ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Εύρεση Προδιαθεσικών Παραγόντων και Εκτίμηση Κινδύνου Εμφάνισης Διαβητικής Αμφιβληστροειδοπάθειας με Χρήση Μεθόδων Τεχνητής Νοημοσύνης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΒΑΣΙΛΙΚΗΣ ΛΙΑΚΩΝΗ

Επιβλέπουσα : Κωνσταντίνα Νικήτα
Καθηγήτρια Ε.Μ.Π.

Αθήνα, Νοέμβριος 2010

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ
ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Εύρεση Προδιαθεσικών Παραγόντων και Εκτίμηση Κινδύνου Εμφάνισης Διαβητικής Αμφιβληστροειδοπάθειας με Χρήση Μεθόδων Τεχνητής Νοημοσύνης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΒΑΣΙΛΙΚΗΣ ΛΙΑΚΩΝΗ

Επιβλέπουσα : Κωνσταντίνα Νικήτα
Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 30η Νοεμβρίου 2010.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Κωνσταντίνα Νικήτα Καθηγήτρια Ε.Μ.Π.	Διονύσιος-Δημήτριος Κουτσούρης Καθηγητής Ε.Μ.Π.	Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2010

(Υπογραφή)

.....

ΒΑΣΙΛΙΚΗ ΛΙΑΚΩΝΗ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Σκοπός της παρούσας διπλωματικής εργασίας ήταν η ανάπτυξη μίας μεθοδολογίας για την εύρεση των προδιαθεσικών παραγόντων της αμφιβληστροειδοπάθειας ως μακροπρόθεσμης επιπλοκής του Διαβήτη Τύπου ΙΙ και την εκτίμηση του κινδύνου εμφάνισής της. Η ανάγκη για την επίτευξη των δύο αυτών στόχων υπαγορεύεται από την έλλειψη αξιόπιστης μεθόδου πρόληψης και θεραπείας της νόσου, τον μη πλήρως κατανοητό μηχανισμό εκδήλωσής της και τις σοβαρές επιπτώσεις της στους πάσχοντες. Η μεθοδολογία που αναπτύχθηκε βασίστηκε σε μεθόδους Τεχνητής Νοημοσύνης και χρησιμοποίησε δεδομένα ασθενών με Διαβήτη τύπου ΙΙ από το Διαβητολογικό Κέντρο του Ιπποκράτειου Νοσοκομείου Αθηνών.

Συγκεκριμένα, υλοποιήθηκε ένας Γενετικός Αλγόριθμος με συνάρτηση καταλληλότητας που βασίζεται στην έννοια της Αμοιβαίας Πληροφορίας. Ο Γενετικός Αλγόριθμος έδωσε σαν έξοδο μερικά υποσύνολα χαρακτηριστικών με μεγάλη προγνωστική αξία για την εμφάνιση της νόσου. Τα υποσύνολα αυτά αξιολογήθηκαν με τη χρήση Τεχνητών Νευρωνικών Δικτύων και επιλέχθηκε αυτό που οδηγεί στη μεγαλύτερη ακρίβεια πρόβλεψης. Τέλος, υλοποιήθηκε το Νευρωνικό Δίκτυο που αποτελεί το τελικό μοντέλο εκτίμησης του κινδύνου εμφάνισης της διαβητικής αμφιβληστροειδοπάθειας και αξιολογήθηκε συστηματικά.

Η παρούσα διπλωματική μπορεί να θεωρηθεί σαν μία επιδημιολογική μελέτη με χρήση "έξυπνων" μεθόδων και μπορεί να συντελέσει στην αποτελεσματικότερη πρόληψη της νόσου. Το μοντέλο πρόβλεψης που υλοποιήθηκε μπορεί να χρησιμοποιηθεί για την υποστήριξη της λήψης προσωποποιημένων ιατρικών αποφάσεων και να συμβάλει στη σωστότερη διαχείριση του Διαβήτη.

Λέξεις Κλειδιά: Διαβήτης Τύπου ΙΙ, Διαβητική Αμφιβληστροειδοπάθεια, Προδιαθεσικοί Παράγοντες, Επιλογή Χαρακτηριστικών, Μοντέλο Πρόβλεψης, Γενετικοί Αλγόριθμοι, Τεχνητά Νευρωνικά Δίκτυα

Abstract

This thesis aims at the development of a methodology in order to identify the risk factors for retinopathy, as a long-term complication of Type II Diabetes, and to estimate a person's risk for its development. The need for the accomplishment of these two objectives is imposed by the lack of reliable prevention and treatment methods, the not fully understood mechanism of the disease development and its serious impact on the person. The methodology was based on Artificial Intelligence methods, and made use of Type II Diabetes patients' data, obtained from the Diabetes Center of Athens Hospital Ippokrateio.

Specifically, a Genetic Algorithm was implemented using a fitness function based on the notion of Mutual Information. The Genetic Algorithm gave as output a few subsets of features with high predictive value for the disease development. These subsets were evaluated using Artificial Neural Networks and the subset that led to the most accurate prediction was selected. Finally, the Neural Network that constitutes the final prediction model for the occurrence of diabetic retinopathy was implemented and evaluated.

This thesis can be considered as an epidemiologic study which makes use of "intelligent" methods and can help towards the more effective personalized prevention of the disease. The prediction model that was implemented can be used to support medical decisions and may contribute to a more efficient Diabetes disease management.

Keywords: Type II Diabetes, Diabetic Retinopathy, Risk factors, Feature Selection, Prediction Model, Genetic Algorithms, Artificial Neural Networks

Ευχαριστίες

Ευχαριστώ θερμά την επιβλέπουσα της παρούσας διπλωματικής Κωνσταντίνα Νικήτα για την εμπιστοσύνη της και την επαγγελματική και ταυτόχρονα φιλική στάση της. Ευχαριστώ επίσης την Κωνσταντία Ζαρκογιάννη για την καθοδήγησή της και το Μάριο Σκευοφύλακα για τη βοήθεια και τις καίριες συμβουλές του. Τέλος, θα ήθελα να ευχαριστήσω το Νίκο Κυρτάτα για την πολύτιμη βοήθεια και την υποστήριξή του.

Στην αδερφή μου,

που είναι πάντα δίπλα μου

και στους γονείς μου,

που με στηρίζουν σε κάθε μου επιλογή.

Πίνακας περιεχομένων

Πρόλογος

Ο Σακχαρώδης Διαβήτης (Diabetes Mellitus), ή πιο συχνά αναφερόμενος απλά ως Διαβήτης, αποτελεί μια χρόνια μεταβολική διαταραχή, η οποία χαρακτηρίζεται από αύξηση της συγκέντρωσης του σακχάρου στο αίμα. Πρόκειται για μία μη ιάσιμη ασθένεια που συνδέεται με πολλές μακροπρόθεσμες επιπλοκές, καθώς οι συχνές διακυμάνσεις της συγκέντρωσης γλυκόζης στο αίμα πέρα των φυσιολογικών ορίων μπορεί να επηρεάσουν την ομαλή λειτουργία του οργανισμού σε βάθος χρόνου με πολλούς τρόπους.

Μία από τις μακροπρόθεσμες επιπλοκές του Διαβήτη είναι η αμφιβληστροειδοπάθεια (retinopathy), η οποία αφορά στη βλάβη του αμφιβληστροειδή χιτώνα του οφθαλμού. Η διαβητική αμφιβληστροειδοπάθεια συνήθως δεν έχει συμπτώματα στα αρχικά στάδια εμφάνισής της. Επηρεάζει σημαντικά την ποιότητα ζωής του ατόμου και αποτελεί την κύρια αιτία τύφλωσης και διαταραχών της όρασης στις αναπτυγμένες χώρες. Υπάρχουν κάποιες γενικές οδηγίες σχετικά με την πρόληψη της αμφιβληστροειδοπάθειας, όπως η διατήρηση της συγκέντρωσης της γλυκόζης σε φυσιολογικά επίπεδα, η κατάλληλη φαρμακευτική αντιμετώπιση του διαβήτη και η διατήρηση της πίεσης του αίματος σε φυσιολογικά επίπεδα. Γενικά, όμως, δεν υπάρχει εγγυημένη μέθοδος για την πρόληψη και τη θεραπεία της και ο μηχανισμός πρόκλησής της δεν έχει διαλευκανθεί πλήρως.

Τα παραπάνω στοιχεία σε συνδυασμό με την ελλιπή ενημέρωση μεταξύ των ασθενών συνθέτουν ένα σκηνικό με προβλήματα που χρήζουν αντιμετώπισης. Αν και πολλές επιστήμες, όπως η ιατρική, η φαρμακολογία και η στατιστική λειτουργούν προς αυτήν την κατεύθυνση, τα προβλήματα αυτά δεν έχουν ακόμη αντιμετωπιστεί πλήρως.

Στόχος της παρούσας διπλωματικής ήταν η ανάπτυξη μίας μεθοδολογίας που θα αποτελέσει ένα βοήθημα για την αντιμετώπιση των προβλημάτων αυτών. Συγκεκριμένα, η μεθοδολογία που αναπτύχθηκε αφορά στην εύρεση των παραγόντων που συνδέονται με την εμφάνιση της διαβητικής αμφιβληστροειδοπάθειας σε ασθενείς με διαβήτη Τύπου ΙΙ και στην υλοποίηση ενός μοντέλου πρόβλεψης/εκτίμησης του κινδύνου εμφάνισης της νόσου με μεθόδους Τεχνητής Νοημοσύνης. Αποτελεί με τον τρόπο αυτό ένα μικρό παράδειγμα της συμβολή της επιστήμης των υπολογιστών σε προβλήματα ιατρικής φύσεως.

Η παρούσα εργασία απαρτίζεται από 6 Κεφάλαια. Είναι δομημένη έτσι ώστε ο αναγνώστης να εισάγεται σταδιακά στο πρόβλημα και στη μεθοδολογία που ακολουθείται για την επίλυσή του. Το Κεφάλαιο 1 αποτελεί μία ευρεία εισαγωγή στο Σακχαρώδη Διαβήτη. Παρουσιάζονται επιπλέον οι επιπλοκές με τις οποίες συνδέεται και στατιστικά στοιχεία του διαβήτη, ώστε να γίνει αντιληπτό το μέγεθος του προβλήματος. Στο τέλος του κεφαλαίου παρουσιάζεται συνοπτικά το αντικείμενο της διπλωματικής.

Το Κεφάλαιο 2 παρουσιάζει στον αναγνώστη το θεωρητικό υπόβαθρο στο οποίο στηρίζεται η μεθοδολογία που αναπτύσσεται στην παρούσα διπλωματική. Αναλύονται οι έννοιες της Ταξινόμησης (Classification), της Επιλογής Χαρακτηριστικών (Feature Selection), καθώς επίσης και υπολογιστικά εργαλεία Τεχνητής Νοημοσύνης, όπως οι Γενετικοί Αλγόριθμοι (Genetic Algorithms) και τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks). Αναφερόμαστε, τέλος, στην Αμοιβαία Πληροφορία (Mutual Information), μία έννοια από το πεδίο της Θεωρίας Πληροφορίας (Information Theory).

Στο Κεφάλαιο 3 παραθέτουμε κάποιες εργασίες που έχουν γίνει και είναι σχετικές με το αντικείμενο της διπλωματικής. Παρουσιάζουμε συνοπτικά εργασίες που έχουν πραγματοποιηθεί σχετικά με την εύρεση των σχετικών με αμφιβληστροειδοπάθεια παραγόντων και με την εκτίμηση του κινδύνου εμφάνισης επιπλοκών του διαβήτη, καθώς επίσης και εργασίες που χρησιμοποιούν κάποιες από τις μεθόδους που χρησιμοποιούνται και στην παρούσα διπλωματική. Γίνεται, επίσης, μία εισαγωγή στη μεθοδολογία που αναπτύσσεται στην παρούσα διπλωματική, στη λογική της και στον τρόπο με τον οποίο φιλοδοξεί να συνεισφέρει και να διαφοροποιηθεί από ότι έχει γίνει προς την ίδια κατεύθυνση.

Το Κεφάλαιο 4 περιλαμβάνει την περιγραφή της μεθοδολογίας και των αλγορίθμων που χρησιμοποιήθηκαν για την επίλυση του προβλήματος. Περιγράφονται αρχικά τα δεδομένα στα οποία βασιζόμαστε και στη συνέχεια παρουσιάζονται αναλυτικά όλα τα στάδια που περιλαμβάνει η μεθοδολογία που αναπτύχθηκε.

Στο Κεφάλαιο 5 παρουσιάζεται η διαδικασία που ακολουθήθηκε για τον έλεγχο της μεθοδολογίας. Περιγράφονται τα κριτήρια στα οποία βασίστηκε η αξιολόγηση και αναλύονται τα αποτελέσματά που προέκυψαν.

Στο Κεφάλαιο 6 συνοψίζονται τα αποτελέσματα και τα συμπεράσματα που προκύπτουν από τη διπλωματική και προτείνονται κάποιες μελλοντικές επεκτάσεις της.

1

Εισαγωγή

Το κεφάλαιο αυτό αποτελεί μία εισαγωγή στο Σακχαρώδη Διαβήτη και στις επιπλοκές του. Ο αναγνώστης εισάγεται έτσι στο πρόβλημα που πραγματεύεται η παρούσα διπλωματική. Στο τέλος του κεφαλαίου παρουσιάζεται συνοπτικά το αντικείμενο της εργασίας.

1.1 Σακχαρώδης Διαβήτης

1.1.1 Ορισμός

Ο Σακχαρώδης Διαβήτης (Diabetes Mellitus) αποτελεί μια μεταβολική διαταραχή που χαρακτηρίζεται από χρόνια υπεργλυκαιμία που οφείλεται είτε σε μειωμένη έκκριση ινσουλίνης, είτε σε μειωμένη δράση ινσουλίνης, είτε σε συνύπαρξη των δύο αυτών δυσλειτουργιών.

Ο οργανισμός μετατρέπει σχεδόν όλες τις τροφές σε γλυκόζη (σάκχαρο). Η γλυκόζη είναι η βασική τροφή, δηλαδή πηγή ενέργειας, των κυττάρων. Για την εισαγωγή όμως της γλυκόζης μέσα στα κύτταρα, είναι απαραίτητη η δράση μίας ορμόνης, της ινσουλίνης. Η ορμόνη αυτή εκκρίνεται από το πάγκρεας, ένα μεγάλο αδένα που βρίσκεται πίσω από το στομάχι. Η ινσουλίνη αποτελεί το "κλειδί" που επιτρέπει την εισαγωγή της γλυκόζης στα κύτταρα του οργανισμού. Όταν το πάγκρεας δεν παράγει αρκετή ινσουλίνη ή η ινσουλίνη που παράγει δεν δρα σωστά, τότε η γλυκόζη δεν εισέρχεται μέσα στα κύτταρα, αυξάνεται η συγκέντρωσή της στο αίμα και "διαβαίνει" μέσω των νεφρών στα ούρα, από όπου προέρχεται και η ονομασία "διαβήτης".

1.1.1.1 Πάγκρεας και Ινσουλίνη

Το πάγκρεας είναι ένας μεγάλος αδένας που βρίσκεται πίσω από το στομάχι. Είναι μεικτός αδένας, δηλαδή είναι και ενδοκρινής και εξωκρινής. Ως εξωκρινής αδένας, το πάγκρεας εκκρίνει το παγκρεατικό υγρό, το οποίο περιέχει πεπτικά υγρά που περνάνε στο λεπτό έντερο και βοηθάνε στη διάσπαση υδατανθράκων, πρωτεϊνών και λιπών. Το μέρος του παγκρέατος στο οποίο οφείλεται η ενδοκρινής λειτουργία του αποτελείται από περίπου ένα εκατομμύριο ομάδες κυττάρων, που ονομάζονται νησίδια του Λάνγκερχανς (islets of Langerhans). Οι ομάδες αυτές μπορούν να διαχωριστούν σε τέσσερις κατηγορίες κυττάρων: Τα α-κύτταρα, τα β-κύτταρα, τα δ-κύτταρα και τα PP-κύτταρα. Στην παρούσα διπλωματική θα ασχοληθούμε μόνο με τις δύο πρώτες κατηγορίες.

Τα α-κύτταρα εκκρίνουν μία ορμόνη που ονομάζεται γλυκαγόνη (glucagon). Η γλυκαγόνη βοηθά στην παρεμπόδιση της υπερβολικής πτώσης της συγκέντρωσης της γλυκόζης στο αίμα, έχει δηλαδή αντίρροπη δράση από αυτή της ινσουλίνης.

Τα β-κύτταρα εκκρίνουν την ινσουλίνη, μία ορμόνη με πολλαπλή δράση. Όπως είπαμε, η ινσουλίνη λειτουργεί σα "μέσο μεταφοράς" της γλυκόζης στα κύτταρα. Η γλυκόζη που δε χρησιμοποιείται άμεσα από τα κύτταρα αποθηκεύεται για μελλοντική χρήση με τη μορφή γλυκογόνου. Το γλυκογόνο είναι ένας πολυσακχαρίτης, αποτελούμενος από πολλά μόρια γλυκόζης και αποθηκεύεται στο ήπαρ και στους μύες. Για τη σύνθεση του γλυκογόνου κυρίαρχο ρόλο παίζει η ινσουλίνη, καθώς είναι υπεύθυνη για τη διέγερση της διαδικασίας σύνθεσής του. Όταν αυξηθούν οι ενεργειακές απαιτήσεις του οργανισμού, ενεργοποιείται μέσω της γλυκαγόνης η διάσπαση του γλυκογόνου σε μόρια γλυκόζης. Το γλυκογόνο που αποθηκεύεται στους μύες προορίζεται μόνο για τοπική κατανάλωση. Το ήπαρ, αντίθετα, έχει την ικανότητα να περάσει την παραγόμενη από τη διάσπαση του γλυκογόνου γλυκόζη στο αίμα με σκοπό τη μεταφορά της στον εγκέφαλο και στους μύες για την κάλυψη των ενεργειακών αναγκών τους.

Ακόμα μία λειτουργία της ινσουλίνης είναι η διέγερση της διαδικασίας σύνθεσης λιπαρών οξέων και τριγλυκεριδίων και η αναστολή της διάσπασης των τελευταίων. Με τον τρόπο αυτό, η ινσουλίνη βοηθάει στην αποθήκευση του λίπους και την παραγωγή πρωτεΐνης στους μύες. Με άλλα λόγια, η ινσουλίνη ελέγχει την αποθήκευση ενέργειας του οργανισμού και συμβάλλει στη διαδικασία ανάπτυξης των μυών.

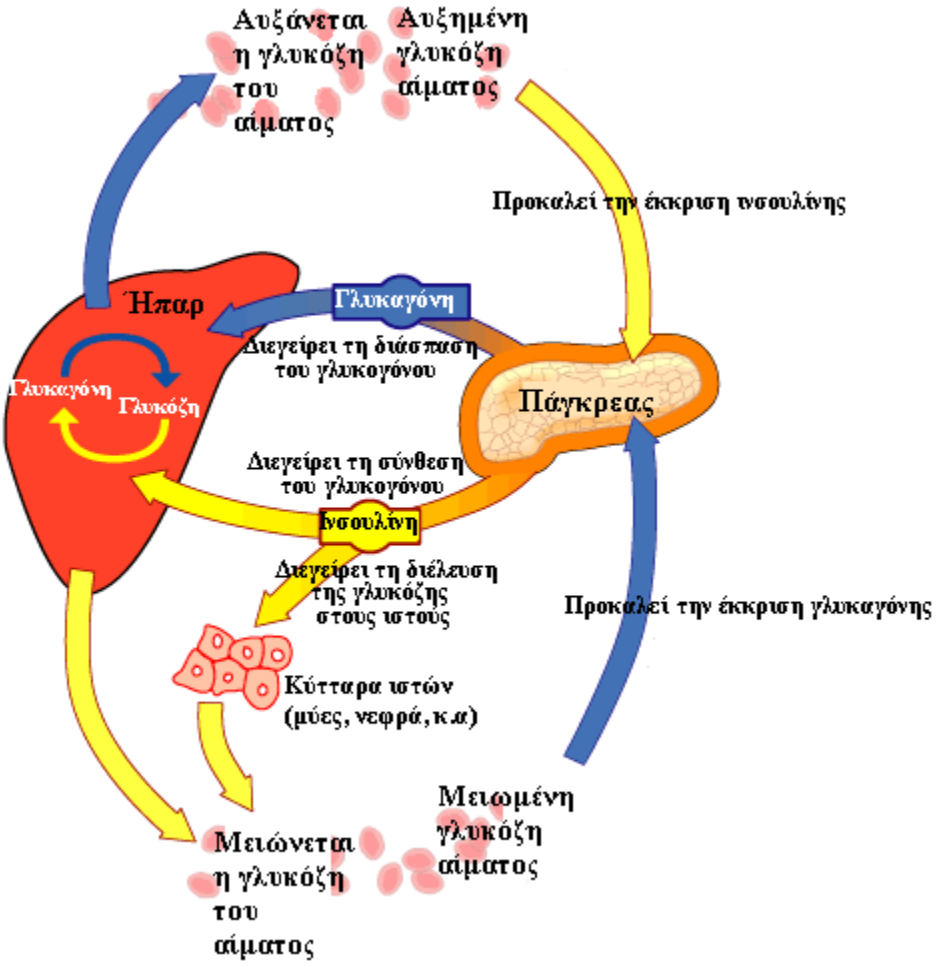
1.1.1.2 Μεταβολισμός υδατανθράκων

Ο έλεγχος του μεταβολισμού πραγματοποιείται βάσει των σχετικών ποσοτήτων ινσουλίνης και γλυκαγόνης. Σε ένα φυσιολογικό άτομο, ο μηχανισμός λειτουργεί ως εξής: Κατά την κατανάλωση και πέψη ενός γεύματος που περιλαμβάνει υδατάνθρακες, τα επίπεδα γλυκόζης του αίματος αυξάνονται. Το γεγονός αυτό επιφέρει την έκκριση της ινσουλίνης και τη μείωση της έκκρισης γλυκαγόνης. Επικρατεί, επομένως, η δράση της ινσουλίνης, η οποία με τη σειρά της μειώνει την

παραγωγή γλυκόζης από το ήπαρ, και ευνοεί τη χρήση της γλυκόζης και τη σύνθεση γλυκογόνου, δηλαδή την αποθήκευση της γλυκόζης στο λίπος, στους μύες και στο ήπαρ. Όσο και η ινσουλίνη και η γλυκόζη είναι σε αφθονία, η σύνθεση του γλυκογόνου συνεχίζει να λαμβάνει χώρα και το ήπαρ να λαμβάνει περισσότερη γλυκόζη από όση εκκρίνει. Κατά την ολοκλήρωση της πέψης, τα επίπεδα γλυκόζης αρχίζουν να πέφτουν, η έκκριση της ινσουλίνης μειώνεται και έτσι, σταματάει η διαδικασία της σύνθεσης γλυκογόνου. Για τις επόμενες 8-12 ώρες που ακολουθούν το γεύμα, οι ανάγκες "καυσίμων " του οργανισμού καλύπτονται από την έκκριση γλυκόζης μέσω του ήπατος. Αν η γλυκόζη του αίματος αρχίσει να πέφτει κάτω από το φυσιολογικό, εκκρίνεται από το πάγκρεας γλυκαγόνη σε μεγάλες ποσότητες. Στην περίπτωση αυτή, ακόμα κι αν η ινσουλίνη είναι σε μη φυσιολογικά υψηλά επίπεδα, επικρατεί η δράση της γλυκαγόνης. Η τελευταία διεγείρει τη διάσπαση του γλυκογόνου σε γλυκόζη με σκοπό την επαναφορά της γλυκόζης αίματος σε φυσιολογικά επίπεδα. Παρακάτω φαίνεται η όλη διαδικασία και σχηματικά.

Με τον τρόπο αυτό, στα μη διαβητικά άτομα η στάθμη της γλυκόζης, δηλαδή του σακχάρου του αίματος, κινείται σταθερά μέσα σε ένα περιορισμένο εύρος. Ακόμα και μετά από ένα πολύ πλούσιο γεύμα, η γλυκόζη του αίματος συνήθως δεν υπερβαίνει τα 150mg/100ml. Η γλυκόζη του αίματος ρυθμίζεται, όπως ένας θερμοστάτης ρυθμίζει τη θερμοκρασία του δωματίου.

Όταν, όμως, η λειτουργία αυτού του "συστήματος αυτόματου ελέγχου" διαταραχθεί, η γλυκόζη που απορροφάται από το γαστρεντερικό σωλήνα δεν μπορεί ούτε να διοχετευθεί στους ιστούς, ούτε να αποθηκευτεί, με αποτέλεσμα την αύξηση της συγκέντρωσής της, δηλαδή την υπεργλυκαιμία. Η παθολογική αυτή κατάσταση συνοψίζεται στον όρο "διαβήτης".



