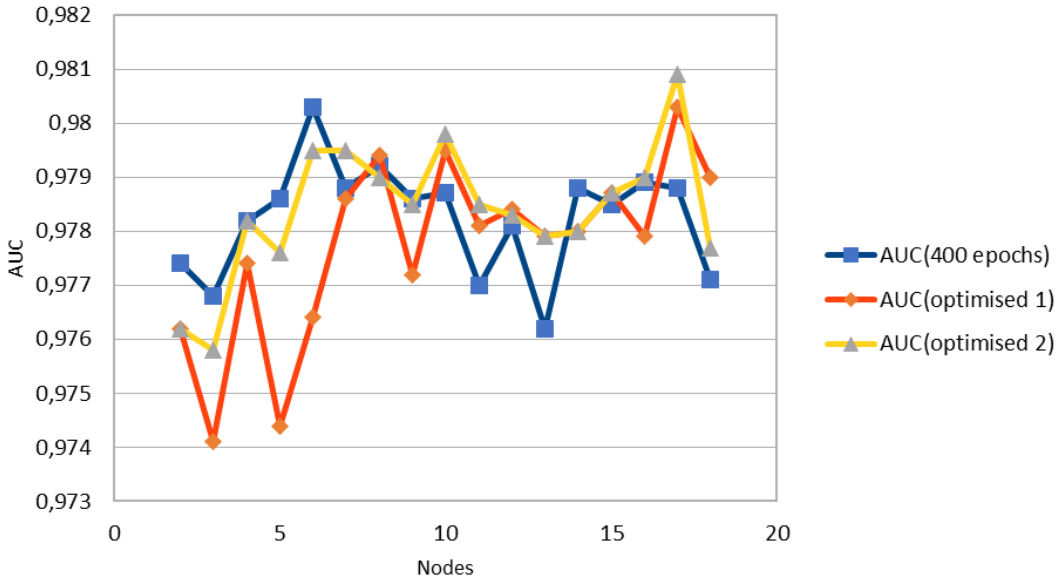
Μελέτη βέλτιστης αρχιτεκτονικής νευρωνικού δικτύου

Από τις προηγούμενες αναλύσεις κρίθηκε ότι η μεταβλητή 4 φέρει μικρή διαχωριστική ικανότητα, οπότε δεν χρησιμοποιήθηκε στους μετέπειτα υπολογισμούς. Αρχικώς διερευνήθηκε η βέλτιστη αρχιτεκτονική του νευρωνικού δικτύου. Για 2, 4, 6 και 8 κρυφά στρώματα παρατηρήθηκε ότι για 6 κρυφά στρώματα το δίκτυο είχε την καλύτερη απόδοση, οπότε και για τους μετέπειτα υπολογισμούς προτιμήθηκε ως αρχιτεκτονική.



Εικόνα 0.12 Σχεδιαγραμματική απεικόνιση της μεταβολής της απόδοσης (εμβαδόν κάτω από την καμπύλη (AUC/ Area Under Curve) του λειτουργικού χαρακτηριστικού δείκτη (ROC/ Receiver Operating Characteristics) του νευρωνικού δικτύου για διαφορετικό αριθμό κρυφών στρωμάτων

Επιπροσθέτως διερευνήθηκε ο βέλτιστος αριθμός κόμβων και εποχών του νευρωνικού δικτύου. Ο αριθμός τον κόμβων εκφράζει την καμπυλότητα του επιπέδου απόφασης και ο αριθμός των εποχών εκφράζει το σύνολο των επαναλήψεων του αλγορίθμου για ολόκληρο το σύνολο δεδομένων εκπαίδευσης. Στις περισσότερες περιπτώσεις χρειάζονται πολλές εποχές για να ολοκληρωθεί η εκπαίδευση ενός νευρωνικού δικτύου με τον αλγόριθμο ανάστροφης μετάδοσης σφάλματος. Όσο αυξάνεται ο αριθμός των κόμβων τόσο αυξάνονται και οι εποχές, καθώς απαιτείται περισσότερος χρόνος για την εκπαίδευση. Αρχικά για σταθερό αριθμό εποχών 400 και για αριθμό κόμβων από 2 έως 18 δεν ήταν ξεκάθαρη ποια αρχιτεκτονική ήταν η καλύτερη. Για την αποφυγή overfitting έγινε βελτιστοποίηση του αριθμού των εποχών, με δύο βελτιστοποιημένες εκδοχές (optimized 1,2). Παρατηρούμε ότι από 7 κόμβους και πάνω η κατανομή σταθεροποιείται, οπότε επιλέγουμε την αρχιτεκτονική με 6 κόμβους και 200 εποχές, καθώς μεγιστοποιεί την απόδοση με ελαχιστοποίηση των βαθμών ελευθερίας.

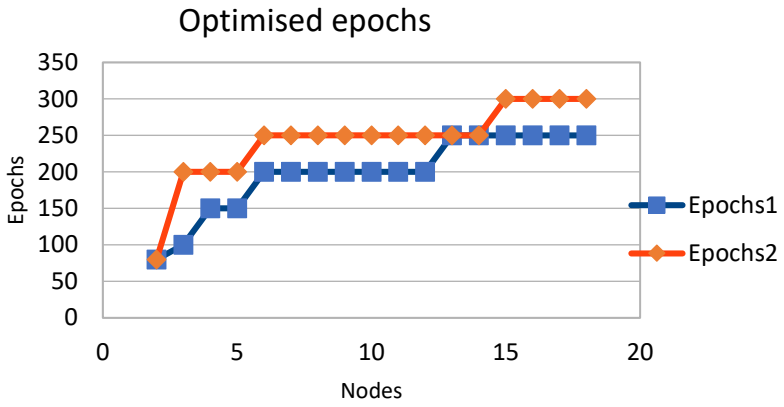


Εικόνα 0.13 Σχεδιαγραμματική απεικόνιση της μεταβολής της απόδοσης (εμβαδόν κάτω από την καμπύλη (AUC/ Area Under Curve) του λειτουργικού χαρακτηριστικού δείκτη (ROC/ Receiver Operating Characteristics) του νευρωνικού δικτύου για διαφορετικό αριθμό κόμβων και εποχών

Number Of Nodes	AUC(400 epochs)	AUC(optimised 1)	AUC(optimised 2)
2	0,9774	0,9762	0,9762
3	0,9768	0,9741	0,9758
4	0,9782	0,9774	0,9782
5	0,9786	0,9744	0,9776
6	0,9803	0,9764	0,9795
7	0,9788	0,9786	0,9795
8	0,9792	0,9794	0,979
9	0,9786	0,9772	0,9785
10	0,9787	0,9795	0,9798
11	0,977	0,9781	0,9785
12	0,9781	0,9784	0,9783
13	0,9762	0,9779	0,9779
14	0,9788	0,978	0,978
15	0,9785	0,9787	0,9787
16	0,9789	0,9779	0,979
17	0,9788	0,9803	0,9809
18	0,9771	0,979	0,9777