Σχήμα 1.1: Σχηματικό διάγραμμα μεταβολισμού γλυκόζης.

1.1.2 Τύποι Σακχαρώδη Διαβήτη

Υπάρχουν τρεις κύριοι τύποι σακχαρώδη διαβήτη:

- Διαβήτης Τύπου Ι.

Χαρακτηρίζεται από παντελή έλλειψη ή ελάχιστη έκκριση ινσουλίνης. Είναι αυτοάνοσο νόσημα, κατά το οποίο καταστρέφονται τα β-κύτταρα του παγκρέατος που είναι υπεύθυνα για την έκκριση ινσουλίνης. Πρόκειται για αρκετά σπάνιο είδος διαβήτη (5-10% των διαβητικών έχει Τύπου Ι) και εμφανίζεται συνήθως σε παιδιά, για αυτό και έχει ονομαστεί και "νεανικός διαβήτης" αλλά μπορεί να εμφανιστεί και σε ενήλικους. Τα περισσότερα άτομα που εμφανίζουν Διαβήτη Τύπου Ι είναι υγιή κατά τα άλλα. Τα αίτια που οδηγούν σε Διαβήτη Τύπου Ι δεν είναι πλήρως κατανοητά, πιστεύεται όμως ότι σχετίζεται με τη διατροφή και με το στρες.

Από τη στιγμή που θα εμφανιστεί ο Διαβήτης Τύπου Ι, συνήθως δεν μπορεί να θεραπευτεί. Λόγω της πλήρης αδυναμίας έκκρισης ινσουλίνης, πρέπει απαραίτητα να συνοδευτεί από χορήγηση ινσουλίνης, αλλιώς μπορεί να αποβεί θανατηφόρος. Για το λόγο αυτό, του έχει δοθεί

και η ονομασία “ινσουλινοεξαρτώμενος διαβήτης”. Η πιο συνηθισμένη μέθοδος χορήγησης ινσουλίνης είναι σε ενέσιμη μορφή. Άλλες μέθοδοι είναι η εισπνεόμενη ινσουλίνη και οι αντλίες συνεχούς έγχυσης ινσουλίνης. Απαραίτητη φυσικά είναι και η τακτική μέτρηση των επιπέδων γλυκόζης του αίματος, μία διαδικασία που ονομάζεται γλυκαιμικός έλεγχος.

- Διαβήτης Τύπου ΙΙ.

Στην περίπτωση αυτή, είτε το πάγκρεας παράγει μειωμένη ινσουλίνη, είτε η ινσουλίνη που παράγεται έχει μειωμένη δράση, με την έννοια της ελαττωμένης ευαισθησίας των κυττάρων σε αυτή. Ο διαβήτης τύπου ΙΙ είναι ο πιο συχνά εμφανιζόμενος (90% των περιπτώσεων) και εμφανίζεται συνήθως σε μεγαλύτερες ηλικίες. Προδιαθεσιακοί παράγοντες θεωρούνται η ηλικία, η παχυσαρκία, ο τρόπος ζωής και το οικογενειακό ιστορικό. Η παχυσαρκία έχει βρεθεί ότι συμβάλλει κατά 55% στην εμφάνιση του Διαβήτη Τύπου ΙΙ [EOE+04]. Επίσης, στατιστικές έρευνες έχουν δείξει ότι άτομα με αρκετή φυσική δραστηριότητα, υγιεινή διατροφή, που δεν καπνίζουν και καταναλώνουν με μέτρο αλκοόλ, έχουν πολύ μικρότερες πιθανότητες να εμφανίσουν Διαβήτη Τύπου ΙΙ. Και αυτός ο τύπος διαβήτη αποτελεί χρόνια κατάσταση, που συνήθως δε θεραπεύεται και χρήζει σωστής αντιμετώπισης, έτσι ώστε να καθυστερηθεί η εξέλιξη της ασθένειας προς το χειρότερο και να αποφευχθούν οι επιπλοκές που συνδέονται με αυτή. Γενικά, οι στόχοι της αντιμετώπισης του Διαβήτη Τύπου ΙΙ είναι η αποφυγή των επιπλοκών και η διατήρηση της ποιότητας ζωής του ατόμου.

Επειδή τα διαβητικά άτομα αυτού του τύπου είναι σε θέση να παράγουν μία ποσότητα ινσουλίνης, δεν εξαρτώνται από τη εξωγενή χορήγηση ινσουλίνης, τουλάχιστον κατά τα πρώτα στάδια της ασθένειας. Για το λόγο αυτό, ο Διαβήτης Τύπου ΙΙ ονομάζεται και “μη ινσουλινοεξαρτώμενος διαβήτης”. Ο Διαβήτης Τύπου ΙΙ αντιμετωπίζεται αρχικά με σωστή διατροφή, σωματική άσκηση και απώλεια βάρους, ειδικά στα παχύσαρκα άτομα. Σε πολλές περιπτώσεις, οι ενέργειες αυτές μπορούν να βοηθήσουν στην αποκατάσταση της ευαισθησίας των κυττάρων στην ινσουλίνη. Απαραίτητος και πάλι είναι ο τακτικός γλυκαιμικός έλεγχος, από τον ίδιο τον ασθενή. Υπάρχουν επίσης διάφορα είδη φαρμακευτικής αγωγής για τα άτομα με Διαβήτη Τύπου ΙΙ, κάθε ένα από τα οποία λειτουργεί με διαφορετικό τρόπο. Κάποια μειώνουν την έκκριση γλυκόζης από τις αποθήκες γλυκογόνου του ήπατος και αυξάνουν τη διέλευση της γλυκόζης στους ιστούς, άλλα αυξάνουν την ευαισθησία των κυττάρων στη γλυκόζη με διάφορους τρόπους, ενώ άλλα αυξάνουν την έκκριση ινσουλίνης από το πάγκρεας. Σε κάποιες περιπτώσεις, όταν πλέον δεν υπάρχει κλινικό όφελος από τη φαρμακευτική αγωγή, κρίνεται απαραίτητη η εξωγενής χορήγηση ινσουλίνης για τη διατήρηση των φυσιολογικών επιπέδων γλυκόζης στο αίμα. Η εφαρμογή μίας φαρμακευτικής αγωγής είναι αρκετά σύνθετη υπόθεση, καθώς πρέπει να διατηρηθεί μία ισορροπία μεταξύ της ανάγκης για ρύθμιση των επιπέδων γλυκόζης του αίματος και της αποφυγής του κίνδυνου υπογλυκαιμίας. Πρέπει, για το λόγο αυτό, να ακολουθείται πρόγραμμα λήψης κατάλληλων φαρμάκων και τακτικών

γευμάτων. Επισημαίνουμε τέλος ότι κάθε διαβητικό άτομο είναι ξεχωριστή περίπτωση και απαιτούνται τακτικές επισκέψεις στο γιατρό που το παρακολουθεί, με σκοπό την υιοθέτηση της καλύτερης κάθε φοράς τακτικής για την αντιμετώπιση της ασθένειας.

- Διαβήτης Κύησης.

Ο Διαβήτης Κύησης εμφανίζεται για πρώτη φορά κατά τη διάρκεια της εγκυμοσύνης. Και αυτός ο τύπος διαβήτη χαρακτηρίζεται από μειωμένη δράση ή έκκριση ινσουλίνης. Παρουσιάζεται στο 3-5% των κυήσεων. Δεν έχει πολύ εμφανή συμπτώματα και συνήθως διαγιγνώσκεται κατά τους προληπτικούς ελέγχους που πραγματοποιούνται κατά τη διάρκεια της εγκυμοσύνης, μέσω διαγνωστικών δοκιμασιών που ανιχνεύουν υψηλά επίπεδα γλυκόζης στο αίμα. Τα αίτια που προκαλούν Διαβήτη Κύησης δεν είναι γνωστά, πιστεύεται όμως ότι οι ορμόνες που παράγονται κατά τη διάρκεια της εγκυμοσύνης μειώνουν την ευαισθησία των κυττάρων στην ινσουλίνη.

Ο Διαβήτης Κύησης αποτελεί μία θεραπεύσιμη ασθένεια. Τις περισσότερες φορές αντιμετωπίζεται απλά με σωστή διατροφή και σωματική άσκηση, υπάρχουν όμως και περιπτώσεις που μπορεί να χρειαστούν αντιδιαβητικά χάπια. Συνήθως υποχωρεί μετά τον τοκετό, οι γυναίκες όμως που τον εμφανίζουν έχουν αυξημένες πιθανότητες να εμφανίσουν Διαβήτη Τύπου ΙΙ αργότερα στη ζωή τους. Όσον αφορά στα παιδιά των γυναικών που εμφάνισαν Διαβήτη Κύησης, υπάρχει ο κίνδυνος να έχουν αυξημένο βάρος πριν γεννηθούν (κάτι που μπορεί να οδηγήσει σε επιπλοκές) και χαμηλά επίπεδα γλυκόζης. Οι κίνδυνοι αυτοί μπορούν να αντιμετωπιστούν μέσω του ελέγχου των επιπέδων γλυκόζης της εγκύου. Μετά την εγκυμοσύνη, τα παιδιά τους έχουν κάποιες πιθανότητες να εμφανίσουν παιδική παχυσαρκία και Διαβήτη Τύπου ΙΙ αργότερα στη ζωή τους.

1.1.3 Συμπτώματα του Σακχαρώδη Διαβήτη

Τα πρώτα συμπτώματα του διαβήτη είναι η αυξημένη όρεξη (πολυφαγία), η αυξημένη δίψα (πολυδιψία), η συχνή και σε μεγάλες ποσότητες ούρηση (πολυουρία), η απώλεια βάρους και η αίσθηση κόπωσης. Επειδή τα υψηλά επίπεδα γλυκόζης μπορούν να προκαλέσουν αλλαγές στο σχήμα του φακού του οφθαλμού, ένα ακόμη σύμπτωμα που μπορεί να παρατηρηθεί μερικές φορές είναι διαταραχή στην όραση. Ειδικά σε περιπτώσεις παρατήρησης γρήγορων μεταβολών στην όραση, θα πρέπει σίγουρα να ελέγχεται η περίπτωση Διαβήτη Τύπου Ι. Υπάρχουν επίσης περιπτώσεις ανθρώπων, συνήθως με Διαβήτη Τύπου Ι, που αρχικά εμφάνισαν διαβητική κετοξέωση, μία μεταβολική δυσλειτουργία, που χαρακτηρίζεται από όξινη μυρωδιά στην αναπνοή, δυσκολία στην αναπνοή, ναυτία, εμετό και στομαχόπονο. Η διαβητική κετοξέωση αποτελεί έκτακτη ιατρική κατάσταση.

Στο Διαβήτη Τύπου Ι τα συμπτώματα αναπτύσσονται συνήθως αρκετά γρήγορα, μέσα σε εβδομάδες ή μήνες. Στο Διαβήτη Τύπου ΙΙ αναπτύσσονται πολύ πιο αργά, υπάρχουν δε και περιπτώσεις κατά τις οποίες κανένα από τα παραπάνω συμπτώματα δεν έγινε αντιληπτό.

1.1.4 Επιπλοκές του Σακχαρώδη Διαβήτη

Ο Σακχαρώδης Διαβήτης συνδέεται με διάφορες επιπλοκές, κάποιες βραχυπρόθεσμες και κάποιες μακροπρόθεσμες.

1.1.4.1 Βραχυπρόθεσμες επιπλοκές

Οι βραχυπρόθεσμες επιπλοκές είναι τα επεισόδια υπεργλυκαιμίας και υπογλυκαιμίας. Υπεργλυκαιμία ονομάζεται η κατάσταση αυξημένης συγκέντρωση γλυκόζης στο αίμα, πέρα από τα φυσιολογικά όρια. Τα συμπτώματα της υπεργλυκαιμίας είναι δίψα, συχνοουρία, ατονία, αδιαθεσία, δύσπνοια, κοιλιακοί πόνοι και γενικευμένοι πόνοι. Αν δεν αντιμετωπιστεί εγκαίρως η υπεργλυκαιμία, μπορεί να εξελιχθεί σε διαβητικό κώμα, που συνεπάγεται απώλεια των αισθήσεων και αποτελούσε τη βασικότερη αιτία θανάτων σε διαβητικούς, πριν τη χρήση της ινσουλίνης στην αντιμετώπιση του διαβήτη.

Η υπογλυκαιμία, αντίθετα, χαρακτηρίζεται από την πτώση των επιπέδων γλυκόζης κάτω από το φυσιολογικά όρια. Συνήθως δεν είναι αποτέλεσμα του ίδιου του διαβήτη, αλλά της θεραπείας του με ινσουλίνη ή με υπογλυκαιμικούς παράγοντες. Η υπογλυκαιμία μπορεί να εμφανιστεί όταν ένα άτομο που λαμβάνει αντιδιαβητική φαρμακευτική αγωγή φάει πολύ λίγο, αν δεν φάει αρκετά έγκαιρα, αν λάβει αυξημένη δόση φαρμακευτικής αγωγής (π.χ. μεγαλύτερη ποσότητα ινσουλίνης από όση χρειάζεται) ή αν ασκηθεί υπερβολικά. Τα συμπτώματα είναι υπερβολική εφίδρωση, ζαλάδα, τρέμουλο, ωχρότητα και πονοκέφαλος. Μπορεί να εμφανιστεί αιφνίδια και αν δεν αντιμετωπιστεί έγκαιρα, μπορεί να οδηγήσει σε απώλεια των αισθήσεων ή ακόμη και σε σπασμούς.

1.1.4.2 Μακροπρόθεσμες Επιπλοκές

Οι συχνές διακυμάνσεις της συγκέντρωσης γλυκόζης στο αίμα πέρα των φυσιολογικών ορίων μπορεί να επηρεάσουν την ομαλή λειτουργία του οργανισμού σε βάθος χρόνου με πολλούς τρόπους. Τα υψηλά επίπεδα γλυκόζης στο αίμα προκαλούν αλλαγές στη σύσταση της μεμβράνης των αιμοφόρων αγγείων, παρεμποδίζοντας τη σωστή λειτουργία της. Επειδή η μεμβράνη των αιμοφόρων αγγείων εκτελεί το πολύ σημαντικό έργο της ανταλλαγής θρεπτικών ουσιών και οξυγόνου μεταξύ τριχοειδών και αγγείων, η διαταραχή αυτή προκαλεί μακροπρόθεσμα διάφορες επιπλοκές, τις κυριότερες από τις οποίες περιγράφουμε συνοπτικά παρακάτω:

- Καρδιαγγειακές παθήσεις (Cardiovascular Disease).

  Ο όρος καρδιαγγειακές παθήσεις περιλαμβάνει την καρδιοπάθεια, το εγκεφαλικό επεισόδιο και άλλες παθήσεις της καρδιάς και του κυκλοφορικού συστήματος, όπως η αθηροσκλήρυνση

(atherosclerosis) και η στένωση των αρτηριών. Διαβητικά άτομα, που δεν ακολουθούν έναν υγιεινό τρόπο ζωής έχουν μεγάλη πιθανότητα εναπόθεσης λιπαρών ουσιών στα τοιχώματα των αρτηριών τους, που περιορίζουν την ελεύθερη ροή του αίματος. Η κατάσταση αυτή ονομάζεται αθηροσκλήρυνση. Αν μία αρτηρία γίνει πολύ στενή, ή αποφραχθεί τελείως, κάποια μέρη του σώματος δε θα μπορούν να λάβουν το οξυγόνο και τα θρεπτικά συστατικά που χρειάζονται. Ανάλογα με το σε ποια αρτηρία συμβαίνει αυτή η διαταραχή προκαλούνται διάφορα προβλήματα. Αν αποφραχθεί μία αρτηρία που οδηγεί στην καρδιά, μπορεί να προκληθεί έμφραγμα του μυοκαρδίου (Myocardial Infarction - MI). Αν κάτι τέτοιο συμβεί σε μία αρτηρία που οδηγεί στον εγκέφαλο, μπορεί να προκληθεί εγκεφαλικό επεισόδιο (Cerebrovascular Accident – CVA), ενώ αν η στένωση ή η απόφραξη συμβεί σε κάποιο από τα άκρα, κάτι που είναι γνωστό ως περιφερική αγγειοπάθεια (Peripheral Vascular Disease - PVD), μπορεί να καταλήξει μέχρι και σε γάγγραινα.

Για την αποφυγή καρδιαγγειακών παθήσεων απαραίτητη είναι η απώλεια βάρους, η ρύθμιση των επιπέδων της γλυκόζης στα φυσιολογικά όρια, η ισορροπημένη διατροφή και η διατήρηση της πίεσης του αίματος και των επιπέδων χοληστερίνης στα φυσιολογικά όρια.

- Νευροπάθεια (Neuropathy).

Ο τρόπος με τον οποίο ο διαβήτης συνδέεται με τη νευροπάθεια δεν είναι πλήρως κατανοητός. Πιστεύεται, όμως, ότι τα υψηλά επίπεδα γλυκόζης προκαλούν χημικές μεταβολές στα νεύρα, οι οποίες μπορεί να βλάψουν την ικανότητά τους να μεταδίδουν σήματα. Επίσης, η μεγάλη συγκέντρωση γλυκόζης μπορεί να προκαλέσει αλλοιώσεις στα αγγεία του αίματος με τον τρόπο που έχουμε ήδη περιγράψει, επηρεάζοντας έτσι την ικανότητα μεταφοράς οξυγόνου, κάτι που έχει σαν αποτέλεσμα την ανεπαρκή οξυγόνωση των νευρικών κυττάρων. Και στην περίπτωση της νευροπάθειας η σωστή διατροφή, η διατήρηση της γλυκόζης σε φυσιολογικά επίπεδα και η σωματική άσκηση αποτελούν το κλειδί για τη μείωση του κινδύνου εμφάνισής της.

- Νεφροπάθεια (Nephropathy/Kidney Disease).

Η νεφροπάθεια μπορεί να συμβεί στον καθένα, αλλά είναι πιο συνηθισμένη σε διαβητικά και σε υπερτασικά άτομα. Η οφειλόμενη στο διαβήτη νεφροπάθεια εξελίσσεται αργά σε βάθος χρόνων και είναι πιο συχνή σε άτομα που έχουν διαβήτη για πάνω από 20 χρόνια. Η διαβητική βλάβη των νεφρών συνήθως οφείλεται σε αλλαγές στα μικρά αιμοφόρα αγγεία που οδηγούν στο σύστημα διήθησης (φιλτραρίσματος) των νεφρών ή σε μικρότερα αιμοφόρα αγγεία μέσα στο ίδιο το σύστημα. Τα νεφρά είναι το όργανο που φιλτράρει και καθαρίζει το αίμα από περιττές ουσίες παράγοντας τα ούρα, ρυθμίζει την ποσότητα υγρών και αλάτων στο σώμα και εκκρίνει διάφορες ορμόνες. Αν τα αιμοφόρα αγγεία δεν μεταφέρουν το απαραίτητο οξυγόνο και τα απαραίτητα θρεπτικά συστατικά, τα νεφρά δεν μπορούν να λειτουργήσουν αποδοτικά. Όταν τα νεφρά δυσλειτουργούν, οι ουσίες που θα έπρεπε να ανακυκλωθούν και να περάσουν

στο αίμα περνούν στα ούρα, ενώ ορισμένες άχρηστες ουσίες περνούν στο αίμα. Μία κατηγορία ουσιών που πρέπει να ανακυκλωθούν είναι οι πρωτεΐνες. Όταν αυτό δεν συμβαίνει και οι νεφροί επιτρέπουν την απέκκριση των πρωτεϊνών στα ούρα, αυτό είναι σημείο ότι οι νεφροί έχουν υποστεί βλάβη. Πολλά άτομα με διαβήτη που εμφανίζουν μικρές ποσότητες της πρωτεΐνης λευκωματίνης στα ούρα τους (μικρολευκωματινουρία) για 5 χρόνια, αναπτύσσουν μετά από 15 χρόνια διαβητική νεφροπάθεια. Γενικά, περίπου ένα στα τρία διαβητικά άτομα αναπτύσσει νεφροπάθεια, αν και όσο η αντιμετώπιση του διαβήτη βελτιώνεται, εμφανίζεται σε όλο και λιγότερα άτομα. Η νεφροπάθεια είναι πιο συχνή στους άντρες, ενώ η ανεπαρκής ρύθμιση των επιπέδων της γλυκόζης αυξάνει πολύ τον κίνδυνο εμφάνισής της και επιταχύνει, επιπλέον, και την εξέλιξή της από τη στιγμή που εμφανίζεται. Για το λόγο αυτό, η τήρηση της κατάλληλης φαρμακευτικής αγωγής και η υιοθέτηση υγιεινής διατροφής, είναι και πάλι ζωτικής σημασίας.

- Αμφιβληστροειδοπάθεια (Retinopathy).

Αφέθηκε σκόπιμα τελευταία, καθώς αποτελεί βασικό κομμάτι της παρούσας διπλωματικής.

Τα μάτια είναι πολύ ευαίσθητα στην επίδραση του διαβήτη. Ο διαβήτης μπορεί να προάγει την εμφάνιση καταρράκτη και αποτελεί παράγοντα κινδύνου για γλαύκωμα. Η πλέον όμως σημαντική επίδραση του διαβήτη στον οφθαλμό είναι η διαβητική αμφιβληστροειδοπάθεια (Diabetic Retinopathy). Η διαβητική αμφιβληστροειδοπάθεια είναι το αποτέλεσμα βλάβης του αμφιβληστροειδή χιτώνα του οφθαλμού. Ο αμφιβληστροειδής είναι ο φωτοευαίσθητος χιτώνας που καλύπτει την εσωτερική επιφάνεια του οφθαλμού. Προσλαμβάνει τις εστιασμένες ακτίνες του φωτός και μετατρέπει την οπτική εικόνα σε νευρικό ερέθισμα, το οποίο ο εγκέφαλος το ερμηνεύει ως όραση. Υπάρχουν τρεις κατηγορίες διαβητικής αμφιβληστροειδοπάθειας: η μη παραγωγική διαβητική αμφιβληστροειδοπάθεια, η παραγωγική διαβητική αμφιβληστροειδοπάθεια και το διαβητικό οίδημα της ωχράς κηλίδας.

Στη μη παραγωγική αμφιβληστροειδοπάθεια (Nonproliferative Diabetic Retinopathy), τα υψηλά επίπεδα της γλυκόζης στο αίμα επηρεάζουν τα μικρά αιμοφόρα αγγεία του αμφιβληστροειδή και του προκαλούν διάφορες δομικές και λειτουργικές αλλαγές. Οι μεταβολές στα τοιχώματα των αιμοφόρων αγγείων επιτρέπουν σε υγρά να διαρρέουν στους ιστούς του αμφιβληστροειδή και έχουν σαν αποτέλεσμα λιγότερο αίμα πλούσιο σε οξυγόνο να κυκλοφορεί στον αμφιβληστροειδή, δύο παράγοντες που του προκαλούν σημαντικές βλάβες.

Η παραγωγική διαβητική αμφιβληστροειδοπάθεια (Proliferative Diabetic Retinopathy - PDR) είναι λίγο πιο περίπλοκη. Σε αυτήν, ο αμφιβληστροειδής χιτώνας για να αντιμετωπίσει την έλλειψη οξυγόνου, απελευθερώνει χημικούς μεσολαβητές που διεγείρουν το σχηματισμό νέων αιμοφόρων αγγείων για να φέρουν περισσότερο οξυγόνο. Αυτό έχει σαν αποτέλεσμα τη δημιουργία νέων εύθραυστων αγγείων, από τα οποία διαρρέουν υγρά, όπως αίμα και παράγωγα αίματος, στον αμφιβληστροειδή. Χωρίς ιατρική αντιμετώπιση, είναι επίσης δυνατόν να

ραγίσουν, οπότε και πληρείται ο οφθαλμός με αίμα, προκαλώντας σημαντική απώλεια όρασης. Επιπλέον, ο ουλώδης ιστός που συνοδεύει το σχηματισμό των νέων αγγείων μπορεί να ασκήσει έλξη στον αμφιβληστροειδή, προκαλώντας αλλοίωση της όρασης, και τελικά να οδηγήσει σε αποκόλληση αυτού από τον ιστό που τον στηρίζει. Η αποκόλληση του αμφιβληστροειδή αποτελεί σημαντικό κίνδυνο για την όραση.

Η διαρροή υγρών από παθολογικά αγγεία στον αμφιβληστροειδή μπορεί να προκαλέσει μια πολύ επικίνδυνη συσσώρευση υγρού στην ωχρά κηλίδα, στην κεντρική, δηλαδή, περιοχή του αμφιβληστροειδή, που είναι υπεύθυνη για τη σαφή και ευκρινή απεικόνιση των λεπτομερειών. Αυτή η συσσώρευση υγρού ονομάζεται οίδημα της ωχράς κηλίδας (macular edema).