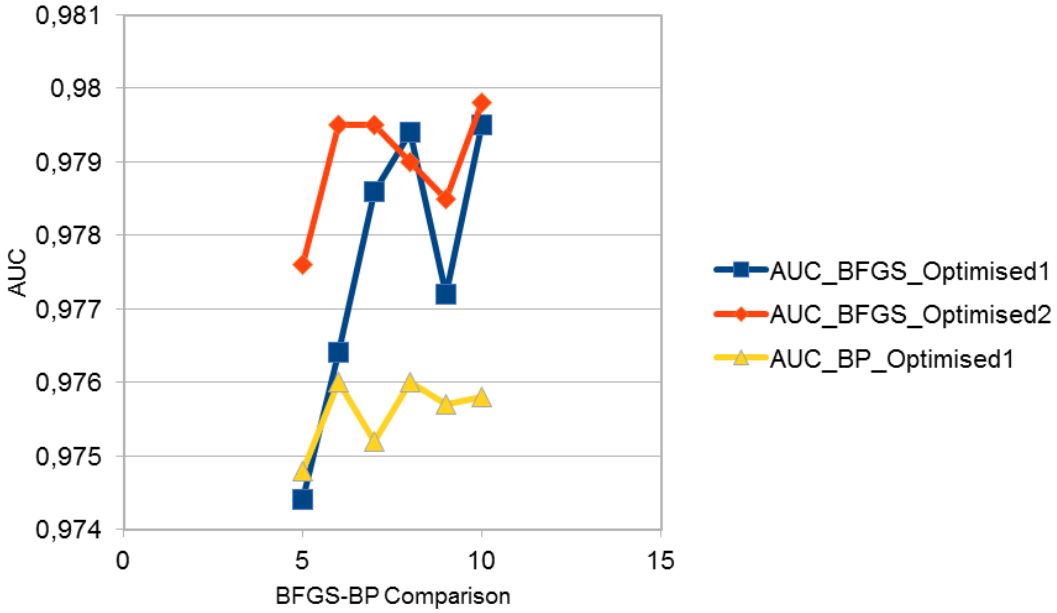
Number of Nodes	Epochs1	Epochs2
2	80	80
3	100	200
4	150	200
5	150	200
6	200	250
7	200	250
8	200	250
9	200	250
10	200	250
11	200	250
12	200	250
13	250	250
14	250	250
15	250	300
16	250	300
17	250	300
18	250	300

Εικόνα 0.14 Μεταβολή της απόδοσης (εμβαδόν κάτω από την καμπύλη (AUC/ Area Under Curve) του λειτουργικού χαρακτηριστικού δείκτη (ROC/ Receiver Operating Characteristics) του νευρωνικού δικτύου για διαφορετικό αριθμό κρυφών στρωμάτων (δεξιά) και οι διαφορετικοί συνδυασμοί αριθμού κόμβων και εποχών που χρησιμοποιήθηκαν κατά τη βελτιστοποίηση



Εικόνα 0.15 Διαφορετικοί συνδυασμοί αριθμού κόμβων και εποχών που χρησιμοποιήθηκαν κατά τη βελτιστοποίηση του νευρωνικού δικτύου

Σύγκριση αλγορίθμων Back Propagation και Broyden-Fletcher-Goldfarb-Shannon (BFGS) για την εκπαίδευση του νευρωνικού δικτύου



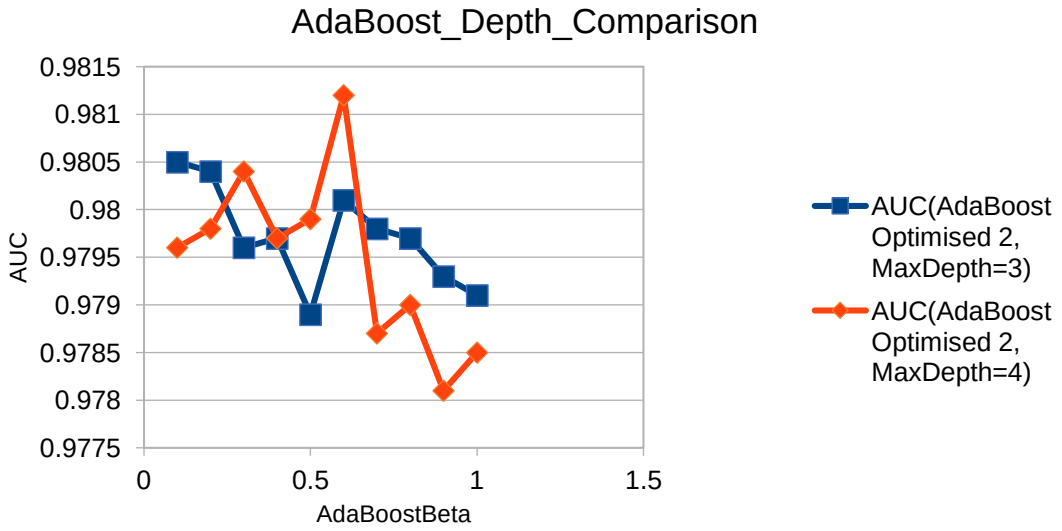
Εικόνα 0.16 Μεταβολή της απόδοσης (εμβαδόν κάτω από την καμπύλη (AUC/ Area Under Curve) του λειτουργικού χαρακτηριστικού δείκτη (ROC/ Receiver Operating Characteristics) για τους αλγόριθμους BFGS και BP του νευρωνικού δικτύου για διαφορετικές βέλτιστες εκδοχές

Φαίνεται ότι καλύτερη απόδοση έχει ο αλγόριθμος BFGS με τη δεύτερη εκδοχή του αριθμού των εποχών. Με τον αλγόριθμο Back Propagation η σύγκλιση είναι γρηγορότερη και οι στατιστικές διακυμάνσεις μεγαλύτερες, ενώ με τον αλγόριθμο Broyden-Fletcher-Goldfarb-Shannon παρατηρούμε μικρότερες στατιστικές διακυμάνσεις, αλλά ταχύτερη σύγκλιση. Επίσης παρατηρείται σταθεροποίηση της

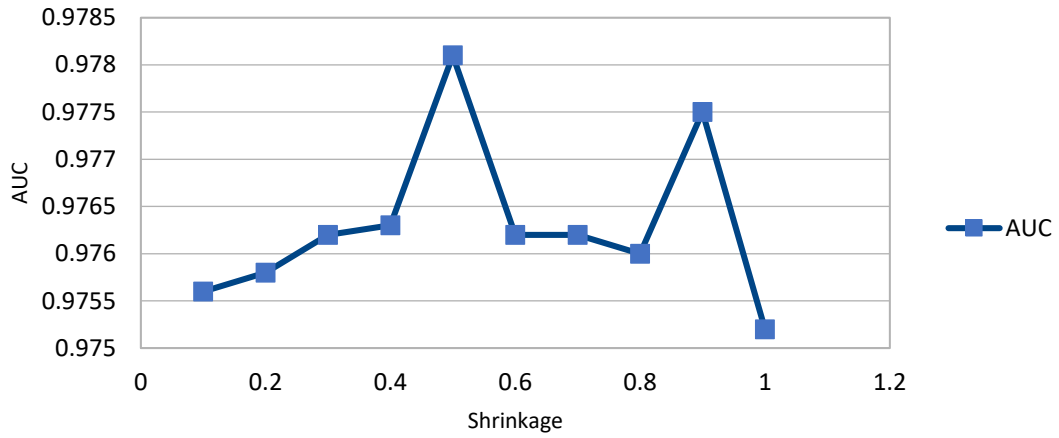
απόδοσης και των δύο συστημάτων για αριθμό κόμβων από εφτά και πάνω. Για να πετύχουμε την καλύτερη απόδοση με όσο το δυνατόν λιγότερους βαθμούς ελευθερίας επιλέχθηκε η αρχιτεκτονική με 6 κρυφά στρώματα, 7 κόμβους, 200 εποχές και αλγόριθμο ρύθμισης βαρών τον BFGS.

5.1.6 Ενισχυμένα Δένδρα αποφάσεως

Στην ανάλυση με χρήση της μεθόδου ενισχυμένων δένδρων απόφασης έγινε εφαρμογή και σύγκριση δύο μεθόδων Boosting, των AdaBoost και Grandient Boost. Για τις δύο μεθόδους πραγματοποιήθηκε βελτιστοποίηση των παραμέτρων του ρυθμού μάθησης shrinkage για GradBoost και AdaBoostBeta για AdaBoost. Όπως ήταν αναμενόμενο ασθενής ταξινομητές, δηλαδή δένδρα με βάθος 3-4 είχανε καλύτερη απόδοση. Συγκεκριμένα για MaxDepth= 4 παρατηρήθηκαν στατιστικές διακυμάνσεις, οπότε προτιμήθηκε μέγιστο βάθος ίσο με 3 και μέθοδο βελτιστοποιήσης την AdaBoost. Η παράμετρος β επιλέχθηκε 0.6, καθώς παρατηρείται σταθεροποίηση της απόδοσης από αυτή την τιμή και μετά.