Η διαβητική αμφιβληστροειδοπάθεια συνήθως δεν έχει συμπτώματα στα αρχικά στάδια εμφάνισής της. Είναι δυνατόν να υπάρχει σοβαρή, απειλητική για την όραση διαβητική αμφιβληστροειδοπάθεια, χωρίς καμιά μεταβολή στην όραση. Όταν εμφανίζονται συμπτώματα, αυτά μπορεί να περιλαμβάνουν μαύρα ή κόκκινα "άστρα" ή στίγματα στο οπτικό πεδίο. Το οίδημα της ωχράς κηλίδας μπορεί να προκαλέσει θόλωση ή παραμόρφωση της όρασης, και οι ευθείες γραμμές να εμφανίζονται ρυτιδωμένες. Το οίδημα της ωχράς κηλίδας μπορεί επίσης να αλλοιώσει την αντίληψη των χρωμάτων. Πολύ σημαντική είναι και πάλι η διατήρηση της συγκέντρωσης της γλυκόζης σε φυσιολογικά επίπεδα, η κατάλληλη φαρμακευτική αντιμετώπιση του διαβήτη, καθώς επίσης και η διατήρηση της πίεσης του αίματος σε φυσιολογικά επίπεδα, αλλά η καλύτερη προστασία κατά της διαβητικής αμφιβληστροπάθειας είναι οι τακτικές οφθαλμολογικές εξετάσεις.

Οι ασθενείς με διαβήτη τύπου Ι γενικά δεν εμφανίζουν αμφιβληστροειδοπάθεια πριν την εφηβεία, αλλά μετά από 15 χρόνια με διαβήτη τύπου Ι, υπάρχει υψηλότερη πιθανότητα εμφάνισης αμφιβληστροειδοπάθειας. Η απειλητική για την όραση παραγωγική αμφιβληστροειδοπάθεια υπάρχει σε 25% όσων έχουν διαβήτη τύπου Ι για περισσότερα από 15 χρόνια.

Οι ασθενείς με διαβήτη τύπου ΙΙ είναι πιο πιθανόν να έχουν αμφιβληστροειδοπάθεια τη στιγμή της διάγνωσης του διαβήτη ή σύντομα μετά από αυτή. Για το λόγο αυτό πρέπει να υποβληθούν σε οφθαλμολογική εξέταση όσο δυνατόν το συντομότερο αφού πληροφορηθούν ότι έχουν διαβήτη.

Δεν υπάρχει εγγυημένος τρόπος να προληφθεί η διαβητική αμφιβληστροειδοπάθεια ή να θεραπευθεί μετά την εγκατάστασή της. Ευτυχώς, τα τελευταία 25 χρόνια, έχουν εμφανιστεί αρκετές θεραπευτικές επιλογές που μπορούν να ελαττώσουν σημαντικά τον κίνδυνο της απώλειας της όρασης από τη διαβητική αμφιβληστροειδοπάθεια. Μία από αυτές είναι η θεραπεία με λέιζερ, η οποία είναι αρκετά αποτελεσματική αν το πρόβλημα διαγνωστεί εγκαίρως, ανώδυνη και στο 80% των περιπτώσεων αποτρέπει περαιτέρω απώλεια της όρασης.

1.1.5 Στατιστικά στοιχεία

Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (World Health Organization - WHO), το 2000 τα άτομα με Σακχαρώδη Διαβήτη ανέρχονταν σε τουλάχιστον 171 εκατομμύρια παγκοσμίως, ή αλλιώς το 2,8% του πληθυσμού. Εκτιμάται ότι μέχρι το 2030 ο αριθμός αυτός θα έχει σχεδόν διπλασιαστεί. Μόνο στην Ευρώπη το 9% του πληθυσμού πάσχει από διαβήτη και το 2025 αναμένεται ο αριθμός των ασθενών να φθάσει τα 40 εκατομμύρια. Στην Ελλάδα εκτιμάται ότι πάσχει το 5.9% του γενικού πληθυσμού.

Ο Σακχαρώδης Διαβήτης παρατηρείται πολύ πιο συχνά στις ανεπτυγμένες χώρες, ιδιαίτερα ο Διαβήτης Τύπου ΙΙ. Αν και ο μηχανισμός εκδήλωσης Διαβήτη δεν έχει πλήρως εξακριβωθεί, λόγω αυτών των παρατηρήσεων φαίνεται να είναι άμεσα συνδεδεμένος με το "Δυτικό τρόπο ζωής", την αστικοποίηση, τη μη ισορροπημένη και ανθυγιεινή διατροφή, την καθιστική ζωή και το στρες. Για το λόγο αυτό του έχει δοθεί και η ονομασία "ασθένεια της ευημερίας". Αποτελεί μία από τις κύριες αιτίες θανάτου παγκοσμίως. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας ο αριθμός των θανάτων που οφείλονταν στο διαβήτη το 2000 ανερχόταν στα 2,9 εκατομμύρια παγκοσμίως.

Η κατάσταση που συνθέτουν τα παραπάνω στοιχεία γίνεται ακόμα δυσχερέστερη με την ελλιπή ενημέρωση μεταξύ των ασθενών. Πρόσφατη μελέτη που διεξήχθη για λογαριασμό της Παγκόσμιας Ομοσπονδίας κατά του Διαβήτη σε πέντε μεγάλες ευρωπαϊκές χώρες και στις ΗΠΑ ανέδειξε αρκετά ενδιαφέροντα στοιχεία όσον αφορά στις αντιλήψεις των ίδιων των πασχόντων για τους κινδύνους που διατρέχουν από τη νόσο. Πιο συγκεκριμένα, και ενώ υπολογίζεται ότι το 74% αυτών θα αναπτύξει μικροαγγειοπάθεια, το 60% περίπου των ερωτηθέντων απάντησε ότι δεν τους ανησυχεί το ενδεχόμενο πιθανής τύφλωσης. Μόλις το 30% ανέφερε ότι έχει πρόβλημα με τη ρύθμιση των επιπέδων του σακχάρου στο αίμα, από αυτούς όμως οι επτά στους δέκα δεν θυμόντουσαν τα αποτελέσματα της τελευταίας μέτρησης γλυκόζης τους. Τέλος, ενώ το 70% ζήτησε πιο αποτελεσματικές θεραπείες για την αποφυγή των επιπλοκών από τη νόσο, οι τέσσερις στους δέκα πιστεύουν ότι οι τελευταίες θα συμβούν ό,τι θεραπευτικά μέτρα κι αν ληφθούν.

1.2 Αντικείμενο της διπλωματικής

Από όσα περιγράφηκαν στην προηγούμενη ενότητα, φαίνεται ότι ο μη πλήρως σαφής μηχανισμός εκδήλωσης του ίδιου του Διαβήτη και της διαβητικής αμφιβληστροειδοπάθειας, οι μη εγγυημένες μέθοδοι πρόληψής της, οι σοβαρές επιπτώσεις της στο άτομο και η ελλιπής ενημέρωση των ασθενών συνθέτουν ένα σκηνικό με προβλήματα που χρήζουν αντιμετώπισης. Γίνεται επίσης φανερή η ανάγκη περαιτέρω διερεύνησης των παραγόντων που συνδέονται με την εμφάνιση αμφιβληστροειδοπάθειας και η επιλογή της κατάλληλης τακτικής όσον αφορά στη διαχείριση του διαβήτη προς αποφυγή της επιπλοκής.

Στόχος της παρούσας διπλωματικής ήταν η ανάπτυξη μίας μεθοδολογίας προς αυτήν την κατεύθυνση. Η μεθοδολογία χωρίζεται πρακτικά σε δύο επιμέρους στάδια. Το πρώτο στάδιο αφορά στην εύρεση των παραγόντων που συνδέονται με την εμφάνιση της διαβητικής αμφιβληστροειδοπάθειας σε ασθενείς με διαβήτη Τύπου ΙΙ. Η διαδικασία αυτή βασίστηκε σε πραγματικά δεδομένα ασθενών και εφήρμοσε "έξυπνες μεθόδους" για την ανακάλυψη συσχετίσεων μεταξύ χαρακτηριστικών του ασθενή (π.χ. το δείκτη μάζας σώματος, το αν καπνίζει, διάφορα αποτελέσματα βιοχημικών εξετάσεων, κ.α.) και την εμφάνιση αμφιβληστροειδοπάθειας. Οι μέθοδοι που χρησιμοποιήθηκαν περιλαμβάνουν την υλοποίηση ενός Γενετικού Αλγορίθμου (Genetic Algorithm), ο οποίος κάνει χρήση της Αμοιβαίας Πληροφορίας (Mutual Information), μίας έννοιας από το πεδίο της Θεωρίας Πληροφορίας (Information Theory), καθώς και την υλοποίηση Τεχνητών Νευρωνικών Δικτύων (Artificial Neural Networks). Με τον τρόπο αυτό μελετάται το ζήτημα των παραγόντων που παίζουν σημαντικό ρόλο στην εκδήλωση της διαβητικής αμφιβληστροειδοπάθειας με τη χρήση πιο σύγχρονων μεθόδων, κάτι που μπορεί να θεωρηθεί ως ένα είδος επιδημιολογικής μελέτης και να αποτελέσει ένα βήμα προς την καλύτερη πρόληψη της ασθένειας.

Στο δεύτερο στάδιο της μεθοδολογίας μας υλοποιήθηκε ένα μοντέλο εκτίμησης της πιθανότητας ενός ασθενή με διαβήτη τύπου ΙΙ να εμφανίσει μακροπρόθεσμα αμφιβληστροειδοπάθεια, βάσει του τρέχοντος "προφίλ" του. Το πρόβλημα δηλαδή που μελετάται είναι: Δεδομένου ορισμένων μετρήσεων σε διάφορα χαρακτηριστικά, ποια η πιθανότητα το συγκεκριμένο άτομο με διαβήτη να εμφανίσει μελλοντικά αμφιβληστροειδοπάθεια. Όπως γίνεται εύκολα αντιληπτό το πρόβλημα αυτό ανάγεται σε πρόβλημα ταξινόμησης (classification). Το υπολογιστικό εργαλείο που χρησιμοποιήσαμε για τον σκοπό αυτό είναι ένα Τεχνητό Νευρωνικό Δίκτυο. Βασιζόμενοι και πάλι σε πραγματικά δεδομένα ασθενών, αναπτύξαμε ένα μοντέλο που να συγκεντρώνει τη γνώση που προκύπτει μέσα από πολλές περιπτώσεις ασθενών και να μπορεί να προβλέπει μακροπρόθεσμα την πιθανότητα εμφάνισης της επιπλοκής. Γίνεται με τον τρόπο αυτό μία απόπειρα "αποτύπωσης" της σχέσης μεταξύ διάφορων χαρακτηριστικών του ασθενή και του ενδεχομένου να εμφανίσει μετά από καιρό αμφιβληστροειδοπάθεια, ένα πρόβλημα του οποίου, όπως εξηγήσαμε, ο βασικός μηχανισμός δεν είναι πλήρως κατανοητός. Επίσης, η ανάπτυξη ενός τέτοιου μοντέλου μπορεί να συνεισφέρει στην ενημέρωση και στην αφύπνιση του ατόμου σχετικά με τον κίνδυνο που ενδέχεται να διατρέχει, με στόχο κάποια αλλαγή στον τρόπο ζωής του, και να συμβάλει στην πιο αποτελεσματική διαχείριση της νόσου (disease management) προς αποφυγή εμφάνισης της επιπλοκής.

2

Θεωρητικό Υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζεται το θεωρητικό υπόβαθρο, πάνω στο οποίο στηρίζεται η παρούσα διπλωματική και η κατανόηση του οποίου κρίνεται απαραίτητη για τον αναγνώστη. Αρχικά αναφερόμαστε συνοπτικά στην έννοια της ταξινόμησης (Classification). Στην ενότητα 2.2 περιγράφεται η επιλογή χαρακτηριστικών (Feature Selection) και οι διάφορες μέθοδοι που έχουν αναπτυχθεί για την υλοποίησή της. Στις ενότητες 2.3 και 2.4 παραθέτουμε το θεωρητικό υπόβαθρο για τους Γενετικούς Αλγόριθμους (Genetic Algorithms) και τα Νευρωνικά Δίκτυα (Neural Networks) αντίστοιχα, καθώς αποτελούν τα "εργαλεία" για τη μέθοδο που αναπτύσσεται στην παρούσα διπλωματική. Τέλος, αναφερόμαστε συνοπτικά στην Αμοιβαία Πληροφορία (Mutual Information), καθώς θα χρησιμοποιηθεί σε ένα από τα στάδια της υλοποίησής μας.

2.1 Ταξινόμηση (Classification)

Η έννοια της ταξινόμησης (classification) αναφέρεται σε μια διαδικασία αντιστοίχισης κάποιων δεδομένων εισόδου σε μία από ένα σύνολο κατηγοριών. Στην καθημερινή μας ζωή η ταξινόμηση είναι μια αρκετά συνηθισμένη εργασία και πολλές δραστηριότητές μας την περιλαμβάνουν, έστω εν μέρει. Μερικά παραδείγματα ταξινόμησης είναι:

- Η κατάταξη ενός ηλεκτρονικού μηνύματος (e-mail) στην ενοχλητική αλληλογραφία (spam) ή όχι.
- Η αποτίμηση ιατρικής διάγνωσης βάσει διαφόρων χαρακτηριστικών ενός ατόμου (αποτελέσματα σε διάφορες εξετάσεις, παρουσία ή απουσία συμπτωμάτων, κλπ.)

- Διαπίστωση απάτης στην χρήση πιστωτικών καρτών.
- Μετατροπή χειρόγραφων σε κείμενα ηλεκτρονικής μορφής.
- Αναγνώριση προσώπων σε εικόνες και αναγνώριση φωνής.
- Οδήγηση αυτόνομου οχήματος (στρίψε δεξιά, αριστερά, προχώρα ευθεία), κ.α.

Ένας αλγόριθμος που υλοποιεί ταξινόμηση ονομάζεται ταξινομητής (classifier). Στη διπλωματική αυτή αναφερόμαστε στην ταξινόμηση ως κομμάτι της μηχανικής μάθησης (machine learning), και πιο συγκεκριμένα της επιβλεπόμενης μηχανικής μάθησης (supervised machine learning).

Ο άνθρωπος προσπαθεί να κατανοήσει το περιβάλλον του παρατηρώντας το και δημιουργώντας μια πιο απλή εκδοχή του. Επιπλέον, έχει τη δυνατότητα να οργανώνει και να συσχετίζει τις διάφορες εμπειρίες του δημιουργώντας νέες δομές. Η πραγματοποίηση των παραπάνω λειτουργιών από ένα υπολογιστικό σύστημα ονομάζεται μηχανική μάθηση [BKB+06].

Η επιβλεπόμενη μάθηση αφορά σε μια διαδικασία μάθησης μέσω ήδη υπάρχουσας γνώσης, με την χρήση παραδειγμάτων. Διαθέτουμε, δηλαδή κάποια δεδομένα, τα δεδομένα εκπαίδευσης (training data), για τα οποία ξέρουμε ποιά είναι η έξοδος και ο ταξινομητής μας καλείται να "μάθει" μέσω αυτών να ταξινομεί σωστά κάθε μελλοντικά δεδομένα εισόδου. Τα "άτομα" των δεδομένων εισόδου μας ονομάζονται στιγμιότυπα (instances). Κάθε στιγμιότυπο περιγράφεται από ένα σετ χαρακτηριστικών (features/attributes). Χαρακτηριστικό θεωρείται μία "συλλογή" από ανεξάρτητες και αμοιβαίως αποκλειόμενες τιμές [JB04]. Τα χαρακτηριστικά μπορεί να έχουν ονομαστικές τιμές (nominal features), όπως για παράδειγμα η ομάδα αίματος που μπορεί να πάρει τις τιμές A, B, AB και O, συνεχείς (continuous), όπως η μέτρηση της πίεσης και δυαδικές (binary), όπως η παρουσία (1) ή η απουσία (0) ενός συμπτώματος. Οι κατηγορίες στις οποίες κατατάσσονται τα στιγμιότυπα ονομάζονται κλάσεις (classes).

Στον παρακάτω πίνακα φαίνεται ένα μικρό παράδειγμα ταξινόμησης. Τα στιγμιότυπα είναι οι ημέρες, τα χαρακτηριστικά είναι ο καιρός, η θερμοκρασία, η υγρασία και ο αέρας και η κλάση είναι το αν η συγκεκριμένη μέρα προσφέρεται για να παίξει κανείς τέννις [E09].

Πίνακας 2.1 : Παράδειγμα Ταξινόμησης.

Ημέρα	Καιρός	Θερμοκρασία	Υγρασία	Αέρας	Τέννις
1	Ηλιοφάνεια	Υψηλή	Υψηλή	Ασθενής	Όχι
2	Ηλιοφάνεια	Υψηλή	Υψηλή	Ισχυρός	Όχι
3	Νεφελώδης	Υψηλή	Υψηλή	Ασθενής	Ναι
4	Βροχή	Μέτρια	Υψηλή	Ασθενής	Ναι

Οι πιο ευρέως γνωστοί ταξινομητές είναι: τα Δέντρα Ταξινόμησης ή Αποφάσεων (Classification or Decision trees), η Μάθηση Κανόνων Ταξινόμησης (Rule Learning), ο Απλός Ταξινομητής Bayes (Naïve-Bayes Classifier), ο Ταξινομητής των k-Κοντινότερων γειτόνων (k-Nearest Neighbor

Classifier), οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVMs), τα Δίκτυα Συναρτήσεων Βάσης Ακτινικού Τύπου (Radial Basis Function Networks – RBF Networks) και τα Νευρωνικά Δίκτυα (Neural Networks - NNs).

Η απόδοση του ταξινομητή εξαρτάται άμεσα από τα δεδομένα εκπαίδευσης. Αφού ολοκληρωθεί η διαδικασία της εκπαίδευσης, ακολουθεί η αξιολόγηση του ταξινομητή. Η αξιολόγηση πραγματοποιείται συνήθως με την χρήση επιπλέον δεδομένων, των δεδομένων έλεγχου (testing data), για τα οποία και πάλι γνωρίζουμε την έξοδο. Στο σημείο αυτό δεν κάνουμε καμία αλλαγή στον ταξινομητή (δεν συνεχίζουμε, δηλαδή να του "μαθαίνουμε" τα δεδομένα), αλλά ελέγχουμε κατά πόσο μπορεί να ταξινομήσει σωστά τα δεδομένα ελέγχου. Στη συνέχεια υπολογίζονται διάφορα μέτρα εκτίμησης της ποιότητας του ταξινομητή και χαράσσονται καμπύλες που αναπαριστούν γραφικά τα ποσοστά των σωστά και των λανθασμένα ταξινομημένων στιγμιοτύπων. Στη διαδικασία αξιολόγησης του ταξινομητή θα αναφερθούμε αναλυτικά στα Κεφάλαια 5 και 6.

2.2 Επιλογή χαρακτηριστικών (Feature Selection)

Στα περισσότερα προβλήματα ταξινόμησης, αναγνώρισης προτύπων, πρόβλεψης και ανακάλυψης γνώσης γενικότερα, η επιλογή χαρακτηριστικών (feature selection) αποτελεί ένα πολύ σημαντικό κομμάτι για την ελαχιστοποίηση της πιθανότητας λάθους και την αύξηση της ακρίβειας. Η επιλογή χαρακτηριστικών αφορά στην εύρεση του συνόλου των χαρακτηριστικών (features) που οδηγούν στο καλύτερο και ακριβέστερο μοντέλο πρόβλεψης. Ένας ακριβέστερος ορισμός είναι ο εξής [PLD05]: Δεδομένου N δειγμάτων εισόδου, ένα σετ M χαρακτηριστικών $X=\{x_i \mid i=1\ldots M\}$ και της μεταβλητής ταξινόμησης (target variable) c, το πρόβλημα της επιλογής χαρακτηριστικών έγκειται στη εύρεση ενός υποσυνόλου $Y=\{x_{i1},x_{i2},\ldots,x_{im}\}$ με $m<M$ χαρακτηριστικά που να χαρακτηρίζει "βέλτιστα" τη c, ή αλλιώς να μεγιστοποιεί την πιθανότητα σωστής ταξινόμησης. Ας σταθούμε για λίγο στην έννοια "βέλτιστα". Επειδή ο αριθμός όλων των δυνατών υποσυνόλων είναι 2^M και το πλήθος όλων των υποσυνόλων με μέχρι m στοιχεία είναι $\sum_{i=1}^{m}\binom{M}{i}$, η εξαντλητική αναζήτηση (exhaustive search) του βέλτιστου υποσυνόλου είναι υπολογιστικά δυσχερής και πρακτικά ανέφικτη. Για παράδειγμα, αν θεωρήσουμε ότι η αξιολόγηση ενός υποψήφιου υποσυνόλου διαρκεί κατά μέσο όρο 1/100 του δευτερολέπτου, η εύρεση του βέλτιστου υποσυνόλου M=40 χαρακτηριστικών θα απαιτούσε 349 χρόνια! Γι' αυτό το λόγο έχουν αναπτυχθεί διάφοροι αλγόριθμοι για την υλοποίηση της επιλογής χαρακτηριστικών, που ενδεχομένως να μην καταλήγουν στην εύρεση του βέλτιστου υποσυνόλου, αλλά ενός υπο-βέλτιστου που μας δίνει πολύ ικανοποιητικά αποτελέσματα. Τη λογική των διαφόρων αλγορίθμων θα αναπτύξουμε σε επόμενη υποενότητα, αφού πρώτα αναλύσουμε τη συνεισφορά της επιλογής χαρακτηριστικών στις εφαρμογές μηχανικής μάθησης και ανακάλυψης γνώσης.

2.2.1 Συνεισφορά της επιλογής χαρακτηριστικών

Στην υποενότητα αυτή παρουσιάζονται οι παράγοντες που επηρεάζει η επιλογή χαρακτηριστικών και οι λόγοι για τους οποίους αποτελεί ένα σημαντικό στάδιο σε κάθε εφαρμογή ταξινόμησης [YH97], [GE03].

Πρώτα απ' όλα τίθεται το ερώτημα: είναι όλα τα χαρακτηριστικά εξίσου σημαντικά και αναγκαία για την εργασία που εκτελούμε; Το ερώτημα αυτό συνδέεται με την "κατάρα της διαστατικότητας" (curse of dimensionality), σύμφωνα με την οποία, η αύξηση του αριθμού των χαρακτηριστικών πρέπει να συνοδεύεται από εκθετική αύξηση του αριθμού των στιγμιότυπων (instances), δηλαδή των δεδομένων μας. Έτσι, στις περιπτώσεις που έχουμε δεκάδες, ή χιλιάδες χαρακτηριστικά πρέπει να διαθέτουμε και ένα πολύ μεγάλο αριθμό δειγμάτων. Κάτι τέτοιο δεν είναι πάντα εφικτό. Το γεγονός αυτό και η χρήση πολλών χαρακτηριστικών δυσχεραίνει την υπολογιστική διαδικασία και αυξάνει τις απαιτήσεις για μνήμη και τον χρόνο επεξεργασίας και εκπαίδευσης του μοντέλου πρόβλεψης. Επιλέγοντας μόνο τα πιο σημαντικά χαρακτηριστικά τα παραπάνω προβλήματα μπορούν να αρθούν.

Ένας ακόμη λόγος για τον οποίο είναι σημαντική η επιλογή χαρακτηριστικών είναι ότι συχνά κάποια χαρακτηριστικά είναι περιττά (redundant), με την έννοια ότι δεν προσφέρουν πολλά στον ταξινομητή (classifier) ή προσδιορίζονται πλήρως από κάποια άλλα χαρακτηριστικά, ή ακόμα και άσχετα (irrelevant) ή "θορυβώδη" (noise features). Η συμπερίληψη τέτοιων χαρακτηριστικών στην εκπαίδευση του μοντέλου καθιστά πιο δύσκολη την κατανόηση των εσωτερικών (underlying) συσχετίσεων μεταξύ των δεδομένων και τη σύλληψη της απαραίτητης πληροφορίας που απαιτείται για τη σωστή ταξινόμηση και, συνεπώς, μειώνει την ακρίβεια της ταξινόμησης. Η επιλογή των σημαντικών χαρακτηριστικών μπορεί με τον τρόπο αυτό να βελτιώσει την απόδοση του ταξινομητή και την ακρίβεια της πρόβλεψης. Κάτι τέτοιο είναι συνώνυμο με την αύξηση της ικανότητας γενίκευσης (generalization capability) του μοντέλου, την ικανότητα του δηλαδή να αποδίδει σωστά όχι μόνο στα δεδομένα εκπαίδευσης, αλλά και σε νέα δεδομένα.

Τέλος, ο περιορισμός των χαρακτηριστικών στα πλέον σημαντικά ενδέχεται να συνδέεται με τη μείωση του κόστους της πραγματοποίησης της ταξινόμησης. Σε πολλές πρακτικές εφαρμογές, όπως η ιατρική διάγνωση, στα χαρακτηριστικά συμπεριλαμβάνονται διάφορα συμπτώματα, καθώς επίσης και αποτελέσματα από διαγνωστικές δοκιμασίες. Κάθε διαγνωστική δοκιμασία σχετίζεται με διαφορετικό κόστος και ρίσκο. Για παράδειγμα, μία επεμβατική διερευνητική εγχείρηση μπορεί να είναι πιο δαπανηρή ή επικίνδυνη από ότι μία εξέταση αίματος. Με την επιλογή χαρακτηριστικών μπορούν να εξαλειφθούν τέτοιου είδους χαρακτηριστικά, με την προϋπόθεση βεβαίως ότι αυτό δεν γίνεται εις βάρος της απόδοσης του μοντέλου.

2.2.2 Αλγόριθμοι - Τεχνικές επιλογής χαρακτηριστικών

Όλοι οι αλγόριθμοι επιλογής χαρακτηριστικών (feature selection algorithms) αποτελούνται ουσιαστικά από:

- Έναν αλγόριθμο αναζήτησης (search algorithm) για την επιλογή των υποψήφιων υποσυνόλων, και
- Ένα κριτήριο αξιολόγησης των υποσυνόλων.

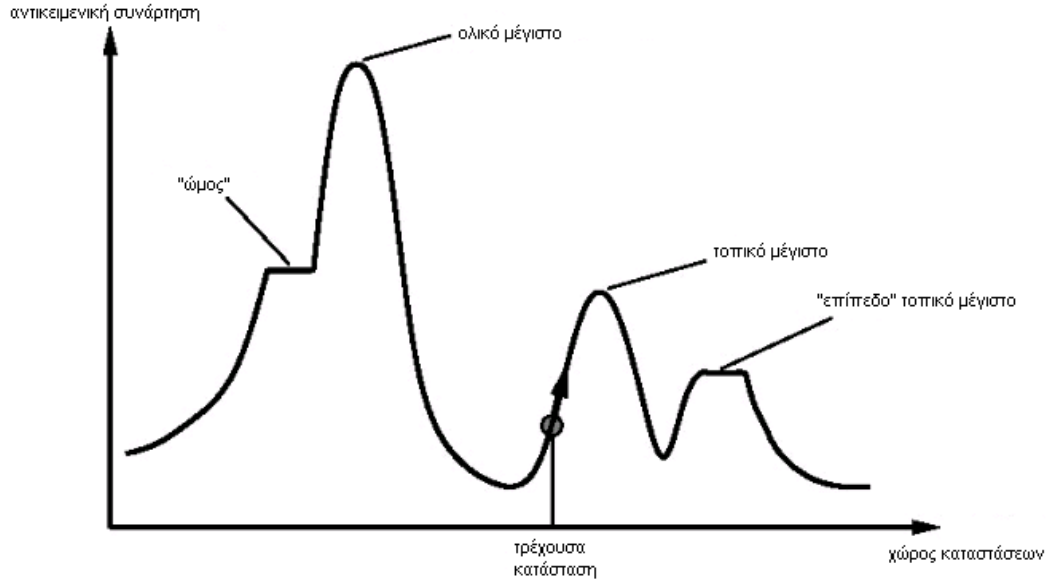
2.2.2.1 Αλγόριθμοι αναζήτησης (search algorithms)

Μερικοί αλγόριθμοι αναζήτησης είναι οι:

- Επιλογή από κάτω προς τα πάνω (Sequential Forward Selection - SFS): Ξεκινάει με ένα άδειο σετ και προσθέτει σειριακά τα χαρακτηριστικά, εξετάζοντας κάθε φορά με την προσθήκη ποιανού μεγιστοποιείται το κριτήριο αξιολόγησης.
- Επιλογή από πάνω προς τα κάτω (Sequential Backward Selection/Elimination - SBS/SBE): Ακολουθεί την ίδια λογική με τον προηγούμενο αλγόριθμο, μόνο που ξεκινάει από ένα πλήρες σετ των δεδομένων και διαδοχικά αφαιρεί χαρακτηριστικά.
- Plus-L-minus-R Αναζήτηση: Αποτελεί μια γενίκευση των δύο προηγούμενων μεθόδων. Πραγματοποιεί L βήματα προς τα πάνω και R βήματα προς τα κάτω.
- Διακλάδωση και Οριοθέτηση (Branch-and-bound): Ο αλγόριθμος αυτός δεν εξετάζει καθόλου τις λύσεις που δεν αναμένεται να δώσουν καλύτερο αποτέλεσμα.
- Ακτινωτή Αναζήτηση (Beam search): Ο αλγόριθμος αυτός σε κάθε επανάληψη αξιολογεί όλες τις πιθανές καταστάσεις που απορρέουν από την προσθήκη ενός χαρακτηριστικού στο υποσύνολο και κρατάει μόνο τις καλύτερες από αυτές στο μέτωπο αναζήτησης.
- Πρώτα στο Καλύτερο Αναζήτηση (Best First): Αντίθετα με τους δύο προηγούμενους αλγόριθμους, ο αλγόριθμος αυτός κρατάει όλες τις καταστάσεις στο μέτωπο αναζήτησης και μπορεί να επιστρέψει σε κάποια από αυτές, αν το μονοπάτι που ακολουθεί αποδειχθεί χειρότερο.
- Δικατευθυντήρια Αναζήτηση (Bidirectional search): Παράλληλη υλοποίηση των SFS και SBS.
- Random Generation plus Sequential Selection: Αποτελεί την εισαγωγή του "τυχαίου" σε SFS και SBS για την αποφυγή "τοπικών ελαχίστων".
- Προσομοιωμένη Ανόπτηση (Simulated Annealing - SA): Δίνει μία πιθανότητα μετάβασης σε χειρότερες καταστάσεις, αφήνοντας έτσι ένα ενδεχόμενο να ξεφύγει η αναζήτηση από "τοπικά ελάχιστα". Η ιδέα της μεθόδου προέρχεται από τις φυσικές διαδικασίες δημιουργίας κρυσταλλικών δομών στη φύση.

- Γενετικοί Αλγόριθμοι (Genetic Algorithms - GAs): Σε αυτήν την κατηγορία θα αναφερθούμε εκτενώς σε επόμενο κεφάλαιο, καθώς αποτελεί μέρος της μεθοδολογίας που ακολουθήθηκε σε αυτή τη διπλωματική.

Στο σημείο αυτό πρέπει να ξεκαθαρίσουμε ότι στην επιλογή χαρακτηριστικών και γενικά στα προβλήματα βελτιστοποίησης, κάποιες φορές θέλουμε να μεγιστοποιήσουμε κάποιο κριτήριο ή κάποια συνάρτηση και άλλες να τα ελαχιστοποιήσουμε, ανάλογα με τη φύση του προβλήματος. Οι δύο αυτές απαιτήσεις αντιμετωπίζονται με τον ίδιο τρόπο. Άλλωστε, η μεγιστοποίηση μιας συνάρτησης f είναι ισοδύναμη με την ελαχιστοποίηση της –f. Στη συνέχεια, ό,τι αναφέρουμε για τον ένα όρο, θα ισχύει και για τον άλλο. Το ίδιο θα τηρήσουμε και για τους όρους "τοπικό μέγιστο" και "τοπικό ελάχιστο". Η έννοια "τοπικό ακρότατο", γενικότερα, αναφέρεται σε μία "καλή" λύση, που όμως είναι απλά καλύτερη από κάποιες άλλες και όχι η καλύτερη από όλες τις δυνατές. Το φαινόμενο αυτό αναπαρίσταται στο παρακάτω σχήμα:



Σχήμα 2.1 : Το σύνολο τιμών μίας αντικειμενικής συνάρτησης,

Οι παραπάνω αλγόριθμοι υπάγονται στις παρακάτω γενικότερες κατηγορίες:

- Σειριακοί αλγόριθμοι, οι οποίοι προσθέτουν ή αφαιρούν χαρακτηριστικά σειριακά. Στην κατηγορία αυτή ανήκουν οι SFS, SBE, Plus-L-minus-R και Bidirectional Search.
- Ευριστικοί αλγόριθμοι, οι οποίοι αξιολογούν τις καταστάσεις βάσει κάποιας πληροφορίας και καθοδηγούν την αναζήτηση "κλαδεύοντας" καταστάσεις, οι οποίες εκτιμάται ότι δε θα οδηγήσουν σε καλή λύση. Στην κατηγορία αυτή ανήκουν ουσιαστικά όλοι οι παραπάνω αλγόριθμοι, αφού δεν εξετάζουν εξαντλητικά όλες τις υποψήφιες λύσεις.
- Τυχαίοι αλγόριθμοι, οι οποίοι επιστρατεύουν τον παράγοντα "τύχη" στο ψάξιμό τους, για να αποφύγουν την παγίδευση σε τοπικά ελάχιστα. Παραδείγματα αυτής της κατηγορίας είναι οι

Random Generation plus Sequential Selection, Προσομοιωμένη Ανόπτηση και οι Γενετικοί Αλγόριθμοι.

Κάθε αλγόριθμος αναζήτησης έχει τα πλεονεκτήματα και τα μειονεκτήματά του. Οι "άπληστοι" (greedy) αλγόριθμοι, όπως είναι οι SFS και SBS δεν είναι βέβαιο ότι θα βρουν μία καλή λύση, καθώς τείνουν να παγιδεύονται σε "τοπικά ελάχιστα". "Άπληστος" ονομάζεται ένας αλγόριθμος που, για την επίλυση ενός προβλήματος, επιλέγει σε κάθε βήμα την καλύτερη λύση που "βλέπει" ελπίζοντας ότι έτσι θα καταλήξει στο ολικό ελάχιστο. Για πολλά προβλήματα, η τακτική αυτή μπορεί όχι μόνο να μη βρει τη βέλτιστη λύση, αλλά μπορεί να καταλήξει και στη χειρότερη δυνατή. Ένα τέτοιο παράδειγμα είναι το πρόβλημα του πλανόδιου πωλητή (Travelling Salesman Problem – TSP).

Επιπλέον, κάποια χαρακτηριστικά μπορεί από μόνα τους να μην προσφέρουν πολλά στον ταξινομητή και να μην πετυχαίνουν υψηλό σκορ στο κριτήριο αξιολόγησης, αλλά σε συνδυασμό με κάποιο άλλο χαρακτηριστικό να δίνουν πολλή πληροφορία για την έξοδο. Παράδειγμα μίας τέτοιας περίπτωσης είναι όταν η έξοδος είναι το λογικό 'Αποκλειστικό Η' (Exclusive OR - XOR) δύο χαρακτηριστικών. Τότε κάθε χαρακτηριστικό από μόνο του δε μας δίνει καμία πληροφορία για την τιμή της εξόδου, μαζί όμως προσδιορίζουν πλήρως την έξοδο. Η SFS, που χαρακτηρίζεται από την αξιολόγηση κάθε χαρακτηριστικού ξεχωριστά και τη διαδοχική πρόσθεση χαρακτηριστικών βάσει του σκορ που πετυχαίνει το καθένα, συχνά αδυνατεί να καλύψει και αυτές τις περιπτώσεις. Από την άλλη μεριά, η SBE δεν μπορεί να επανεκτιμήσει χαρακτηριστικά που έχει ήδη απορρίψει. Οι εκθετικοί και οι τυχαίοι αλγόριθμοι συνήθως δίνουν καλά αποτελέσματα.

2.2.2.2 Κριτήρια Αξιολόγησης

Βάσει του είδους του κριτηρίου αξιολόγησης που χρησιμοποιούν οι αλγόριθμοι επιλογής χαρακτηριστικών χωρίζονται σε δύο κατηγορίες:

- Αλγόριθμοι-φίλτρα (filters), και
- Αλγόριθμοι-περιτυλίγματα (wrappers)

Οι filter αλγόριθμοι αξιολογούν τα υποψήφια υποσύνολα βάσει των εγγενών ιδιοτήτων των χαρακτηριστικών. Η αξιολόγηση δηλαδή γίνεται ανεξάρτητα από τον ταξινομητή και ελέγχονται διάφορα μέτρα που αφορούν στην περιεχόμενη πληροφορία των χαρακτηριστικών.

Τα κριτήρια αξιολόγησης των filter τεχνικών χωρίζονται στις παρακάτω κατηγορίες [DL97]:

- Μέτρα απόστασης ή διαχωριστικότητας (Distance or Separability Measures).
  Εκτιμούν τη διαχωριστικότητα μεταξύ των κλάσεων που παρέχει κάθε υποψήφιο υποσύνολο (Euclidian distance, Inter-class distance, Probabilistic distance, Class separability). Για ένα πρόβλημα δύο κλάσεων, ένα χαρακτηριστικό X προτιμάται από κάποιο άλλο χαρακτηριστικό Y, αν η διαφορά μεταξύ των υπό συνθήκη πιθανοτήτων των δύο κλάσεων

δεδομένου του X, δηλαδή η διαφορά $|P(Class\_1/X) - P(Class\_2/X)|$, είναι μεγαλύτερη από ότι δεδομένου του Y.

- Μέτρα πληροφορίας (Information Measures).

  Εξετάζουν πόση πληροφορία παρέχουν τα χαρακτηριστικά για την έξοδο, ή αλλιώς ποιο είναι το κέρδος πληροφορίας (information gain) κάθε χαρακτηριστικού. Βασίζονται σε έννοιες όπως η αμοιβαία πληροφορία (Mutual Information) και η εντροπία (Entropy). Το κέρδος πληροφορίας ενός χαρακτηριστικού X ορίζεται ως η διαφορά μεταξύ της εκ των προτέρων αβεβαιότητας για την έξοδο και της αναμενόμενης αβεβαιότητας δεδομένου της γνώσης του X. Επιλέγονται τα χαρακτηριστικά με το μεγαλύτερο κέρδος πληροφορίας.

- Μέτρα εξάρτησης ή συσχέτισης (Dependence or Correlation Measures).

  Αξιολογούν την ικανότητα να προβλεφθεί η τιμή μιας μεταβλητής από την τιμή μιας άλλης μεταβλητής. Χρησιμοποιούν κλασσικούς συντελεστές στατιστικής εξάρτησης και συσχέτισης (dependence, correlation). Αν η συσχέτιση ενός χαρακτηριστικού X με την κλάση εξόδου είναι υψηλότερη από ότι η συσχέτιση ενός χαρακτηριστικού Y, τότε το X προτιμάται από το Y.

- Μέτρα συνέπειας (Consistency Measures).

  Τα μέτρα αυτά είναι σχετικά καινούρια. Ελέγχουν διάφορες προτάσεις "συνέπειας" μεταξύ των χαρακτηριστικών, όπως για παράδειγμα, αν κάποια χαρακτηριστικά έχουν ίδια τιμή για διαφορετική έξοδο. Στην περίπτωση αυτή τα απορρίπτει ως "ασυνεπή". Τα μέτρα αυτά βασίζονται πολύ στα δεδομένα εκπαίδευσης (training data) και καταλήγουν στο μικρότερο υποσύνολο χαρακτηριστικών που ικανοποιούν τον αποδεκτό βαθμό ασυνέπειας (inconsistency rate), τον οποίο ορίζει ο χρήστης.

Στην επιλογή χαρακτηριστικών έχει αποδειχτεί ότι η επιλογή των ατομικά καλύτερων χαρακτηριστικών δεν οδηγεί απαραίτητα σε καλή απόδοση ταξινόμησης. Με άλλα λόγια, "τα m καλύτερα χαρακτηριστικά, δεν είναι τα καλύτερα m χαρακτηριστικά" [PLD05]. Γι' αυτό το λόγο, πέρα από τις σχέσεις χαρακτηριστικών – εξόδου, είναι σημαντικό να εξεταστούν και οι αλληλεπιδράσεις μεταξύ των χαρακτηριστικών και να μειωθεί και ο πλεονασμός (redundancy) μεταξύ τους. Πιο συγκεκριμένα, η ιδέα αυτήν βασίζεται στην υπόθεση ότι ένα καλό υποσύνολο χαρακτηριστικών περιλαμβάνει χαρακτηριστικά που είναι στενά συσχετισμένα με τη μεταβλητή εξόδου και ασυσχέτιστα μεταξύ τους. Στην κατεύθυνση αυτή έχει προταθεί το κριτήριο Minimum-redundancy-Maximum-relevance, που μπορεί να περιλαμβάνει μέτρα από οποιαδήποτε από τις παραπάνω κατηγορίες και που συνήθως δίνει καλά αποτελέσματα.

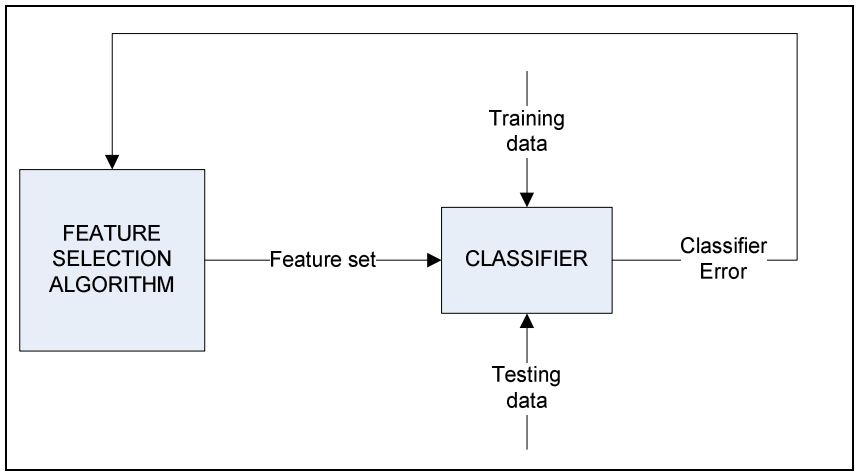
Οι filters τεχνικές χαρακτηρίζονται γενικά από γρήγορη εκτέλεση, καθώς πραγματοποιούνται ανεξάρτητα από τον αλγόριθμο εκπαίδευσης και δεν περιλαμβάνουν τις επαναλήψεις που απαιτεί η εκπαίδευση. Ακόμη ένα προσόν τους είναι η γενικότητα, αφού η αξιολόγηση βασίζεται στα ίδια τα δεδομένα. Από την άλλη μεριά, αυτή η ανεξάρτητη αντιμετώπιση μπορεί να αποτελέσει τελικά

μειονέκτημα, καθώς η επιλογή του καλύτερου υποσυνόλου μπορεί να μην είναι ανεξάρτητη του ταξινομητή [YH97]. Μην ξεχνάμε ότι ο τελικός μας στόχος είναι η ταξινόμηση, και μπορεί κάθε ταξινομητής να αποδίδει καλύτερα με διαφορετικό υποσύνολο χαρακτηριστικών. Μπορεί, με άλλα λόγια, η εύρεση των "καλύτερων" χαρακτηριστικών να μην συνεπάγεται απαραίτητα/να μην είναι ισοδύναμη με την επίτευξη της "καλύτερης" ακρίβειας ταξινόμησης [KJ97]. Στα μειονεκτήματα των filter προσεγγίσεων συγκαταλέγεται και η τάση τους να επιλέγουν σχετικά μεγάλα υποσύνολα.

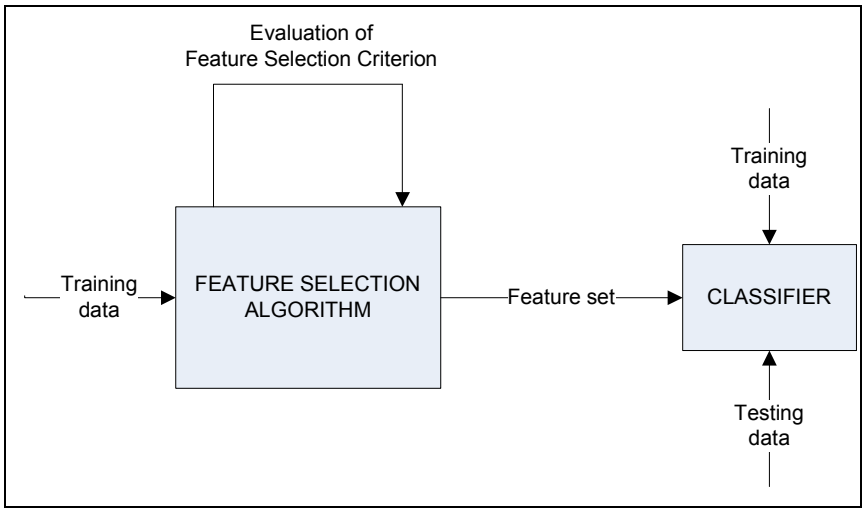
Αντιθέτως, στις wrappers τεχνικές η διαδικασία της επιλογής των σημαντικών χαρακτηριστικών είναι προσκολλημένη στη διαδικασία της εκπαίδευσης και το κριτήριο αξιολόγησης των υποψήφιων υποσυνόλων είναι η ίδια η απόδοση του ταξινομητή (classifier error). Οφείλουν το όνομά τους στο γεγονός ότι λειτουργούν σαν "περιτύλιγμα" γύρω από τον ταξινομητή. Οι τεχνικές αυτές είναι συνήθως πιο αποτελεσματικές και παρέχουν μεγαλύτερη ακρίβεια, καθώς βασίζονται στην "ιδιαίτερη σχέση" μεταξύ ταξινομητή και δεδομένων. Εφόσον ο στόχος της επιλογής χαρακτηριστικών είναι η ακριβέστερη ταξινόμηση, η χρήση της ίδιας της απόδοσης του ταξινομητή θεωρείται από πολλούς πιο σωστή αντιμετώπιση [KJ97]. Η ιδέα είναι ότι ο ταξινομητής αντιμετωπίζεται σαν ένα "μαύρο κουτί". Ο ταξινομητής εκπαιδεύεται με κάποια δεδομένα εκπαίδευσης (training data) για διάφορα υποσύνολα χαρακτηριστικών. Ο κίνδυνος να βασιστούμε υπερβολικά στα δεδομένα εκπαίδευσης (να συμβεί δηλαδή "overfitting") μπορεί να αποφευχθεί με μεθόδους στις οποίες θα αναφερθούμε εκτενέστερα στην ενότητα των Νευρωνικών Δικτύων. Η απόδοση του ταξινομητή αξιολογείται βάσει κάποιων άλλων δεδομένων (testing data) και το υποσύνολο για το οποίο επιτυγχάνεται η καλύτερη ακρίβεια ταξινόμησης, ή αλλιώς το μικρότερο λάθος ταξινόμησης (classifier error) είναι το υποσύνολο που επιλέγεται τελικά.

Μειονεκτήματα των τεχνικών αυτών είναι ότι έχουν αρκετά μεγάλες υπολογιστικές απαιτήσεις και ότι χρειάζονται περισσότερο χρόνο εκτέλεσης, καθώς για την αξιολόγηση κάθε υποψήφιου υποσυνόλου πρέπει να επαναλαμβάνεται ολόκληρη η διαδικασία εκπαίδευσης. Χαρακτηρίζονται, τέλος, από έλλειψη γενικότητας, αφού το αποτέλεσμά τους βασίζεται μόνο σε έναν συγκεκριμένο ταξινομητή.

Παρακάτω παραθέτουμε σχηματικά τη λογική των δύο τεχνικών που περιγράφηκαν.



Σχήμα 2.2: Wrapper προσέγγιση.



Σχήμα 2.3: Filter προσέγγιση.

Στον παρακάτω πίνακα φαίνεται μια συνοπτική σύγκριση των διαφόρων κριτηρίων αξιολόγησης όσον αφορά στη γενικότητά τους (Generalization), δηλαδή κατά πόσο το επιλεγμένο υποσύνολο είναι κατάλληλο για αρκετούς ταξινομητές, στη χρονική πολυπλοκότητά τους (Time complexity), δηλαδή πόσο χρόνο χρειάζονται για την εκτέλεσή τους και στην ακρίβειά τους (Accuracy), πόσο δηλαδή ακρίβεια πρόβλεψης παρέχεται με τη χρήση του επιλεγμένου υποσυνόλου. Η παύλα "-" στην τελευταία στήλη σημαίνει ότι δεν μπορεί να ειπωθεί κάποιο γενικό συμπέρασμα για την ακρίβεια του αντίστοιχου κριτηρίου [DL97].