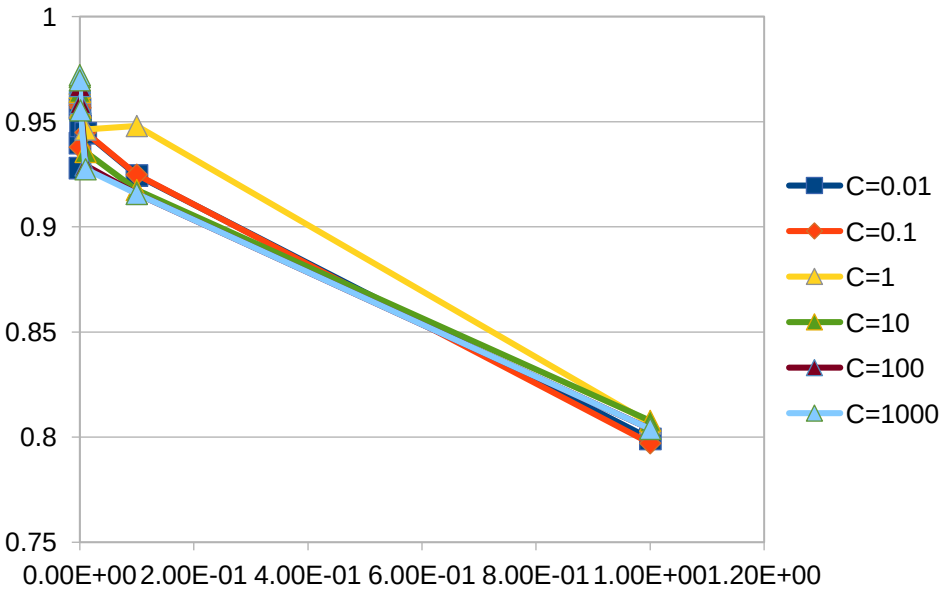
Εικόνα 0.17 Μεταβολή της απόδοσης (εμβαδόν κάτω από την καμπύλη (AUC/ Area Under Curve) του λειτουργικού χαρακτηριστικού δείκτη (ROC/ Receiver Operating Characteristics) με αλλαγές στην τιμή της μεταβλητής β του αλγορίθμου AdaBoost και για MaxDepth 3 και 4



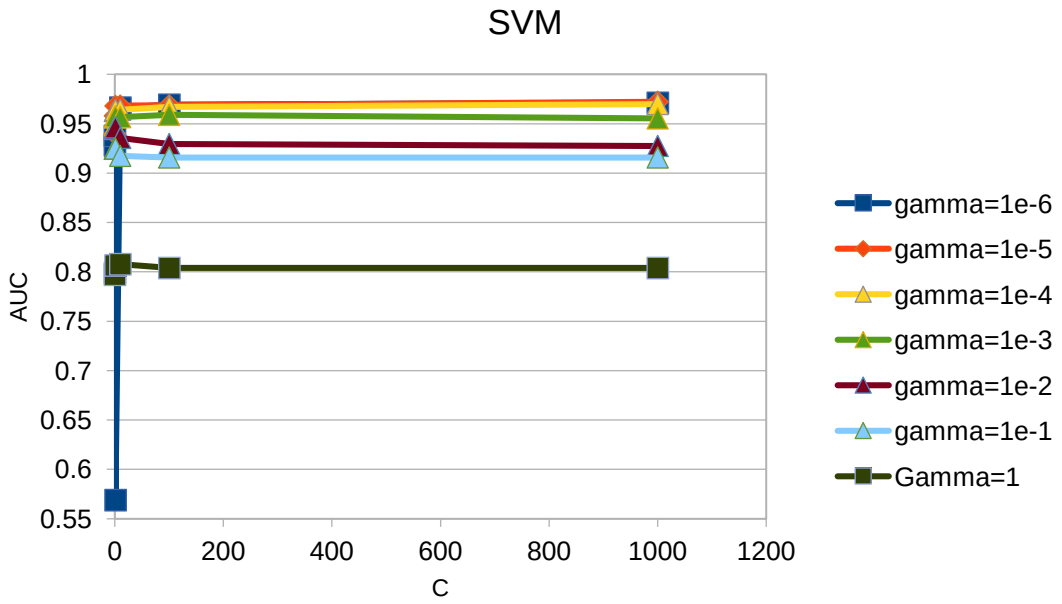
Εικόνα 0.18 Μεταβολή της απόδοσης (εμβαδόν κάτω από την καμπύλη (AUC/ Area Under Curve) του λειτουργικού χαρακτηριστικού δείκτη (ROC/ Receiver Operating Characteristics) για αλλαγές της μεταβλητής Shrinkage του αλγορίθμου BP