Πίνακας 2.2: Συνοπτική Σύγκριση Κριτηρίων Αξιολόγησης

Κριτήριο Αξιολόγησης	Γενικότητα	Χρονική Πολυπλοκότητα	Ακρίβεια
Μέτρα απόστασης	Ναι	Χαμηλή	-
Μέτρα πληροφορίας	Ναι	Χαμηλή	-
Μέτρα εξάρτησης	Ναι	Χαμηλή	-
Μέτρα συνέπειας	Ναι	Μέτρια	-
Απόδοση ταξινομητή	Όχι	Πολύ Υψηλή	Πολύ Υψηλή

Όπως περιγράψαμε, κάθε λογική έχει τα πλεονεκτήματα και τα μειονεκτήματά της και το ποια θα επιλεχθεί τελικά άπτεται στην κρίση του ερευνητή και στις ανάγκες του εκάστοτε προβλήματος. Η νέα τάση είναι ο συνδυασμός των δύο αυτών προσεγγίσεων και η δημιουργία ενός υβριδικού μοντέλου συνήθως με έναν από τους δύο παρακάτω τρόπους:

- Εφαρμογή της wrapper λογικής και "τοπική" βελτίωση των αποτελεσμάτων της με εφαρμογή της filter λογικής.
- Εφαρμογή της filter λογικής ως "προπαρασκευαστικό" στάδιο, για απόρριψη κάποιων λύσεων και περιορισμό του χώρου αναζήτησης και εφαρμογή στη συνέχεια της wrapper λογικής, βασιζόμενοι στα αποτελέσματα της πρώτης.

Ο συνδυασμός των δύο προσεγγίσεων φαίνεται να είναι πολύ αποτελεσματική πρακτική, και την ιδέα αυτή θα ακολουθήσουμε και στην παρούσα διπλωματική.

2.3 Γενετικοί Αλγόριθμοι (Genetic Algorithms)

Οι γενετικοί αλγόριθμοι είναι μια κατηγορία επίλυσης προβλημάτων, της οποίας ο μηχανισμός βασίζεται στη Δαρβινική θεωρία της εξέλιξης (evolution) και στη φυσική επιλογή. Η φύση έχει έναν πολύ ισχυρό μηχανισμό εξέλιξης των οργανισμών που βασίζεται στον ακόλουθο κανόνα: οι οργανισμοί που δεν μπορούν να επιβιώσουν στο περιβάλλον τους πεθαίνουν, ενώ οι υπόλοιποι πολλαπλασιάζονται μέσω της αναπαραγωγής. Οι απόγονοι παρουσιάζουν μικρές διαφοροποιήσεις από τους προγόνους τους και τελικά, σε βάθος χρόνου, υπερισχύουν αυτοί που συγκεντρώνουν τα καλύτερα χαρακτηριστικά, δηλαδή τα καλύτερα προσόντα για επιβίωση. Αυτός ο κανόνας ονομάζεται φυσική επιλογή. Κάποιες φορές συμβαίνουν τυχαίες μεταλλάξεις στους οργανισμούς. Τις περισσότερες φορές τα μεταλλαγμένα όντα οδηγούνται στο θάνατο, ενώ σε κάποιες πιο σπάνιες περιπτώσεις η μετάλλαξη οδηγεί στη δημιουργία νέων, "καλύτερων" οργανισμών.

Αυτή ακριβώς τη λογική υιοθετούν και οι γενετικοί αλγόριθμοι. Δημιουργείται αρχικά, με τυχαίο συνήθως τρόπο, ένα σύνολο από υποψήφιες λύσεις του δεδομένου προβλήματος. Οι λύσεις αυτές βαθμολογούνται από μια συνάρτηση καταλληλότητας (fitness function). Η βαθμολόγηση αυτή

αποτελεί κριτήριο αξιολόγησης για κάθε λύση και δηλώνει την εγγύτητά της ως προς κάποια αποδεκτή λύση. Στη συνέχεια, από τον αρχικό αυτό πληθυσμό (population) δημιουργούνται ζευγάρια, όχι απαραίτητα μοναδικών γονέων (parents), με προτεραιότητα στις λύσεις που πέτυχαν υψηλά σκορ στη συνάρτηση καταλληλότητας. Κάθε ζευγάρι "ζευγαρώνει" (crossover) και δίνει δύο νέες λύσεις (offsprings) και δημιουργείται έτσι ο νέος πληθυσμός. Στους απογόνους υπάρχει μία μικρή πιθανότητα να εμφανιστεί μια μετάλλαξη (mutation), η οποία αλλάζει τη νέα λύση με τυχαίο τρόπο. Η παραπάνω διαδικασία επαναλαμβάνεται για το νέο πληθυσμό και οι συνθήκες τερματισμού ποικίλουν, με πιο συνηθισμένη τη σύγκλιση όλων των λύσεων σε μία.

Η γενική μορφή ενός γενετικού αλγόριθμου είναι η εξής [SS89]:

1. Δημιούργησε έναν αρχικό πληθυσμό Π, με N υποψήφιες λύσεις.
2. Για i=1 έως (Πλήθος_γενεών) do:
   (a) Αρχικοποίησε ένα σετ ζευγαρώματος M= Ø και ένα σετ απογόνων O.
   (b) For j = 1 έως N do:
   Προσέθεσε $f(a_j)/f_{aver}$ φορές την υποψήφια λύση a_i στο M.
   (c) For j=1 έως N/2 do:
   Διάλεξε ένα ζευγάρι λύσεων a_i και a_k από το M και κάνε
   O = O ∪ crossover(a_i, a_k) με πιθανότητα Pc.
   (d) For i=1 to N do:
   For j=1 to (Μέγεθος_Χρωμοσώματος) do
   Άλλαξε το j-στο bit στο a_i ∈ O με πιθανότητα P_m
   (e) Ανανέωσε τον πληθυσμό Π, συνδυάζοντας το Π με το O.

Όπου $f(a_j)$ είναι η τιμή της συνάρτησης καταλληλότητας για την a_j λύση και $f_{aver} = \sum_{j=1}^{N} f(a_j)/N$ η μέση τιμή των σκορ της συνάρτησης καταλληλότητας του κάθε πληθυσμού.

Σε μερικές υλοποιήσεις, κάποιες λύσεις του αρχικού πληθυσμού, αυτές με τις μεγαλύτερες βαθμολογίες στη συνάρτηση καταλληλότητας, περνάνε κατευθείαν στην επόμενη "γενιά" (generation) και αποτελούν τα "elite" μέλη του πληθυσμού. Παρακάτω αναφερόμαστε πιο αναλυτικά σε κάθε κομμάτι της δομής του γενετικού αλγορίθμου.

2.3.1 Αναπαράσταση Υποψήφιων Λύσεων

Στους γενετικούς αλγόριθμους κάθε υποψήφια λύση, ή αλλιώς κάθε υποψήφιο χρωμόσωμα κατ' αναλογία με τη βιολογία, αναπαρίσταται από μία συμβολοσειρά (string). Η συμβολοσειρά αυτή είναι συνήθως δυαδική (bit-string), μπορεί όμως να έχουμε και πιο σύνθετες μορφές αναπαράστασης. Στην περίπτωση της επιλογής χαρακτηριστικών, που μας ενδιαφέρει στη παρούσα διπλωματική, κάθε υποψήφια λύση, δηλαδή κάθε υποψήφιο υποσύνολο, αναπαρίσταται από μια δυαδική

συμβολοσειρά, στην οποία το ψηφίο "1" στην i θέση του χρωμοσώματος δηλώνει την παρουσία του i χαρακτηριστικού στο υποσύνολο, ενώ το ψηφίο "0" την απουσία του.

2.3.2 Συνάρτηση Καταλληλότητας

Η συνάρτηση καταλληλότητας αποτελεί το κριτήριο για την αξιολόγηση των χρωμοσωμάτων. Δέχεται ως είσοδο το χρωμόσωμα και επιστρέφει έναν αριθμό που υποδηλώνει το βαθμό καταλληλότητας του. Το πεδίο τιμών της είναι συνήθως το διάστημα των πραγματικών αριθμών. Ο τρόπος υλοποίησής της εξαρτάται από το εκάστοτε πρόβλημα και μπορεί να είναι από απλός μέχρι πολύπλοκος. Γενικά πρέπει να αντικατοπτρίζει ρεαλιστικά την αξία της υποψήφιας λύσης.

2.3.3 Επιλογή Γονέων (Selection)

Η διαδικασία της επιλογής (selection) αφορά στην απόδοση πιθανότητας στα μέλη του πληθυσμού για το αν θα επιλεχθούν προς αναπαραγωγή ή όχι. Κάποια χρωμοσώματα με υψηλή τιμή στη συνάρτηση καταλληλότητας ενδέχεται να επιλεχθούν περισσότερες από μία φορά, ενώ κάποια άλλα με χαμηλότερη τιμή στη συνάρτηση καταλληλότητας ενδέχεται να επιλεχθούν λιγότερο έως καθόλου. Τα προς αναπαραγωγή χρωμοσώματα αντιγράφονται σε ένα σετ ζευγαρώματος (mating pool), μεγέθους ίσο με τον αρχικό πληθυσμό. Για την επιλογή των χρωμοσωμάτων που θα αντιγραφούν στο σετ ζευγαρώματος έχουν αναπτυχθεί διάφορες τεχνικές. Οι πιο συνηθισμένες είναι οι παρακάτω:

- Επιλογή ρουλέτας (roulette wheel selection).

  Παράγεται το άθροισμα S όλων των τιμών της συνάρτησης καταλληλότητας των υποψήφιων λύσεων. Επιλέγεται ένας τυχαίος αριθμός n, από 0 έως S, χρησιμοποιώντας συνάρτηση ομοιόμορφης κατανομής για τη δημιουργία των τυχαίων αριθμών. Εξετάζεται επαναληπτικά κάθε υποψήφια λύση και η τιμή της προστίθεται σε έναν μετρητή K, μέχρι το σημείο όπου το K γίνει μεγαλύτερο ή ίσο του n. Τότε επιλέγεται η αντίστοιχη λύση, ο K μηδενίζεται και αρχίζει εκ νέου η εξέταση των υποψήφιων λύσεων μέχρι να επιλεχθεί το επιθυμητό πλήθος υποψήφιων λύσεων. Η λογική αυτή βασίζεται στην υπόθεση ότι οι υποψήφιες λύσεις με μεγάλο σκορ έχουν μεγαλύτερη πιθανότητα να αυξήσουν πολύ την τιμή του K ώστε αυτός να ξεπεράσει το n και επομένως να επιλεγούν. Ονομάζεται έτσι γιατί ουσιαστικά προσομοιώνει μια ρουλέτα, κάθε τμήμα της οποίας αντιστοιχεί σε ένα χρωμόσωμα και είναι ανάλογο της τιμή του σκορ του.
- Επιλογή αναλογικής καταλληλότητας (Fitness Proportionate Selection).

  Η πιθανότητα επιλογής ενός χρωμοσώματος a_j υπολογίζεται από τη σχέση: $P(a_j) = \frac{f(a_j)}{\Sigma_{j=1}^{N} f(a_j)}$, όπου $f(a_j)$ είναι η τιμή της συνάρτησης καταλληλότητας για το a_j χρωμόσωμα. Η

πιθανότητα δηλαδή να επιλεγεί είναι ευθέως ανάλογη της καταλληλότητας του και αντιστρόφως ανάλογη της καταλληλότητας των υπολοίπων χρωμοσωμάτων.

- Επιλογή τουρνουά (Tournament Selection).

  Επιλέγονται τυχαία τα καταλληλότερα χρωμοσώματα ενός προκαθορισμένου από τον χρήστη πλήθους (tournament size) και στη συνέχεια επιλέγονται τα καλύτερα από αυτά για να γίνουν γονείς.

- Στοχαστική Ομοιόμορφη (Stochastic Uniform Selection).

  "Σχεδιάζεται" μία ευθεία γραμμή και κάθε υποψήφιο χρωμόσωμα αντιστοιχεί σε ένα κομμάτι της, μήκους ανάλογου της τιμής του στη συνάρτηση καταλληλότητας. Ο αλγόριθμος "κινείται" κατά μήκος της γραμμής βήμα-βήμα κάνοντας ίσου μεγέθους βήματα για κάθε χρωμόσωμα. Σε κάθε του βήμα, τοποθετεί στο σετ ζευγαρώματος τον γονέα στου οποίου το κομμάτι προσγειώθηκε. Το πρώτο βήμα είναι ένας τυχαίος αριθμός, μικρότερος από το μέγεθος του βήματος.

- Επιλογή Κατάταξης (Ranking Selection).

  Τα χρωμοσώματα ταξινομούνται και βαθμολογούνται. Αυτό με τη χειρότερη τιμή στη fitness function λαμβάνει την τιμή 1, το δεύτερο χειρότερο λαμβάνει την τιμή 2 κ. τ. λ. και το καλύτερο λαμβάνει την τιμή N, όπου N το πλήθος των χρωμοσωμάτων του πληθυσμού. Μετά από αυτό, όλα τα χρωμοσώματα έχουν μια πιθανότητα να επιλεχθούν, ανάλογη της κατάταξής τους. Για παράδειγμα:

  $p = \frac{1}{N}[s - 2(s-1)\frac{r(a_j)-1}{N-1}]$, όπου N ο αριθμός των χρωμοσωμάτων στον πληθυσμό, $r(a_j) \in \{1,2,\ldots,N\}$ η σειρά κατάταξης του a_j χρωμοσώματος και $s \in [1,2]$ μία παράμετρος. Η τεχνική αυτή "κανονικοποιεί" κατά κάποιο τρόπο τα σκορ και βοηθάει στις περιπτώσεις που η τιμές της συνάρτησης καταλληλότητας διαφέρουν πολύ. Αν για παράδειγμα, το καλύτερο χρωμόσωμα έχει πολύ υψηλή τιμή και τα επόμενα καλύτερα έχουν πολύ χαμηλότερες τιμές, τότε τα δεύτερα έχουν πολύ χαμηλή πιθανότητα να διαλεχτούν.

Σε όλες τις παραπάνω τεχνικές φαίνεται η πρακτική υλοποίηση της ιδέας της φυσικής επιλογής, κατά την οποία ο καλύτερος επιβιώνει και αναπαράγεται.

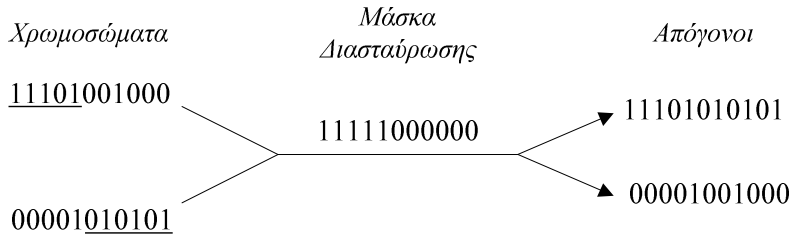
Έχουμε ήδη αναφερθεί στην ιδέα του ελιτισμού (elitism). Επειδή όταν φτιάχνουμε τον καινούριο πληθυσμό ως αποτέλεσμα διασταυρώσεων του προηγούμενου, κινδυνεύουμε να χάσουμε το καλύτερο αποτέλεσμα, έχει εισαχθεί η μέθοδος του ελιτισμού. Σύμφωνα με αυτή, το καλύτερο χρωμόσωμα, ή ένας μικρός αριθμός των καλύτερων, αντιγράφονται κατευθείαν στην επόμενη γενιά και μετά πραγματοποιείται η διαδικασία της επιλογής με κάποιον από τους τρόπους που περιγράψαμε. Ο αριθμός όμως των "elite" χρωμοσωμάτων πρέπει να είναι μικρός, έτσι ώστε να μην "κυριαρχούν" (dominate) στον πληθυσμό και να αποφεύγεται ο κίνδυνος παγίδευσης σε τοπικά ελάχιστα.

2.3.4 Διασταύρωση (Crossover)

Διασταύρωση (crossover) ονομάζεται η διαδικασία κατά την οποία δημιουργούνται δύο απόγονοι από δύο γονείς-υποψήφιες λύσεις, αντιγράφοντας επιλεγμένα bits από κάθε γονέα με τρόπο τέτοιο ώστε το i-οστό bit του κάθε απογόνου να είναι το i-οστό bit ενός εκ των δύο γονέων του. Το πώς κάθε γονέας θα συνεισφέρει στον απόγονο καθορίζεται από τη μάσκα διασταύρωσης (crossover mask). Βάσει της μάσκας διασταύρωσης έχουμε τα παρακάτω είδη διασταύρωσης:

- Διασταύρωση ενός σημείου (Single-point crossover).

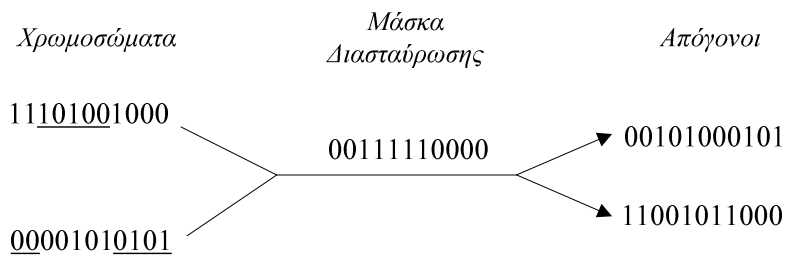
  Σε αυτό το είδος διασταύρωσης, η μάσκα διασταύρωσης αποτελείται από ένα πλήθος συνεχόμενων άσσων (“1”) ακολουθούμενο από συνεχόμενα μηδενικά (“0”), όσα χρειάζονται για να συμπληρωθεί το μέγεθος των χρωμοσωμάτων. Ο ένας απόγονος παίρνει τα bits που αντιστοιχούν σε άσσους από τον ένα γονέα και τα bits που αντιστοιχούν σε μηδενικά από τον άλλο γονέα. Για τον άλλο απόγονο ακολουθείται η αντίστροφη διαδικασία, παίρνει δηλαδή τα bits που δεν χρησιμοποιήθηκαν για τη δημιουργία του πρώτου.



Σχήμα 2.4: Παράδειγμα Διασταύρωσης ενός σημείου.

- Διασταύρωση δύο σημείων (Two-point crossover).

  Στη διασταύρωση δύο σημείων η μάσκα διασταύρωσης ξεκινά με έναν αριθμό μηδενικών, ακολουθείται από έναν αριθμό άσσων και συμπληρώνεται με τον απαιτούμενο αριθμό μηδενικών. Ο ένας απόγονος παίρνει από τον ένα γονέα τα bits που αντιστοιχούν στους άσσους και από τον άλλο αυτά που αντιστοιχούν σε μηδενικά. Το ίδιο ισχύει και για τον άλλο απόγονο, αλλά αντίστροφα. Οι αριθμοί των μηδενικών και των άσσων επιλέγονται κάθε φορά με τυχαίο τρόπο.



Σχήμα 2.5: Παράδειγμα Διασταύρωσης δύο σημείων.

- Ομοιόμορφη διασταύρωση (Uniform Crossover).

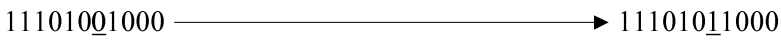
  Η μάσκα διασταύρωσης είναι μια τυχαία ακολουθία από μηδενικά και άσσους, ομοιόμορφα κατανεμημένα. Η δημιουργία των απογόνων γίνεται με τρόπο αντίστοιχο με αυτούς που περιγράφηκαν προηγουμένως.

Σχήμα 2.6: Παράδειγμα Ομοιόμορφης Διασταύρωσης.

Πρέπει, τέλος, να επισημάνουμε ότι η διασταύρωση είναι μια στοχαστική διαδικασία, που συμβαίνει με μία προκαθορισμένη πιθανότητα (crossover rate). Το γεγονός αυτό συμβάλλει στην ευελιξία του γενετικού αλγορίθμου και στην αξία του ως μέθοδο επίλυσης διαφόρων προβλημάτων.

2.3.5 Μετάλλαξη (Mutation)

Η μετάλλαξη, όπως και στη φύση, είναι μια στοχαστική (τυχαία) διαδικασία. Επιλέγεται τυχαία ένα bit ενός χρωμοσώματος και αλλάζει τιμή, δηλαδή από "0" γίνεται "1", ή αντίστροφα. Αυτές οι τυχαίες αλλαγές των υποψήφιων λύσεων που παρέχει η διαδικασία της μετάλλαξης οδηγούν την αναζήτηση σε νέες περιοχές του χώρου αναζήτησης και αποτρέπουν το ενδεχόμενο της παγίδευσης σε τοπικά μέγιστα. Η πιθανότητα να συμβεί μετάλλαξη (mutation rate) είναι πολύ μικρότερη της πιθανότητας για διασταύρωση.



Σχήμα 2.7: Παράδειγμα Μετάλλαξης.

2.3.6 Κριτήρια Τερματισμού (Stopping Criteria)

Υπάρχουν διάφορα κριτήρια για το πότε θα τερματίσει ο γενετικός αλγόριθμος. Παραθέτουμε μερικά:

- Σύγκλιση όλων των υποψήφιων λύσεων σε μία.
- Μέγιστος αριθμός επαναλήψεων, δηλαδή γενεών.
- Χρονικό όριο (time limit). Μέγιστος, δηλαδή, χρόνος σε δευτερόλεπτα που θα τρέξει ο γενετικός αλγόριθμος.

- Όριο στην τιμή της συνάρτησης καταλληλότητας (fitness function limit). Αν η τιμή του καλύτερου σκορ είναι μεγαλύτερη (ή μικρότερη, ανάλογα με το αν ο στόχος μας είναι η μεγιστοποίηση της συνάρτησης καταλληλότητας ή η ελαχιστοποίηση της αντίστοιχα) ή ίση του προκαθορισμένου από τον χρήστη ορίου, τότε ο αλγόριθμος τερματίζει.
- Ανοχή στη μεταβολή της τιμής της συνάρτησης καταλληλότητας (fitness function tolerance). Αν είτε για κάποιο προκαθορισμένο αριθμό γενεών, είτε για κάποιο προκαθορισμένο χρονικό διάστημα, η τιμή της συνάρτησης καταλληλότητας βελτιώνεται λιγότερο από το όριο ανοχής, ο αλγόριθμος τερματίζει.

2.3.7 Σύγκλιση του Πληθυσμού

Ένας αποδοτικός γενετικός αλγόριθμος θα πρέπει μετά από κάποιο αριθμό επαναλήψεων να συγκλίνει σε ένα ολικό μέγιστο (ή ελάχιστο). Συχνά στους γενετικούς αλγόριθμους εμφανίζονται δύο προβλήματα που χρήζουν αντιμετώπισης. Αυτά είναι η πρόωρη σύγκλιση (premature convergence) και η αργή σύγκλιση (slow convergence).

Κατά την πρόωρη σύγκλιση, ο γενετικός αλγόριθμος καταλήγει πολύ γρήγορα γύρω από κάποια λύση, η οποία όμως ενδέχεται να αποτελεί τοπικό μέγιστο. Η λειτουργία της μετάλλαξης του προσφέρει μεν τη δυνατότητα να ξεφύγει από αυτό το τοπικό μέγιστο, αλλά είναι αρκετά μικρή η πιθανότητα να συμβεί. Έτσι, αρκετές φορές ο αλγόριθμος τερματίζει χωρίς να έχει προλάβει να ξεφύγει. Το φαινόμενο αυτό παρουσιάζεται στις περιπτώσεις που η συνάρτηση καταλληλότητας παρουσιάζει πολύ απότομες μεταβολές, επομένως πολλά τοπικά μέγιστα. Η κατάσταση αυτή μπορεί να αντιμετωπιστεί με την προσέγγιση της συνάρτησης καταλληλότητας από μία άλλη συνάρτηση, λιγότερο απότομη. Ακόμη ένας τρόπος είναι ο καθορισμός κάποιων ορίων όσον αφορά στο πόσες φορές επιλέγεται ένα χρωμόσωμα για εισαγωγή στο σετ ζευγαρώματος.

Η αργή σύγκλιση είναι το ακριβώς αντίθετο φαινόμενο. Όταν δηλαδή μετά από αρκετές επαναλήψεις ο γενετικός αλγόριθμος εξακολουθεί να μη συγκλίνει. Το φαινόμενο αυτό παρατηρείται όταν η συνάρτηση καταλληλότητας εμφανίζει μικρές κλίσεις, με αποτέλεσμα οι "καλές" και οι "κακές" τιμές της να έχουν μικρές διαφορές. Αντιμετωπίζεται με την αντιστοίχηση της συνάρτησης καταλληλότητας, με μία με πιο έντονη κλίση.

2.3.8 Χάσμα γενεών (Generation Gap)

Ως χάσμα γενεών (generation gap) ορίζεται το ποσοστό των χρωμοσωμάτων που ανανεώνεται σε κάθε γενιά, προς το συνολικό πλήθος των χρωμοσωμάτων. Συνήθως ο συντελεστής αυτός ισούται με τη μονάδα. Υπάρχει, όμως, και η μέθοδος μερικής ανανέωσης (steady-state replacement) σύμφωνα με την οποία σε έναν πληθυσμό συνυπάρχουν δύο γενεές. Επιτρέπεται έτσι η δυνατότητα στους απογόνους να "ανταγωνιστούν" τους γονείς τους για την επικράτηση του καλύτερου. Στην περίπτωση που τα χάσμα γενεών είναι διάφορο της μονάδας, εκτός από το πώς θα επιλεχθούν τα

χρωμοσώματα που θα αποτελέσουν γονείς (selection), πρέπει να καθοριστεί και ο τρόπος με τον οποίο θα επιλεχθούν τα ίσου πλήθους χρωμοσώματα που θα παραχωρήσουν τη θέση τους στους νέους απογόνους. Συνήθως η επιλογή γίνεται τυχαία, ή με πιθανότητα αντιστρόφως ανάλογη της τιμής τους στη συνάρτηση καταλληλότητας.

2.3.9 Επιλογή των βέλτιστων παραμέτρων

Είναι σημαντικό να επισημάνουμε ότι η επιλογή των βέλτιστων παραμέτρων αποτελεί ακόμα ανοιχτό ερευνητικό θέμα [IML+01]. Υπάρχουν μερικοί εμπειρικοί κανόνες τους οποίους αναφέρουμε παρακάτω [G86]:

- Μέγεθος Πληθυσμού (Population Size).

  Το μέγεθος του πληθυσμού, το πλήθος δηλαδή των υποψήφιων λύσεων – χρωμοσωμάτων κάθε γενιάς, επηρεάζει την απόδοση του γενετικού αλγορίθμου. Γενικά, οι Γενετικοί Αλγόριθμοι δεν αποδίδουν καλά με πολύ μικρούς πληθυσμούς, γιατί τότε περιορίζεται το μέτωπο αναζήτησης. Ένας μεγαλύτερος πληθυσμός προσφέρει τη δυνατότητα να εξεταστούν περισσότερες λύσεις και αποτρέπει την πρόωρη σύγκλιση σε τοπικά ελάχιστα. Από την άλλη μεριά, ένας πολύ μεγάλος πληθυσμός απαιτεί περισσότερους υπολογισμούς για κάθε γενεά και ενδεχομένως να συνοδεύεται από αργή σύγκλιση. Στις περισσότερες περιπτώσεις που συναντήθηκαν στη βιβλιογραφία, ο πληθυσμός περιλάμβανε 20 με 100 χρωμοσώματα.
- Πιθανότητα Διασταύρωσης (Crossover Rate).

  Όπως έχουμε αναφέρει, η πιθανότητα διασταύρωσης καθορίζει τη συχνότητα με την οποία εκτελείται η διασταύρωση. Όσο μεγαλύτερη είναι η πιθανότητα διασταύρωσης, τόσο πιο γρήγορα εισάγονται νέα "άτομα" στον πληθυσμό. Αν η πιθανότητα διασταύρωσης είναι πολύ μεγάλη, ενδέχεται λύσεις με πολύ καλό σκορ να απομακρύνονται πριν η διαδικασία της επιλογής προλάβει να παράξει βελτιώσεις. Αν όμως η πιθανότητα διασταύρωσης είναι πολύ μικρή, η αναζήτηση περιορίζεται λόγω του χαμηλού ρυθμού εξέτασης νέων υποψήφιων λύσεων. Για την επιλογή του σωστού ρυθμού διασταύρωσης απαιτείται πειραματισμός, ενώ σημαντικό ρόλο στη ρύθμιση αυτής της παραμέτρου φαίνεται να παίζει το μέγεθος του πληθυσμού. Εμπειρικά και πειραματικά [G86] έχουν προκύψει οι παρακάτω τιμές ρυθμού διασταύρωσης ανάλογα με την τιμή στην οποία κυμαίνεται το μέγεθος του πληθυσμού:

Πίνακας 2.3: Ενδεικνυόμενη τιμή Πιθανότητας Διασταύρωσης βάσει του Μεγέθους του Πληθυσμού.

Μέγεθος Πληθυσμού	Πιθανότητα Διασταύρωσης
30	0.88
50	0.5
80	0.3

Η παραπάνω εξάρτηση μπορεί να εξηγηθεί ως εξής: Όταν το μέγεθος του πληθυσμού είναι σχετικά μικρό, η αναζήτηση περιλαμβάνει λιγότερα "δείγματα" υποψήφιων λύσεων. Επομένως, χρειαζόμαστε μεγαλύτερο ρυθμό διασταύρωσης για να διευρύνουμε το μέτωπο αναζήτησης και για να αποφύγουμε το ενδεχόμενο κάποιο χρωμόσωμα να "επιβληθεί" στον πληθυσμό και να έχουμε έτσι πρόωρη σύγκλιση σε τοπικό ελάχιστο.

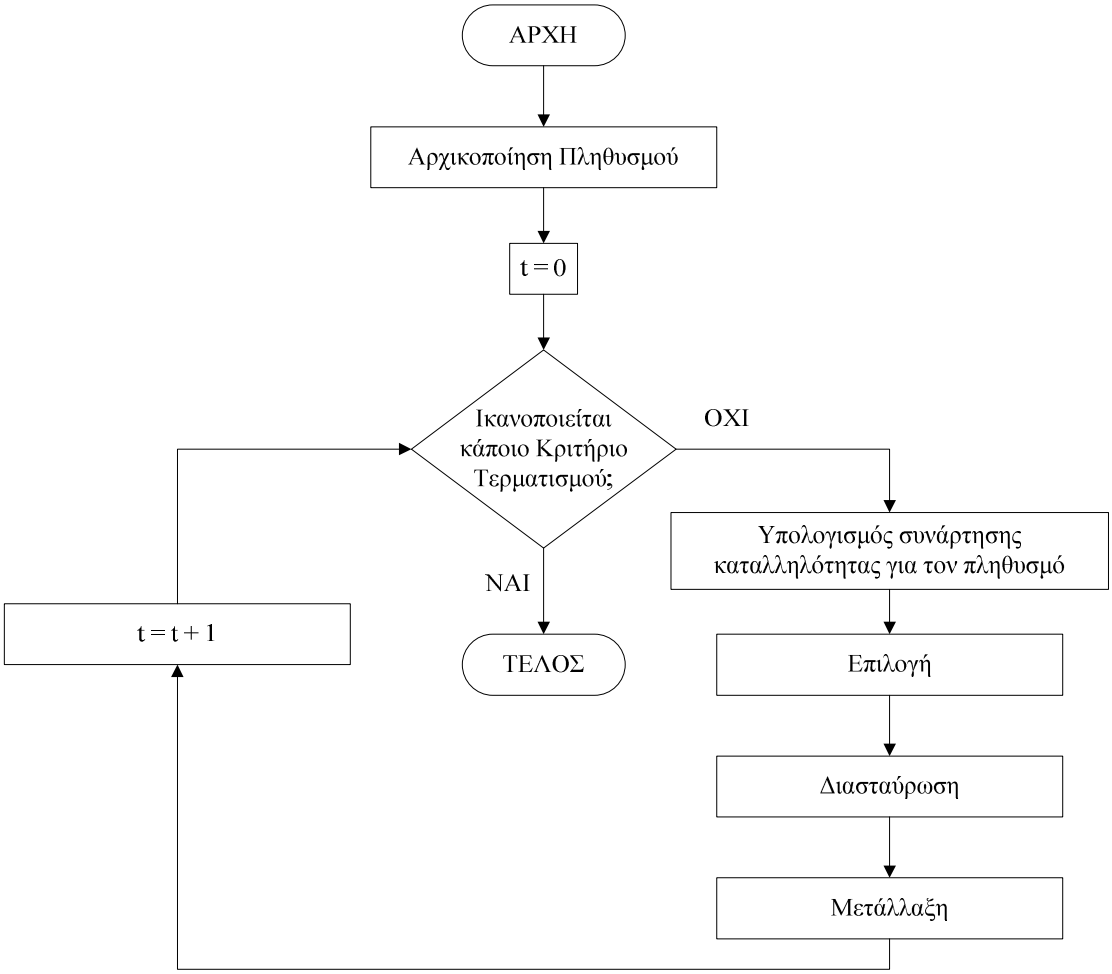
- Πιθανότητα Μετάλλαξης (Mutation Rate).

  Η μετάλλαξη αυξάνει την "ευελιξία" της αναζήτησης και αποτρέπει το ενδεχόμενο να διατηρεί συνεχώς την ίδια τιμή ένα συγκεκριμένο bit των χρωμοσωμάτων. Πολύ μεγάλη τιμή πιθανότητας μετάλλαξης προσδίδει μεγάλη τυχαιότητα στην αναζήτηση, ενώ η απουσία μετάλλαξης έχει αποδειχθεί ότι επιφέρει χαμηλή απόδοση, γεγονός που φανερώνει τη συμβολή της στην "αναζωογόνηση" του πληθυσμού. Εμπειρικά έχει προκύψει ότι η πιθανότητα μετάλλαξης καλό είναι να μην ξεπερνάει την τιμή 0,1. Μία ενδεικτική τιμή που χρησιμοποιείται συχνά σε εφαρμογές είναι 0,001.

- Χάσμα Γενεών (Generation Gap).

  Η παράμετρος αυτή καθορίζει το ποσοστό του πληθυσμού που θα αντικαθίσταται σε κάθε γενεά. Αν το Χάσμα Γενεών ισούται με 1, σε κάθε επανάληψη ανανεώνεται ολόκληρος ο πληθυσμός, ενώ αν ισούται με 0,5, ο μισός πληθυσμός επιζεί στην επόμενη γενιά. Έχει παρατηρηθεί ότι η υιοθέτηση μεγάλου Χάσματος Γενεών οδηγεί συνήθως σε καλύτερη απόδοση, σε συνδυασμό όμως με τη στρατηγική του "ελιτισμού", έτσι ώστε να εξαλειφθεί ο κίνδυνος απώλειας κάποιων καλών λύσεων.

Παραθέτουμε, τέλος, ένα διάγραμμα ροής που συνοψίζει τη διαδικασία εκτέλεσης ενός γενετικού αλγορίθμου [SR06]:



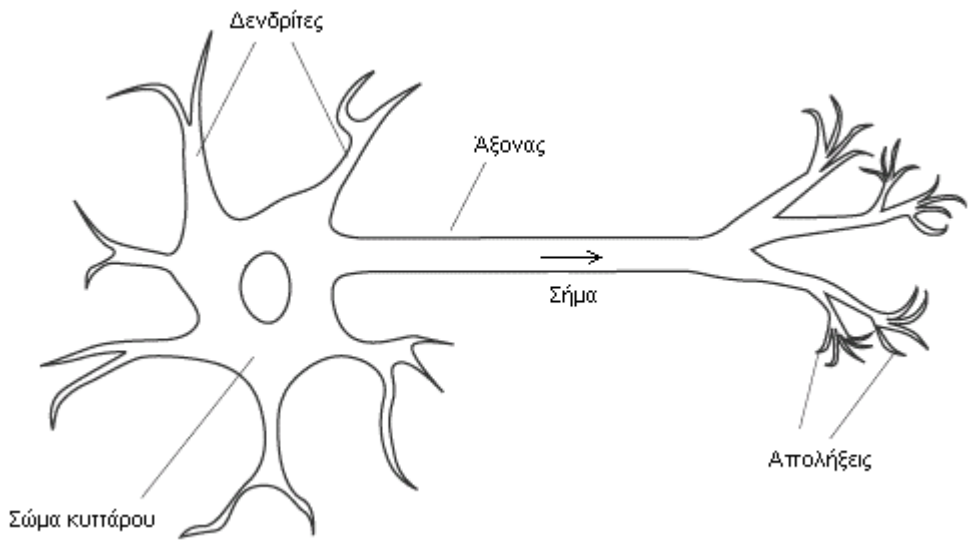
Σχήμα 2.8: Διάγραμμα Ροής Γενετικού Αλγορίθμου.

2.4 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα είναι ένα ιδιαίτερο υπολογιστικό εργαλείο που έχει αναπτυχθεί τα τελευταία χρόνια και γνωρίζει όλο και μεγαλύτερη αποδοχή από διάφορους τομείς της επιστήμης και της ανθρώπινης δραστηριότητας. Αποτελούν μία ιδιαίτερη προσέγγιση στην επίλυση περίπλοκων προβλημάτων και στη δημιουργία συστημάτων με νοημοσύνη, της οποίας η λειτουργία και η δομή βασίζεται στη λειτουργία του ανθρώπινου εγκεφάλου. Μπορούν να οριστούν σαν δομές που περιλαμβάνουν πολλές απλές μονάδες επεξεργασίας συνδεδεμένες μεταξύ τους, ικανές να πραγματοποιούν περίπλοκους υπολογισμούς επεξεργασίας δεδομένων. Στη συνέχεια θα αναλύσουμε όσο πληρέστερα μπορούμε το θεωρητικό υπόβαθρο των νευρωνικών δικτύων. Τα νευρωνικά δίκτυα αποτελούν πλέον ένα ευρύ και αυτόνομο επιστημονικό πεδίο, η πλήρης κάλυψη του οποίου ξεφεύγει από τους στόχους αυτής της διπλωματικής. Αρχικά θα περιγράψουμε συνοπτικά τη δομή και τη λειτουργία του βιολογικού νευρώνα, για να εισαχθούμε στη συνέχεια στην έννοια του τεχνητού νευρώνα και του τεχνητού νευρωνικού δικτύου.

2.4.1 Βιολογικός Νευρώνας

Η βασική δομική μονάδα του νευρικού μας συστήματος είναι ο νευρώνας (neuron). Ένας τυπικός βιολογικός νευρώνας αποτελείται από το σώμα (body), δηλαδή τον πυρήνα του, τον άξονα (axon), που είναι η έξοδος του νευρώνα και το μέσο σύνδεσής του με άλλους νευρώνες, και τους δενδρίτες (dendrites) μέσω των οποίων λαμβάνει σήματα ηλεκτροχημικής φύσεως από γειτονικούς νευρώνες, αποτελούν δηλαδή το σημείο εισόδου του νευρώνα. Σε κάθε δενδρίτη υπάρχει ένα απειροελάχιστο κενό που ονομάζεται σύναψη (synapse). Οι συνάψεις μέσω διάφορων χημικών διαδικασιών επιταχύνουν ή επιβραδύνουν τη ροή των ηλεκτρικών σημάτων προς το σώμα του νευρώνα.



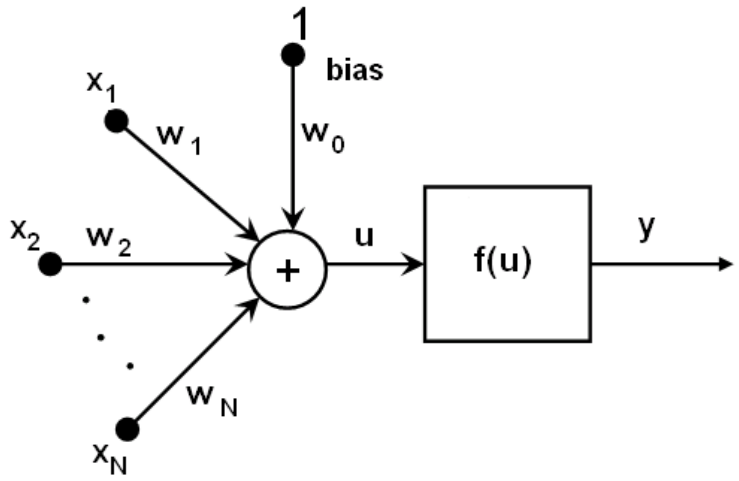
Σχήμα 2.9: Αναπαράσταση βιολογικού νευρώνα.

Αν το άθροισμα των σημάτων που δέχεται ο νευρώνας στους δενδρίτες του είναι αρκετά ισχυρό έτσι ώστε να ξεπεραστεί μία συγκεκριμένη τιμή κατωφλίου, τότε ο νευρώνας ενεργοποιείται και μεταδίδει ένα αντίστοιχο σήμα κατά μήκος του άξονά του και το μεταβιβάζει σε άλλα νευρικά κύτταρα των οποίων οι δενδρίτες συνδέονται με κάποιες από τις απολήξεις του άξονα. Με τη σειρά τους αυτά ενδέχεται να ενεργοποιηθούν και να πραγματοποιήσουν την ίδια διαδικασία. Γίνεται φανερό ότι ο κάθε νευρώνας λειτουργεί σαν διακόπτης (on/off). Αν το σύνολο των διεγέρσεων που δέχεται ξεπεράσει το κατώφλι του, τότε "κλείνει" και μεταφέρει το σήμα. Ο εγκέφαλός μας αποτελείται από δισεκατομμύρια νευρώνες, από πολλές δηλαδή τέτοιες απλές μονάδες, που συνδεδεμένες όλες μαζί μπορούν να επιτελέσουν εξαιρετικά περίπλοκες διαδικασίες.

Στο μηχανισμό που περιγράψαμε βασίζεται η λειτουργία των τεχνητών νευρωνικών δικτύων. Στη συνέχεια θα περιγράψουμε τη λογική τους αναφερόμενοι στα μέρη που τα απαρτίζουν.

2.4.2 Μοντέλο Τεχνητού Νευρώνα

Το πιο απλό νευρωνικό δίκτυο που μπορεί να μελετηθεί και να σχεδιαστεί είναι αυτό που αποτελείται από έναν μόνο νευρώνα. Ο τεχνητός νευρώνας (artificial neuron) είναι ένα υπολογιστικό μοντέλο με μέρη που μιμούνται αυτά του βιολογικού νευρώνα. Δέχεται κάποια σήματα εισόδου x_1, x_2, …, x_n τα οποία αντιστοιχούν σε κάποιες μεταβλητές. Κάθε τέτοιο σήμα εισόδου μεταβάλλεται μέσω κάποιας τιμή βάρους w_i (weight), που έχει τον αντίστοιχο ρόλο που έχει η σύναψη στους βιολογικούς νευρώνες. Αντίστοιχα με την επιταχυντική ή επιβραδυντική λειτουργία της σύναψης, η τιμή του βάρους μπορεί να είναι θετική ή αρνητική, δηλαδή ενισχυτική ή αποσβεννύμενη. Το σώμα του τεχνητού νευρώνα χωρίζεται σε δύο μέρη, τον αθροιστή (sum), ο οποίος προσθέτει τα ήδη επηρεασμένα από τα βάρη σήματα εισόδου παράγοντας μία ποσότητα S, και τη συνάρτηση ενεργοποίησης (activation function), που λειτουργεί σαν ένα φίλτρο, αντίστοιχα με τη λειτουργία του κατωφλίου στους βιολογικούς νευρώνες, και είτε επιτρέπει τη διέλευση του αθροίσματος των εισόδων, είτε όχι. Η τελική τιμή της εξόδου του νευρώνα διαμορφώνεται συναρτήσει του S και της τιμής κατωφλίου της συνάρτησης ενεργοποίησης. Μερικές φορές, εκτός από τα σήματα εισόδου και τα αντίστοιχα βάρη τους, ο νευρώνας έχει και μία επιπλέον είσοδο που επιδρά συνεχώς με μία τιμή x_0=1 με ένα βάρος w_0, που ονομάζεται πόλωση (bias).



Σχήμα: Δομή Τεχνητού Νευρωνίου

Διατυπώνουμε και μαθηματικά τα παραπάνω:

Η διέγερση u δίνεται από τη σχέση:

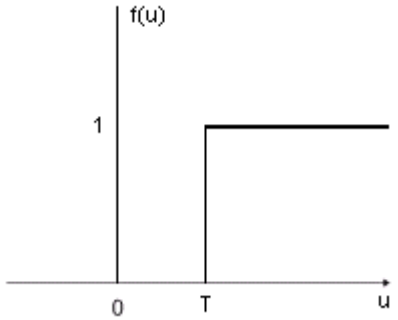
u = $\sum_{i=1}^{N}(w_i\, x_i - w_0)$, και η έξοδος y από τη σχέση:

y = $f(\sum_{i=1}^{N} w_i\, x_i - w_0)$

Τρεις τυπικές περιπτώσεις συναρτήσεων ενεργοποίησης είναι:

- Η βηματική συνάρτηση (step function), η οποία δίνει στην έξοδο αποτέλεσμα (συνήθως 1) μόνο αν η τιμή του αθροιστή είναι μεγαλύτερη από μία τιμή κατωφλίου T. Έχει την παρακάτω έκφραση:

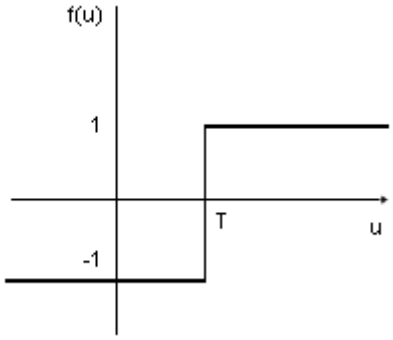
  f(u) = $\begin{Bmatrix} 1, & \alpha\nu\ u \geq T \\ 0, & \alpha\nu\ u < T \end{Bmatrix}$



Σχήμα 2.10: Βηματική Συνάρτηση.

- Η συνάρτηση προσήμου, (sign function), η οποία δίνει στην έξοδο θετική ή αρνητική πληροφορία αν η τιμή του αθροιστή είναι μεγαλύτερη ή μικρότερη από μία τιμή κατωφλίου T. Έχει την παρακάτω έκφραση:

  f(u) = $\begin{Bmatrix} 1, & \alpha\nu\ u \geq T \\ 0, & \alpha\nu\ u = 0 \\ -1, & \alpha\nu\ u < T \end{Bmatrix}$

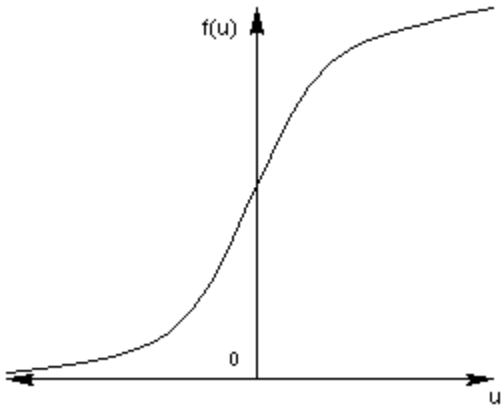


Σχήμα 2.11: Συνάρτηση προσήμου.

- Σιγμοειδής συνάρτηση (sigmoid function), η οποία αποτελεί μια οικογένεια συναρτήσεων που έχουν τη χαρακτηριστική γραφική παράσταση σχήματος πεπλατυσμένου "S". Μια σιγμοειδής συνάρτηση είναι η λογιστική (logistic function), η οποία έχει τη γενική μορφή:

  f(u) = $\frac{1}{1+e^{-au}}$

  , όπου α είναι ένας συντελεστής που ρυθμίζει την ταχύτητα μετάβασης μεταξύ των δύο ασυμπτωτικών τιμών. Η γραφική της αναπαράσταση φαίνεται παρακάτω:



Σχήμα 2.12: Λογιστική συνάρτηση.

Άλλες συναρτήσεις αυτής της οικογένειας είναι η αντίστροφη εφαπτομένη (arctan) και η υπερβολική εφαπτομένη, οι οποίες δίνονται από τις παρακάτω σχέσεις αντίστοιχα:

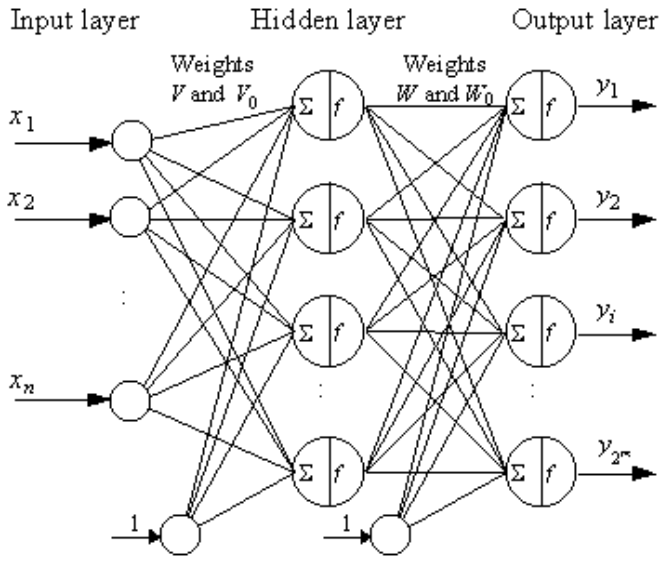
f(u) = $\frac{2}{\pi} tan^{-1}(au)$ και f(u) = tanh (au).