5.1.7 SVM

Η TMVA υποστηρίζει μόνο γκαουσιανό πυρήνα για μη γραμμικό SVM και για τη βελτιστοποίηση της απόδοσης είναι απαραίτητη η διερεύνηση των τιμών των παραμέτρων πυρήνα (gamma) και παραμέτρου κόστους C. Από την εικόνα 5.19 παρατηρούμε μία γενική τάση αύξησης της παραμέτρου C να οδηγεί σε μείωση της απόδοσης με βέλτιστη απόδοση για C = 1 και από την εικόνα 5.20 ότι αύξηση της παραμέτρου gamma οδηγεί σε σταθεροποίηση της απόδοσης.



Εικόνα 0.19 Μεταβολή της απόδοσης για διάφορες τιμές της παραμέτρου κόστους C



Εικόνα 0.20 Μεταβολή της απόδοσης για διάφορες τιμές της παραμέτρου γ

6. Συμπεράσματα

Όλες οι μέθοδοι ταξινόμησης παρουσίασαν καλή απόδοση με καλύτερη τα ενισχυμένα δένδρα αποφάσεως (BDT). Με χρήση διερεύνησης για το ποιες μεταβλητές φέρουν μεγαλύτερη διαχωριστική ικανότητα με μέτρο απόδοσης το εμβαδόν κάτω από την καμπύλη (AUC/ Area Under Curve) του λειτουργικού χαρακτηριστικού δείκτη (ROC/ Receiver Operating Characteristics, η ανάλυση συνεχίστηκε δίχως χρήση της μεταβλητής 4. Έπειτα από διερεύνηση για το νευρωνικό δίκτυο επιλέχθηκε αρχιτεκτονική 6 κρυφά στρώματα, 7 κόμβους, 200 εποχές και αλγόριθμο ρύθμισης βαρών τον BFGS. Για τα ενισχυμένα δένδρα απόφασης επιλέχθηκε χρήση της μεθόδου βελτιστοποίησης AdaBoost με τιμή παραμέτρου β = 0.6 και βάθος (MaxDepth) 3. Για γραμμικό SVM έγινε προτίμηση μικρής τιμής παραμέτρου ανεκτικότητας σφάλματος C ίσο με 1 και παραμέτρου πυρήνα gamma ίσο με 200.

Τα αποτελέσματα μας δείχνουν ότι με χρήση απλών στατιστικών και αλγορίθμων μηχανικής μάθησης υποψήφια σήματα πάλσαρ μπορούν να καταταχθούν σε σήμα και σε θόρυβο με πολύ υψηλή απόδοση.

7. References

A. Hoecker, P. S.-V. (2007, March 4). *TMVA - Toolkit for Multivariate Data Analysis*. Ανάκτηση από https://arxiv.org/abs/physics/0703039v5

Bates, S. D. (2012). The high time resolution universe pulsar survey vi. an artificial neural network and timing of 75 pulsars. *Monthly Notices of the Royal Astronomical Society*, vol. 427, no. 2, pp. 1052-1065.

Broyden, C. (1970). The Convergence of a Class of Double-rank Minimization Algorithms. *J.Inst. of Math. and App. 6.*, 76 .

Brun, F. a. (1997). ROOT - An Object Oriented Data Analysis Framework. *Nucl.Inst. Meth. in Phys. Res.*, A 389, 81.

Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Miningand Knowledge Discovery, 2, 1* .

Choi, S. (χ.χ.). Διαφάνειες Διαλέξεων Μαθήματος "MachineLearning". Korea: Department of Computer Science, Pohang University of Science and Technology.

Cofield, C. (2016, April). *space.com*. Ανάκτηση από https://www.space.com/32661-pulsars.html

Cofield, C. (April 22, 2016). What are pulsars? <*https://www.space.com/32661-pulsars.html*>.

D.R., L. (2009). Radio Pulsar Statisitics.". *InNeutron Stars and Pulsars, editedby W. Becker, volume 357 ofAstrophysics and Space Science Library* , pages1–17. Springer.

D.R., L. (2008). "Binary and Millisecond Pulsars. *Living Reviews in Relativity*, 11 (8):1–90.

Drucker, H. (ICML 1997). Improving regressors using boosting techniques. ,*In D. H. Fisher (Ed.), Proceedingsof the fourteenth international conference on machine learning* (σ. (pp. 107115). Nashville, TN, USA, July 812.: Morgan Kaufmann, ISBN 1558604.

Dua, D. a. (2018). UCI Machine Learning Repository. Irvine, CA.

Eatough R.P., M. N. (2001). Selection of radio pulsar candidates using artificial neu-ral networks. *Monthly Notices of the Royal Astronomical Society*, 407(4):2443–2450.

Eatough R.P., M. N. (2010). Selection of radio pulsar candidates using artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 407(4):2443–2450.

Fletcher, R. (1970). A New Approach to Variable MetricAlgorithms. *Computer J. 13*, 317 .

Ghosh. (2007). Rotation and Accretion Powered Pulsars. *World Scientific Series in Astronomy and Astrophysics*, volume 10 of World Scientific Publishing Co., first edition.

Goldfarb, D. (1970). A Family of Variable Metric UpdatesDerived by Variational Means. *Math. Comp. 24*, 23.

Gutierrez-Osuna, R. (χ.χ.). Διαφάνειες Διαλέξεων Μαθήματος "PatternAnalysis". Department of Computer Science and Engineering, Texas A&M University.

Haykin, S. (1996). Adaptive Filter Theory,3rd edition. *Englewood Cliffs, NJ:Prentice-Hall*.

Haykin, S. (1999). Neural Networks, A Comprehensive Foundation, 2nd Edition. *Prentice Hall International*.

Hessels J.W.T., R. S. (2007). A 1.4GHz Arecibo Survey for Pulsars in Globular Clusters. *Astrophysical Journal*, 670:363–378.

Hogden J., V. W. (2012). Comparison of Radio-frequency Interference Mitigation Strategies for Dis-persed Pulse Detection. *The Astrophysical Journal*, 747(2):141.

Ian H.Witten, E. F. (2005). *Data Mining, 2nd edition.* Morgan Kaufmann Publishers, Elsevier.

Jin, R. (χ.χ.). Διαφάνειες διαλέξεων Μαθήματος "DataMining". Computer Science Kent State University: 2007.

Keith M.J., J. A. (2010). The High Time Resolution Universe Pulsar Survey -I. System Configuration and Initial Discoveries. *Monthly Notices of the RoyalAstronomical Society*, 409(2):619–627.

Keith, M. J. (2010). The High Time Resolution Universe Pulsar Survey - I. System Configuration and Initial Discoveries'. *Monthly Notices of the Royal Astronomical Society*, vol. 409, pp. 619-627.

Kotsiantis, S. (2007). Supervised Machine Learning: A review of Classification techniques. *Informatica31*, 249-268.