Οι σιγμοειδείς συναρτήσεις είναι οι πλέον χρησιμοποιούμενες στην κατασκευή των τεχνητών νευρωνικών δικτύων. Είναι συνεχείς και παραγωγίσιμες σε όλο το πεδίο ορισμού και περιορίζουν την είσοδο μεταξύ των τιμών 0 και 1 (ή -1 και 1). Η μεγάλη τους αξία έγκειται στο γεγονός ότι η παράγωγός τους έχει σχήμα "καμπάνα" και αυτό τους δίνει τη δυνατότητα να καταστέλουν τις μεγάλες τιμές και ταυτόχρονα να δίνουν πολύ ικανοποιητικές εξόδους για μικρές τιμές εισόδου.

Όλες οι παραπάνω συναρτήσεις είναι μη γραμμικές, ιδιότητα απαραίτητη για την κατασκευή τεχνητών νευρωνικών δικτύων και τη μοντελοποίηση μη γραμμικών φαινομένων. Αντίθετα, μια γραμμική συνάρτηση ενεργοποίησης θα παρήγαγε πάντα έξοδο ανάλογη της εισόδου, ιδιότητα αρκετά περιοριστική.

2.4.3 Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks)

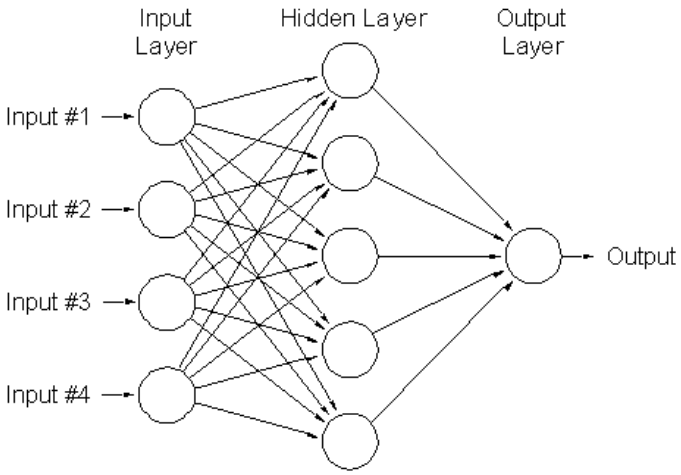
Τα τεχνητά νευρωνικά δίκτυα (artificial neural networks) είναι συστήματα που αποτελούνται από πολλούς τεχνητούς νευρώνες, οργανωμένους με παρόμοιο τρόπο με αυτόν του ανθρώπινου εγκεφάλου. Συνήθως οι τεχνητοί νευρώνες είναι οργανωμένοι σε έναν αριθμό στρωμάτων ή επιπέδων (layers). Το πρώτο επίπεδο ονομάζεται επίπεδο εισόδου (input layer) και σε αυτό πραγματοποιείται η εισαγωγή των δεδομένων. Τα στοιχειά του δεν επιτελούν τους υπολογισμούς που περιγράψαμε στην προηγούμενη ενότητα, επομένως ουσιαστικά δεν είναι νευρώνες. Μετά το επίπεδο εισόδου μπορούν να ακολουθούν ένα ή περισσότερα ενδιάμεσα ή κρυφά επίπεδα (hidden layers) και στο τέλος υπάρχει το επίπεδο εξόδου (output layer). Αν ένα τεχνητό νευρωνικό δίκτυο δε διαθέτει κανένα κρυφό επίπεδο ονομάζεται μονοστρωματικό, ενώ αν διαθέτει ένα ή περισσότερα κρυφά επίπεδα ονομάζεται πολυστρωματικό (multi-layer).



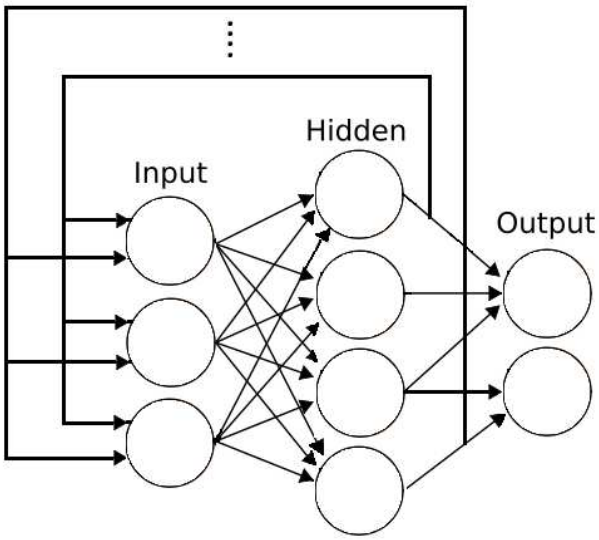
Σχήμα 2.13: Ένα τεχνητό νευρωνικό δίκτυο.

Όταν οι νευρώνες συνδέονται με όλους τους υπόλοιπους ονομάζονται πλήρως συνδεδεμένοι (fully connected). Σε κάθε άλλη περίπτωση ονομάζονται μερικώς συνδεδεμένοι (partially connected). Μία περίπτωση μερικής διασύνδεσης είναι όταν οι νευρώνες κάθε επιπέδου είναι πλήρως συνδεδεμένοι με αυτούς του επόμενου επιπέδου και όχι με προηγούμενου. Όταν συμβαίνει αυτό, όταν δηλαδή η πληροφορία ρέει μόνο προς το επίπεδο εξόδου, τα τεχνητά νευρωνικά δίκτυα χαρακτηρίζονται ως πρόσθιας τροφοδότησης (feedforward). Αντιθέτως, αν έχουμε την έξοδο νευρώνων συνδεδεμένη με νευρώνες προηγούμενου επιπέδου, ή αν έχουμε συνδέσεις μεταξύ νευρώνων του ίδιου επιπέδου, χαρακτηρίζονται ως δίκτυα με ανατροφοδότηση (feedback ή recurrent). Η πιο απλή μορφή τεχνητών νευρωνικών δικτύων είναι τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης. Αξίζει να αναφερθεί ότι τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης πολλών στρωμάτων που ενσωματώνουν τη σιγμοειδή συνάρτηση αποδεικνύεται ότι έχουν πολλές δυνατότητες αναπαράστασης συναρτήσεων. Το βασικό Θεώρημα λέει, συγκεκριμένα, ότι τα δίκτυα αυτής της μορφής μπορούν να προσεγγίσουν μία οποιαδήποτε ομαλή συνάρτηση, όσο πιο κοντά επιθυμούμε [Δ07]. Για αυτό το λόγο, τα δίκτυα αυτά καλούνται "Καθολικοί Προσεγγιστές" ("Universal Approximators"). Η σημασία του παραπάνω ισχυρισμού και της απόδειξής του είναι προφανής. Εξαιρετικό ενδιαφέρον παρουσιάζει επίσης το γεγονός ότι αρκούν μόλις δύο επίπεδα (χωρίς να μετράμε το επίπεδο εισόδου).

Κατά την υλοποίηση των νευρωνικών δικτύων κάθε είδους προκύπτουν δύο βασικά θέματα: η μάθηση και η τοπολογία του. Επειδή τα δύο αυτά θέματα συνδέονται μεταξύ τους και είναι άμεσα συνυφασμένα με τη σωστή συμπεριφορά του προς υλοποίηση νευρωνικού δικτύου, αναλύονται μαζί στην επόμενη ενότητα.



Σχήμα 2.14: Τεχνητό Νευρωνικό Δίκτυο πρόσθιας τροφοδότησης.



Σχήμα 2.15: Τεχνητό Νευρωνικό Δίκτυο με ανατροφοδότηση.

2.4.4 Μάθηση και Τοπολογία

Μάθηση (learning) ή εκπαίδευση (training) ονομάζεται η διαδικασία τροποποίησης της τιμής των βαρών, έτσι ώστε το νευρωνικό δίκτυο να έχει την επιθυμητή συμπεριφορά. Υπάρχουν τρεις κατηγορίες μάθησης:

- Μάθηση με επίβλεψη (supervised learning). Δίνονται στο δίκτυο μερικά ζευγάρια εισόδου – επιθυμητής εξόδου και αυτό παράγει με την επίδραση των τρεχόντων βαρών μία έξοδο που διαφέρει αρχικά από την επιθυμητή. Η διαφορά αυτή ονομάζεται σφάλμα (error) και με τη χρήση ενός από τους πολλούς αλγόριθμους εκπαίδευσης γίνεται αναπροσαρμογή των βαρών με σκοπό τη μείωση αυτής της διαφοράς.
- Βαθμολογημένη Μάθηση (graded learning). Η περίπτωση αυτή μοιάζει με την προηγούμενη, μόνο που στη βαθμολογημένη μάθηση η έξοδος χαρακτηρίζεται ως "καλή" ή

"κακή" βάσει μιας κλίμακας και τα βάρη αναπροσαρμόζονται βάσει αυτού του χαρακτηρισμού.

- Μάθηση χωρίς επίβλεψη (unsupervised learning). Το είδος αυτό αναφέρεται στις περιπτώσεις όπου δεν είναι γνωστά τα διανύσματα εξόδου. Το νευρωνικό δίκτυο καλείται να "μάθει" κάποια χαρακτηριστικά, ουσιαστικά να μάθει να κατηγοριοποιεί τα δεδομένα εισόδου. Η μάθηση αυτή βασίζεται στην ικανότητα του δικτύου να αυτό-οργανώνεται με βάση την είσοδο και τελικά σε ένα συγκεκριμένο σύνολο εισόδων να επιδρά ένας συγκεκριμένος νευρώνας.

Συνήθως χρησιμοποιείται η Μάθηση υπό επίβλεψη. Για την υλοποίησή της υπάρχουν διάφοροι αλγόριθμοι. Αναφέρουμε συνοπτικά μερικούς:

- Κανόνας Δέλτα (Delta Rule Learning). Η διαφορά μεταξύ τρέχουσας και επιθυμητής εξόδου ελαχιστοποιείται μέσω μιας διαδικασίας ελαχίστων τετραγώνων.
- Αλγόριθμος ανάστροφης μετάδοσης λάθους (Backpropagation). Η μεταβολή των βαρών γίνεται υπολογίζοντας τη συνεισφορά κάθε βάρους στο τελικό σφάλμα, προχωρώντας "προς τα πίσω" στο δίκτυο. Στον αλγόριθμο αυτό θα αναφερθούμε εκτενέστερα σε επόμενη παράγραφο.
- Ανταγωνιστική μάθηση (Competitive Learning). Οι τεχνητοί νευρώνες "ανταγωνίζονται" κατά κάποιο τρόπο ο ένας τον άλλο και κάθε φορά τροποποιεί τα βάρη του αυτός που έχει τη μεγαλύτερη απόκριση για δοθείσα έξοδο.
- Τυχαία μάθηση (Random Learning). Οι μεταβολές στα βάρη εισάγονται τυχαία και οι μεταβολές απορρίπτονται ή υιοθετούνται ανάλογα με το αν η έξοδος βελτιώνεται ή όχι βάσει κάποιων καθορισμένων από τον χρήστη κριτηρίων.

Ο στόχος της εκπαίδευσης του νευρωνικού είναι η σωστή συμπεριφορά σε μελλοντικές εισόδους. Δύο φαινόμενα που πρέπει να αποφευχθούν είναι η ατελής μάθηση ή υποπροσαρμογή (underfitting) και η υπερπροσαρμογή (overfitting). Η υποπροσαρμογή αναφέρεται στις περιπτώσεις που το δίκτυο δεν είναι αρκετά περίπλοκο, έτσι ώστε να μοντελοποιήσει σωστά το πρόβλημα και έτσι αδυνατεί να χειριστεί σωστά στις εισόδους. Υπερπροσαρμογή συμβαίνει όταν αντιθέτως το δίκτυο είναι περισσότερο πολύπλοκο από ότι πρέπει και έχει "μάθει" υπερβολικά καλά τα δεδομένα εκπαίδευσης μαζί με το θόρυβο που ενδεχομένως να περιέχουν. Αυτό έχει σαν αποτέλεσμα να έχει "απομνημονεύσει" τα δεδομένα εισόδου, να έχει χάσει την ικανότητα γενίκευσης και να δίνει λανθασμένες προβλέψεις για άλλα δεδομένα εισόδου.

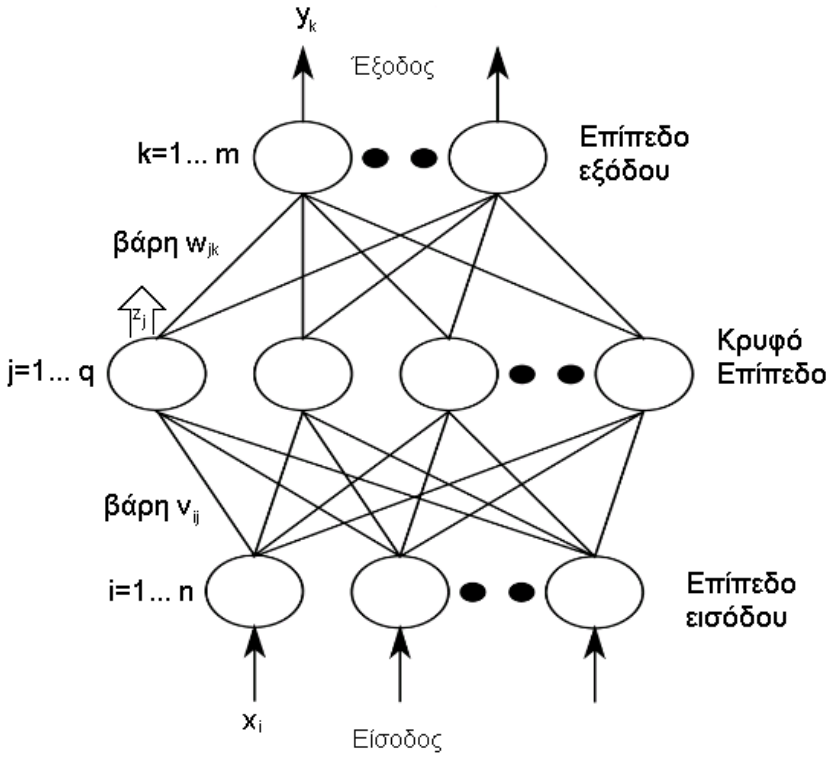
Για την αποφυγή των παραπάνω φαινομένων απαιτούνται, πρώτα από όλα, αρκετά δεδομένα εκπαίδευσης. Επίσης, είναι σημαντικό τα δεδομένα εκπαίδευσης να καλύπτουν το φάσμα των τιμών εισόδων που πρόκειται να χρησιμοποιηθούν μελλοντικά στο δίκτυο. Τα τεχνητά νευρωνικά δίκτυα μπορούν να γενικεύουν σωστά μέσα στο φάσμα τιμών στο οποίο έχουν εκπαιδευτεί, αλλά δεν έχουν την ικανότητα να επεκτείνουν την πρόβλεψη τους πέρα από αυτό το εύρος με την ίδια ακρίβεια.

Σημαντικό ρόλο για την αποφυγή φαινομένων όπως η υπερπροσαρμογή και η υποπροσαρμογή παίζει και η τοπολογία του δικτύου. Με τον όρο τοπολογία εννοούμε πόσα κρυφά επίπεδα θα έχει το δίκτυο, πόσους νευρώνες θα διαθέτει το κάθε επίπεδο και πως θα είναι συνδεδεμένοι οι νευρώνες μεταξύ τους. Δεν υπάρχει κάποιος συγκεκριμένος κανόνας για τον προσδιορισμό αυτών των παραμέτρων και συνήθως απαιτείται πειραματισμός για την εύρεση των παραμέτρων που οδηγούν στο βέλτιστο αποτέλεσμα. Όσον αφορά στη συνδεσμολογία μεταξύ των νευρώνων, η πιο συνηθισμένη περίπτωση είναι ο κάθε νευρώνας να συνδέεται με όλους τους νευρώνες του επόμενου επιπέδου. Σχετικά με το πλήθος των νευρώνων και των κρυφών επιπέδων πρέπει να σημειώσουμε ότι περισσότεροι νευρώνες και περισσότερα κρυφά επίπεδα απαιτούν περισσότερους υπολογισμούς και έχουν την τάση να υπερπροσαρμόζονται στα δεδομένα εκπαίδευσης, αλλά επιτρέπουν την επίλυση πιο περίπλοκων προβλημάτων. Ο προσδιορισμός του πλήθους των νευρώνων εισόδου και εξόδου είναι αρκετά απλός και εξαρτάται από το ίδιο το πρόβλημα. Ένα πολύ απλοϊκό παράδειγμα είναι η ταξινόμηση ενός αριθμού κρασιών σε "καλά" ή "κακά" βάσει κάποιων χαρακτηριστικών τους. Το πλήθος των εισόδων θα ισούται με το πλήθος των χαρακτηριστικών που εξετάζονται (οξύτητα, επίγευση, ένταση, χρώμα, κ.α.), και για έξοδο θα έχουμε ένα νευρώνα, με την τιμή "0" να συμβολίζει τη μία κλάση ("κακό") και με "1" τη δεύτερη ("καλό").

Σχετικά όμως με το πλήθος των νευρώνων των κρυφών επιπέδων, δεν υπάρχει τέτοια σαφής καθοδήγηση. Υπάρχουν όμως κάποιοι εμπειρικοί κανόνες που βάζουν κάποια όρια στην αρχιτεκτονική των δικτύων. Για παράδειγμα, στα προβλήματα ταξινόμησης θεωρείται γενικά καλό ο αριθμός των νευρώνων των κρυφών επιπέδων να είναι μικρότερος από τον αριθμό των διανυσμάτων εκπαίδευσης, έτσι ώστε να αποφεύγεται η αποκλειστική συσχέτιση ενός νευρώνα με ένα διάνυσμα εκπαίδευσης, δηλαδή η "απομνημόνευση". Ένας εμπειρικός κανόνας είναι ο εξής [BKB+06]: Σε προβλήματα με δεδομένα εκπαίδευσης που περιλαμβάνουν θόρυβο, καλό είναι τα δεδομένα εκπαίδευσης να είναι τουλάχιστον 30 φορές περισσότερα από το συνολικό πλήθος των βαρών που περιλαμβάνει το δίκτυο. Αν τα δεδομένα εκπαίδευσης δεν περιλαμβάνουν θόρυβο, αρκούν τουλάχιστον 5 φορές περισσότερα δεδομένα εκπαίδευσης. Η μείωση του αριθμού των βαρών για την έλλειψη ικανοποιητικού αριθμού δεδομένων εκπαίδευσης, προφανώς δεν αποτελεί καλή λύση. Ένας άλλος εμπειρικός κανόνας που συναντήσαμε στη βιβλιογραφία είναι ο εξής [GG10]: Ο συνολικός αριθμός των παραμέτρων του δικτύου πρέπει να είναι μικρότερος από το μισό του πλήθους των δεδομένων εκπαίδευσης. Παρατηρούμε ότι δεν υπάρχει πλήρης συμφωνία μεταξύ των δύο αυτών εμπειρικών κανόνων και οφείλουμε να τονίσουμε και πάλι ότι δεν υπάρχει γενική λύση και ότι απαιτείται πειραματισμός για την εύρεση της χρυσής τομής μεταξύ παραμέτρων και δεδομένων εκπαίδευσης.

2.4.5 Ανάστροφη Μετάδοση Λάθους (Back propagation)

Θεωρήσαμε σκόπιμο να συμπεριλάβουμε στην ανάλυσή μας την Ανάστροφη Μετάδοση Λάθους (Back propagation), εφόσον αποτελεί τον πιο γνωστό αλγόριθμο εκπαίδευσης νευρωνικών δικτύων πολλών επιπέδων.



Σχήμα 2.16: Τεχνητό Νευρωνικό Δίκτυο πρόσθιας τροφοδότησης.

Η διαδικασία ενός κύκλου εκπαίδευσης ενός τεχνητού νευρωνικού δικτύου πολλών στρωμάτων (Multi-Layer Perceptron – MLP) περιλαμβάνει δύο στάδια. Αρχικά εισάγονται στην είσοδο τα δεδομένα κάποιου διανύσματος εκπαίδευσης. Οι νευρώνες του επιπέδου εισόδου περνάνε την είσοδο στο επόμενο, κρυφό επίπεδο. Η διαδικασία αυτή επαναλαμβάνεται και στα επόμενα επίπεδα, μέχρι το επίπεδο εξόδου. Το στάδιο αυτό ονομάζεται πρόσθιο πέρασμα (forward pass). Οι υπολογισμοί ξεκινάνε με τυχαίες τιμές βαρών. Η είσοδος ενός κρυφού νευρώνα j δίνεται από τη σχέση:

$$input_j = \sum_{i=1}^{n} v_{ij} x_i$$

όπου v_{ij} είναι το βάρος της σύνδεσης μεταξύ των νευρώνων i και j, και x_i το σήμα εισόδου του νευρώνα i. Η έξοδος του ίδιου νευρώνα είναι:

$$z_j = f(input_j) = f(\sum_{i=1}^{n} v_{ij} x_i)$$

Η έξοδος αυτή αποστέλλεται σε όλους τους νευρώνες του επόμενου επιπέδου. Η συνάρτηση f, όπως έχουμε ήδη αναφέρει, είναι συνήθως μία σιγμοειδής συνάρτηση. Εδώ πρέπει να αναφέρουμε ακόμα

ότι τα δίκτυα που εκπαιδεύονται με Ανάστροφη Μετάδοση Λάθους, θα πρέπει να διαθέτουν μονότονα αύξουσα και παραγωγίσιμη συνάρτηση ενεργοποίησης σε όλο το φάσμα των τιμών εισόδου, χαρακτηριστικά που οι σιγμοειδείς συναρτήσεις έχουν. Η παραπάνω σχέση εφαρμόζεται σε όλους τους νευρώνες, εκτός από αυτούς του επιπέδου εισόδου, που απλά μεταφέρουν τα δεδομένα στο επόμενο επίπεδο. Για τους νευρώνες του επιπέδου εξόδου έχουμε αντίστοιχα:

$$input_k = \sum_{j=1}^{q} w_{jk} z_j$$

$$y_k = f(input_k) = f(\sum_{j=1}^{q} w_{jk} z_j)$$

Για ένα νευρώνα εξόδου k είναι γνωστό το επιθυμητό αποτέλεσμα t_k. Αυτό συγκρίνεται με την πραγματική έξοδο και παράγεται το σφάλμα: $e_k = t_k - y_k$. Τα βάρη επαναπροσδιορίζονται με σκοπό την ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος: $E = \frac{1}{2} e_k^2$, βάσει της σχέσης:

$w_{jk}(νέο) = w_{jk}(παλιό) + \Delta w_{jk}$ με $\Delta w_{jk} = d \cdot \delta_k \cdot z_j$

, όπου d είναι μία σταθερά που ρυθμίζει το ρυθμό μεταβολής των βαρών και ονομάζεται ρυθμός μάθησης (learning rate), z_j η πραγματική έξοδος του νευρώνα j και δ_k ο ρυθμός μεταβολής του μέσου τετραγωνικού σφάλματος ως προς την είσοδο στο νευρώνα k. Ο τελευταίος προκύπτει από παραγώγιση του E και δίνεται από τη σχέση:

$$\delta_k = (t_k - y_k) f'(input_k)$$

Αντίστοιχα, επαναπροσδιορίζονται και οι νευρώνες των κρυφών επιπέδων, μέσω των σχέσεων:

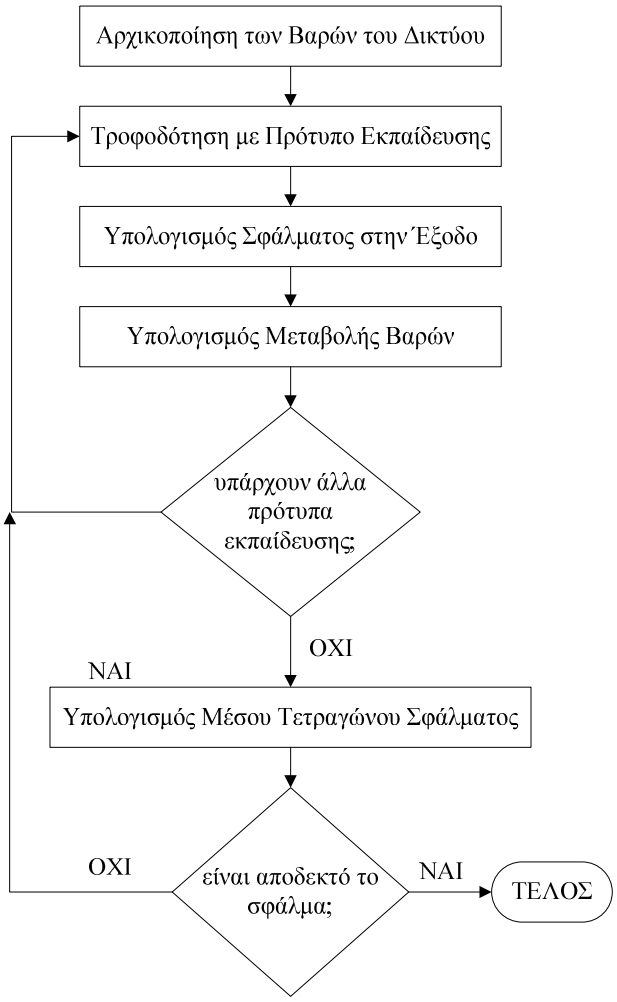
$$\Delta v_{ij} = d \cdot \delta_j \cdot x_i$$

Και

$$\delta_j = f'(input_j) \sum_{k=1}^{m} w_{jk} \delta_k$$

Με αυτόν τον τρόπο μπορούν να αναπροσαρμοστούν αρχικά οι νέες τιμές των βαρών που συνδέουν το επίπεδο εξόδου με το προηγούμενο κρυφό επίπεδο, στη συνέχεια να επαναπροσδιοριστούν τα βάρη που συνδέουν το κρυφό επίπεδο με το προηγούμενό του, κ.ο.κ. μέχρι να φτάσουμε στο επίπεδο εισόδου. Αυτή η διαδικασία αναπροσαρμογής των βαρών συνιστά το δεύτερο στάδιο της εκπαίδευσης και ονομάζεται ανάστροφο πέρασμα (backward pass) ή ανάστροφη μετάδοση (back propagation), ακριβώς επειδή "πάμε προς τα πίσω" στο δίκτυο και διορθώνουμε τα βάρη. Τελικά, η Ανάστροφη Μετάδοση Λάθους είναι μια διαδικασία βελτιστοποίησης, η οποία ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα E μεταξύ της εξόδου του δικτύου και της επιθυμητής εξόδου για p διανύσματα δεδομένων εκπαίδευσης: $E = \frac{1}{p} \sum_p \sum_{k=1}^{m} (t_{k(p)} - y_{k(p)})^2$, κάνοντας κάθε φορά τοπικές

αλλαγές. Η συνολική διαδικασία της Ανάστροφης Μετάδοσης Λάθους φαίνεται στο παρακάτω διάγραμμα:



Σχήμα 2.17: Η διαδικασία εκπαίδευσης Ανάστροφης Μετάδοσης Λάθους.