L, L. (2012). A Search for Radio Pulsars: from Millisecond Pulsars to Magnetars. Swinburne University, Faculty of Information and CommunicationTechnology.

Large M.I., V. A. (1968). Pulsar Search at the MolongloRadio Observatory. *Nature*, 220:753–756.

Lee, K. J. (2013). PEACE: pulsar evaluation algorithm for candidate extraction a software package for post-analysis processing of pulsar survey candidates. *Monthly Notices of the Royal Astronomical Society*, vol. 433, no. 1, pp. 688-694.

Lorimer. (2008). Binary and Millisecond Pulsars. *Living Reviews in Relativity*, 11(8):1–90.

Lorimer D.R. and Kramer M. (2005). Handbook of pulsar astronomy. *Cambridge Uni-versity Press*.

Lyon, R. J. (2015). *PulsarFeatureLab*. Ανάκτηση από https://figshare.com/articles/Pulsar_Feature_Lab/1536472/1

Lyon, R. J. (2016). Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, 459 (1), 1104-1123.

Lyon, R. J. (2016). *Why are pulsars hard to find?* University of Manchester.

Manchester R.N., H. G. (2005a). The Australia TelescopeNational Facility Pulsar Catalogue. *The Astronomical Journal*, 129(4):1993.

Manchester R.N., H. G. (2005b). The Australia TelescopeNational Facility Pulsar Catalogue. *"The Astronomical Journal*, 129(4):1993–2006.

Manchester R.N., L. A. (2001). The Parkes multi-beam pulsar survey - I. Observing and data analysissystems, discovery and timing of 100 pulsars. *Monthly Notices of the RoyalAstronomical Society*, 328(1):17–35.

Mehryar Mohri, A. R. (2012). Foundations of Machine Learning. *MIT Press*.

Morello V., B. E. (2014). SPINN: a straightforward machine learning solution to the pulsar candi-date selection problem. *Monthly Notices of the Royal Astronomical Society*, 443(2):1651–1662.

Morello, V. (2014). SPINN: a straightforward machine learning solution to the pulsar candidate selection problem. *Monthly Notices of the Royal Astronomical Society*, vol. 443, no. 2, pp. 1651-1662.

Nello Cristianini, J. S.-T. (2000). Introduction to Support Vector Machines. *Cambridge University Press*.

Petros Xanthopoulos, P. M. (2013). Robust Data Mining. *Springer*.

Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. *inB. Scholkopf, C. Burges and A. Smola, eds., Advances in Kernel Methods – Support VectorLearning, MIT Press*, ch.12, pp. 185.

Popescu, J. F. (2003). *Gradient Directed Regularization for Linear Regression andClassification.* Statistics Department, Stanford University.

R. J. Lyon, B. W. (2016). Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, 459 (1), 1104-1123.

S. Keerthi, S. S. (1999). *Improvements to Platt's SMO algorithm for SVM classifier design.* Technical Report CD-99-14, Dept. of Mechanical andProduction Engineering, Natl. Univ. Singapore, Singapore.

Schapire, Y. F. (1997). J. of Computer and System Science 55. 119.

Sergios Theodoridis, K. K. (2008). PatternRecognition. *AcademicPress,Elsevier*.

Shannon, D. (1970). Conditioning of Quasi-Newton Methods for Function Minimization. *Math. Comp. 24*, 647.

Stearns, W. B. (1985). Adaptive Signal Processing. *Englewood Cliffs, NJ: Prentice-Hall*.

Stovall K., L. D. (2013). Searching for millisecond pul-sars: surveys, techniques and prospects. *Classical and Quantum Gravity*, 30(22):224003.

Stuart Russell, P. N. (2005). Τεχνητή Νοημοσύνη Μια σύγχρονη προσέγγιση. *2η Αμερικάνικη έκδοση, εκδόσεις Κλειδάριθμος,*.

Thornton. (2013). *The High Time Resolution Radio Sky.* University of Manchester, Jodrell Bank Centre for Astrophysics School of Physics and Astronomy.