Συχνά σαν κριτήριο τερματισμού της διαδικασίας της εκπαίδευσης χρησιμοποιείται ο υπολογισμός του μέσου τετραγωνικού σφάλματος ενός άλλου σετ δεδομένων, των δεδομένων αξιολόγησης (validation data) και όχι το σφάλμα των δεδομένων εκπαίδευσης. Η πρακτική αυτή βοηθάει την αποφυγή φαινομένων υπερπροσαρμογής, καθώς διατηρώντας σαν κριτήριο τερματισμού το σφάλμα των δεδομένων εκπαίδευσης, υπάρχει ο κίνδυνος απομνημόνευσης των δεδομένων αυτών από το δίκτυο.

2.4.6 Ιδιότητες των Νευρωνικών Δικτύων

Τα τεχνητά νευρωνικά δίκτυα παρουσιάζουν γενικά τις παρακάτω ιδιότητες:

- Μη γραμμικότητα. Αποτελεί πολύ σημαντική ιδιότητα όταν το προς μελέτη πρόβλημα είναι μη γραμμικής φύσεως, κάτι που ισχύει για τα περισσότερα προβλήματα του πραγματικού κόσμου.

- Ικανότητα εκμάθησης μέσω παραδειγμάτων (learn by example). Η ιδιότητα αυτή είναι πολύ σημαντική για τις χιλιάδες περιπτώσεις των προβλημάτων του πραγματικού κόσμου στα οποία δεν είναι γνωστή η σχέση εισόδου-εξόδου. Το σύστημα έχει τη δυνατότητα να "εκπαιδευτεί" πάνω στα παραδείγματα και να μπορεί να προβλέπει την έξοδο σε μελλοντικές εισόδους. Η ιδιότητα αυτή καθιστά τα νευρωνικά δίκτυα πολύ χρήσιμα στον τομέα της ταξινόμησης.
- Ανοχή σε σφάλματα. Αυτό σημαίνει ότι αν κάποιοι νευρώνες ή/και συνδέσεις τεθούν εκτός λειτουργίας, αυτό δεν είναι ικανό να διαταράξει τη λειτουργία και την αποτελεσματικότητα του συστήματος.
- Ικανότητα αναγνώρισης προτύπων (pattern recognition). Από τη στιγμή που θα εκπαιδευτούν τα νευρωνικά δίκτυα, δεν επηρεάζονται από ελλιπή ή/και με θόρυβο δεδομένα.
- Ικανότητα θεώρησής τους ως κατανεμημένη μνήμη (distributed memory) και ως μνήμη συσχέτισης (associative memory). Ως κατανεμημένη γιατί η πληροφορία που περιέχουν δεν βρίσκεται αποθηκευμένη σε ένα σημείο τους, αλλά είναι διάχυτη σε όλα τα βάρη που τα απαρτίζουν. Ως μνήμη συσχέτισης γιατί δεν αποθηκεύουν την πληροφορία με τον παραδοσιακό τρόπο, αλλά μέσω συσχετίσεων που προκύπτουν από τα δεδομένα εκπαίδευσης (training data). Με τον τρόπο αυτό, η ανάκληση της πληροφορίας δε γίνεται με τη διεύθυνση της πληροφορίας, αλλά με το περιεχόμενό της, όπως συμβαίνει και στον ανθρώπινο εγκέφαλο. Η ιδιότητα αυτή συνδέεται με την προηγούμενη και καθιστά ικανά τα νευρωνικά δίκτυα να παράγουν τη σωστή έξοδο, ακόμα και όταν η είσοδος περιλαμβάνει θόρυβο ή είναι ελλιπής.

  Για παράδειγμα, αν μία μνήμη έχει αποθηκευμένα τα ονόματα Γιάννης, Νίκος, Αλεξάνδρα, Κατερίνα και Χριστίνα, αν δοθεί το λανθασμένο όνομα Αλξνανρα, μία μνήμη συσχέτισης είναι σε θέση να ανακαλέσει το σωστό όνομα Αλεξάνδρα.

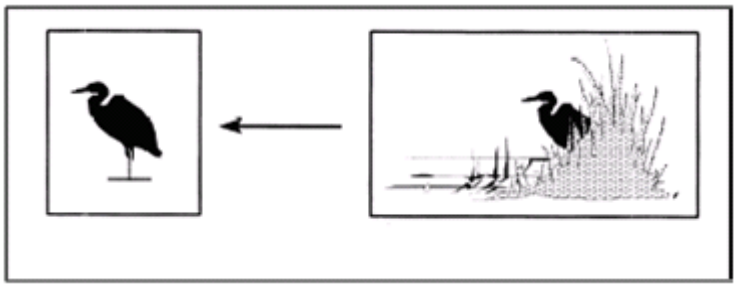
2.4.7 Πρακτικές εφαρμογές των Νευρωνικών Δικτύων

Γενικά τα νευρωνικά δίκτυα γνωρίζουν μεγάλη αποδοχή σε προβλήματα που δεν είναι πλήρως κατανοητά και που περιλαμβάνουν μη προβλέψιμες λειτουργίες. Μερικές γενικές κατηγορίες τέτοιων προβλημάτων είναι η ταξινόμηση (classification), αναγνώριση (recognition), αποτίμηση (assessment), πρόβλεψη (prediction). Παραθέτουμε μερικές κατηγορίες τέτοιων προβλημάτων:

- Ταξινόμηση (Classification). Στην κατηγορία αυτή έχουμε αναφερθεί στην Ενότητα 2.1.
- Ομαδοποίηση (Clustering). Ομαδοποίηση ονομάζεται η ταξινόμηση για τις περιπτώσεις που δε διαθέτουμε κάποια δεδομένα, για τα οποία να γνωρίζουμε ποια είναι η έξοδος. Είναι δηλαδή η μη επιβλεπόμενη (unsupervised) μορφή της ταξινόμησης. Στις περιπτώσεις αυτές,

το σύστημα καλείται να "ανακαλύψει" τους κανόνες ταξινόμησης, ομαδοποιώντας τα δεδομένα βάσει εγγενών διαφορών και ομοιοτήτων.

- Αναγνώριση (recognition) – Συσχέτιση (Association). Τα τεχνητά νευρωνικά δίκτυα μπορούν να εκπαιδευτούν στην "απομνημόνευση" κάποιων προτύπων, έτσι ώστε αν τους δοθεί σαν είσοδος κάποια αλλοιωμένη εκδοχή τους, να μπορούν να τη συσχετίσουν με ό,τι κοντινότερο υπάρχει στη μνήμη τους και να το δώσουν σαν έξοδο, ή αλλιώς να το αναγνωρίσουν. Η τεχνική αυτή είναι χρήσιμη σε εφαρμογές συμπλήρωσης και αναγνώρισης εικόνας. Παρακάτω φαίνεται ένα σχηματικό παράδειγμα:



Σχήμα 2.18: Αναγνώριση προτύπου, συμπλήρωση εικόνας

- Προσέγγιση συναρτήσεων (function approximation). Στην περίπτωση αυτή, στόχος είναι η προσέγγιση μίας μη γραμμικής συνάρτησης της μορφής $\vec{y} = f(\vec{x})$ από μία συνάρτηση F. Το δίκτυο εκπαιδεύεται αρχικά με επαρκή ζεύγη $\{\vec{x}, \vec{y}\}$, και στη συνέχεια είναι σε θέση να προβλέψει την έξοδο για άλλες εισόδους, προσεγγίζοντας έτσι τη συνάρτηση.
- Αποτίμηση (assessment). Η αποτίμηση αφορά στη λήψη της καλύτερης απόφασης από ένα σύνολο αποφάσεων βάσει κάποιων κριτηρίων. Για παράδειγμα, η αποτίμηση της καταλληλότητας ενός αυτοκινήτου, βάσει των αναγκών του πελάτη (μέγεθος, ασφάλεια, κατανάλωση, καυσίμου, κ.α.). Εφαρμογές αποτίμησης μπορούν συχνά να αναχθούν στην κατηγορία της ταξινόμησης.
- Πρόβλεψη (prediction). Τα τεχνητά νευρωνικά δίκτυα μετά από κατάλληλη εκπαίδευση είναι σε θέση να εκτιμήσουν όσο το δυνατό πιο πιστά την εξέλιξη μιας διεργασίας, βασιζόμενα σε προηγούμενες παρατηρήσεις.

Παραθέτουμε μερικές ανθρώπινες δραστηριότητες στις οποίες βρίσκουν εφαρμογή τα νευρωνικά δίκτυα:

- Ιατρική: κατηγοριοποίηση ιατρικών εικόνων που προέρχονται από διάφορα απεικονιστικά όργανα, επισήμανση "ύποπτων" περιοχών.
- Τραπεζικός τομέας: γνησιότητα υπογραφής και διαπίστωση απάτης στη χρήση πιστωτικών καρτών.
- Τηλεπικοινωνίες, πληροφορική, κ.α.: αναγνώριση ήχου, εικόνας και γραπτού χειρόγραφου ή τυπωμένου κειμένου.

- Μηχανολογία, Παραγωγή: παρακολούθηση, επιθεώρηση και έλεγχος προϊόντων, ανίχνευση ανωμαλιών.
- Οικονομία/Επιχειρήσεις: πρόβλεψη τιμών μετοχών, πρόβλεψη πωλήσεων, ανάλυση χρηματοοικονομικών διεργασιών.
- Μετεωρολογία: πρόβλεψη καιρού.
- Ασφάλεια: εντοπισμός κίνησης, ταύτιση δακτυλικών αποτυπωμάτων.
- Έρευνα: μοντελοποίηση περίπλοκων προβλημάτων και φαινομένων και πρόβλεψη μελλοντικής συμπεριφοράς.

2.5 Αμοιβαία Πληροφορία

Ακολουθεί μία συνοπτική ανάλυση μερικών εννοιών της θεωρίας πληροφορίας, που πρόκειται να χρησιμοποιηθούν στη μεθοδολογία που θα αναπτυχθεί.

Έστω η διακριτή τυχαία μεταβλητή (τ.μ.) $X = \{x_k | k=1,\ldots,n\}$ και $p_k = P(x_k)$ η πιθανότητα η μεταβλητή X να πάρει την τιμή x_k (συνάρτηση πυκνότητας της X). Ως εντροπία (Entropy) ορίζεται η ποσότητα:

$$H(X) = -\sum_{k=1}^{n} P(x_k) log_2[P(x_k)]$$

Η εντροπία μετράει την αβεβαιότητα που σχετίζεται με μία τυχαία μεταβλητή. Βασίζεται στην κατανομή της πιθανότητάς της μεταβλητής και όχι στις τιμές που μπορεί να πάρει. Όσο μεγαλύτερη είναι η τιμή της εντροπίας μίας μεταβλητής (ή ενός χαρακτηριστικού) τόσο μεγαλύτερη αβεβαιότητα έχουμε για αυτή, δηλαδή τόσο λιγότερο αξιόπιστες είναι οι προβλέψεις μας για αυτή [Jak05]. Μία άλλη ερμηνεία της εντροπίας είναι: "Η μέση ποσότητα πληροφορίας που περιέχει μία τυχαία μεταβλητή" ή "Το ποσό πληροφορίας που λείπει, όταν δε γνωρίζουμε μία τυχαία μεταβλητή". Η έννοια της εντροπίας εισήχθη από τον Claude E.Shannon το 1948 με τη δημοσίευσή του "A Mathematical Theory of Communication".

Αν η βάση του λογαρίθμου είναι 2, τότε η εντροπία μετράται σε bits, αν είναι e, μετράται σε nat και αν είναι 10 σε δεκαδικά ψηφία. Στη συνέχεια, θα χρησιμοποιούμε σα βάση το 2 για τους ορισμούς μας, αλλά τα ίδια ισχύουν και για τις υπόλοιπες. Η μέγιστη τιμή της εντροπίας είναι 1 και η ελάχιστη 0. Όταν $p_k=0$, έχουμε $H(0) = \lim_{p\to 0^+} p \log p = 0$.

Ένα παράδειγμα για να καταλάβουμε την έννοια της εντροπίας είναι η ρίψη ενός κέρματος. Αν το κέρμα είναι αμερόληπτο, η εντροπία του ισούται με 1 bit. Η αβεβαιότητα που έχουμε για το αποτέλεσμά του είναι μέγιστη. Αλλιώς: κάθε μας πρόβλεψη δεν είναι αξιόπιστη. Αν το κέρμα δεν είναι αμερόληπτο, η αβεβαιότητά μας για το αποτέλεσμά του είναι μικρότερη, καθώς μπορούμε με μεγαλύτερη ακρίβεια να προβλέψουμε το αποτέλεσμά του θα στοιχηματίζουμε στο πιο συχνά εμφανιζόμενο αποτέλεσμα και είναι πιο πιθανό να το πετυχαίνουμε. Η εντροπία του δηλαδή

μειώθηκε. Μέγιστη τιμή εντροπίας έχουμε όταν κάθε ενδεχόμενο είναι ισοπίθανο, ενώ ελάχιστη όταν η μεταβλητή "συγκεντρώνεται" γύρω από μία τιμή.

Δε θα επεκτείνουμε τους ορισμούς μας σε συνεχείς μεταβλητές, καθώς κατά την ανάπτυξη της μεθοδολογίας μας πραγματοποιείται διακριτοποίηση κατά την προεπεξεργασία των δεδομένων μας και έτσι έχουμε να κάνουμε μόνο με διακριτές μεταβλητές.

Ορίζουμε στη συνέχεια την εντροπία υπό συνθήκη (Conditional Entropy). Η εντροπία υπό συνθήκη H(X/Y) μετράει την πληροφορία που παίρνουμε από μία μεταβλητή X, δεδομένου ότι η μεταβλητή Y είναι ήδη γνωστή, δηλαδή την αβεβαιότητα που μένει για το X, όταν το Y είναι ήδη γνωστό [Jak05]. Πιο συγκεκριμένα, αν H(X/Y=y) είναι η εντροπία της μεταβλητής X όταν η Y παίρνει την τιμή y, τότε η εντροπία H(X/Y) προκύπτει αθροίζοντας τα H(X/Y=y) για όλες τις δυνατές τιμές του Y. Έχουμε:

$$H(X/Y) = \sum_{y \in Y} P(y) H(X/Y = y) = -\sum_{y \in Y} P(y) \sum_{x \in X} P(x/y) \log [P(x/y)]$$

$$H(X/Y) = \sum_{x \in X, y \in Y} P(x,y) log P(x/y)$$

Τα x∈X και y∈Y χρησιμοποιούνται για να συμβολίσουμε όλες τις δυνατές τιμές που μπορούν να πάρουν τα X και Y. Η πιθανότητα P(x/y) ονομάζεται δεσμευμένη ή υπό συνθήκη πιθανότητα (Conditional Probability) και εκφράζει την πιθανότητα η μεταβλητή X να πάρει την τιμή x, δεδομένου ότι η Y παίρνει την τιμή y. Η πιθανότητα P(x,y) ονομάζεται από κοινού πιθανότητα (Joint Probability) και εκφράζει την πιθανότητα οι μεταβλητές X και Y να πάρουν ταυτόχρονα τις τιμές x και y αντίστοιχα.

Για παράδειγμα, αν ξέρουμε το χρώμα της δεξιάς μας κάλτσας, δε μαθαίνουμε τίποτα καινούριο με το να κοιτάξουμε το χρώμα της αριστερής μας κάλτσας, μιας και –συνήθως– είναι ίδιες. Η υπό συνθήκη εντροπία τους είναι 0. Αν πάλι οι μεταβλητές X και Y είναι πλήρως ανεξάρτητες μεταξύ τους, τότε H(X/Y)=H(X), αφού η αβεβαιότητα για τη X δε μειώνεται καθόλου λόγω της γνώσης της Y.

Η από κοινού εντροπία (Joint Entropy) μετράει την εντροπία της σύνθετης κατανομής των X και Y, δηλαδή την εντροπία των X και Y μαζί. Δίνεται από τη σχέση:

$$H(X,Y) = \sum_{x \in X, y \in Y} P(x,y) log P(x,y)$$

Αν τα X και Y είναι ανεξάρτητα, τότε η από κοινού εντροπία τους, δηλαδή η συνολική αβεβαιότητα για αυτά, θα ισούται με το άθροισμα των εντροπιών καθενός ξεχωριστά:

$$H(X,Y) = H(X) + H(Y)$$

Μπορούμε τώρα να προχωρήσουμε στην έννοια της αμοιβαίας πληροφορίας. Αμοιβαία Πληροφορία (Mutual Information) είναι η διαφορά της αβεβαιότητας που έχουμε για το X μείον της

αβεβαιότητας που έχουμε για το X όταν το Y είναι ήδη γνωστό. Ουσιαστικά μετράει πόση πληροφορία παίρνουμε από το Y για το X, δηλαδή το κέρδος πληροφορίας (information gain) που έχουμε για μία μεταβλητή από τη γνώση μίας άλλης. Δίνεται από τη σχέση:

$$I(X;Y) = \sum_{x}\sum_{y} P(X,Y) log_2 \frac{P(X,Y)}{P(X)P(Y)}$$

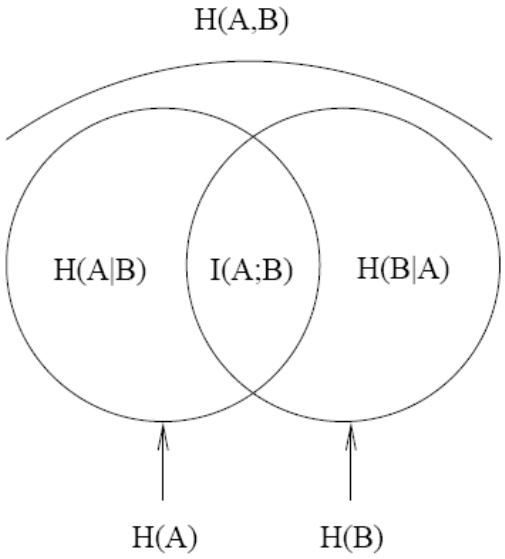
και μπορεί ισοδύναμα να εκφραστεί ως:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X/Y) = H(Y) - H(Y/X).$$

Όπως φαίνεται από τις παραπάνω σχέσεις, η αμοιβαία πληροφορία είναι συμμετρική, δηλαδή ισχύει ότι I(X;Y)=I(Y;X). Επίσης ισχύει ότι I(X;Y)≥0.

Η αμοιβαία πληροφορία ποσοτικοποιεί την πληροφορία που μοιράζονται δύο μεταβλητές, εκφράζει δηλαδή πόσο η γνώση της μίας μειώνει την αβεβαιότητά μας για την άλλη. Αν οι μεταβλητές είναι ανεξάρτητες, τότε η γνώση της X δε δίνει καμία πληροφορία για την Y και αντίστροφα και η αμοιβαία πληροφορία τους είναι ίση με 0. Αντίθετα, αν οι X και Y είναι ίδιες, τότε η γνώση της X προσδιορίζει πλήρως της Y και το αντίστροφο.

Παρακάτω φαίνεται μία γραφική αναπαράσταση της σχέσης των παραπάνω μεγεθών για δύο τυχαίες μεταβλητές A και B [JB04]:



Σχήμα 2.19: Γραφική αναπαράσταση των μεγεθών.

3

Πρόβλεψη εμφάνισης επιπλοκών του Σακχαρώδη Διαβήτη και Μέθοδοι Τεχνητής Νοημοσύνης

Στο κεφάλαιο αυτό παρουσιάζονται εργασίες που έχουν γίνει σε θεματικές περιοχές σχετικές με την παρούσα διπλωματική. Αρχικά, κάνουμε μία συγκριτική παρουσίαση εργασιών που αφορούν στην εύρεση παραγόντων σχετικών με το Σακχαρώδη Διαβήτη και τις διάφορες επιπλοκές του και στην πρόβλεψη της εμφάνισής τους. Στη συνέχεια, αναφερόμαστε σε εργασίες που χρησιμοποιούν ή μελετάνε τις μεθόδους Τεχνητής Νοημοσύνης που χρησιμοποιούνται και στη διπλωματική. Τέλος, γίνεται μία εισαγωγή στη μεθοδολογία που αναπτύσσεται στην παρούσα διπλωματική και που παρουσιάζεται διεξοδικά στο επόμενο κεφάλαιο, επισημαίνοντας τους γενικούς στόχους της, τη λογική της και τον τρόπο με τον οποίο φιλοδοξεί να συνεισφέρει και να διαφοροποιηθεί από ό,τι έχει γίνει προς την ίδια κατεύθυνση.

3.1 Επιλογή παραγόντων που συνδέονται με την εμφάνιση επιπλοκών του Διαβήτη

Όπως έχουμε ήδη αναφέρει, ο μηχανισμός πυροδότησης του Διαβήτη και των διαφόρων επιπλοκών με τις οποίες σχετίζεται δεν έχει διαλευκανθεί πλήρως. Πρόκειται, επιπλέον, για μη ιάσιμες καταστάσεις, για τις οποίες δεν υπάρχουν εγγυημένες μέθοδοι πρόληψης. Με στόχο την άρση των παραπάνω, την εύρεση αποτελεσματικών μεθόδων πρόληψης και, ιδανικά, το ενδεχόμενο

ανακάλυψης μίας θεραπείας κατά της νόσου, απαραίτητο βήμα είναι η εύρεση των παραγόντων που ενθαρρύνουν την εμφάνιση των επιπλοκών. Στην κατεύθυνση αυτή έχουν πραγματοποιηθεί διάφορες μελέτες, βασιζόμενες, κυρίως, σε στατιστικές μεθόδους. Αναφερόμαστε κυρίως σε μελέτες σχετικές με την εμφάνιση της διαβητικής αμφιβληστροειδοπάθειας, καθώς η συγκεκριμένη επιπλοκή αποτελεί το κύριο αντικείμενο της διπλωματικής.

Μία από τις πρώτες πληθυσμιακές μελέτες που έχουν διεξαχθεί σχετικά με την εμφάνιση διαβητικής αμφιβληστροειδοπάθειας είναι αυτή που πραγματοποιήθηκε στο νότιο Wisconsin σε άτομα που διαγνώστηκαν ως διαβητικά πριν την ηλικία των 30 χρόνων και λάμβαναν ινσουλίνη εξωγενώς. Πρόκειται για την επιδημιολογική μελέτη "The Wisconsin Epidemiologic Study of Diabetic Retinopathy IX" [KKM+89]. Από τα 217 άτομα που δεν είχαν αμφιβληστροειδοπάθεια στην πρώτη εξέταση, τα 160 (59%) την είχαν εμφανίσει όταν επανεξετάστηκαν μετά από 4 χρόνια. Επίσης, από τους 713 που δεν είχαν παραγωγική διαβητική αμφιβληστροειδοπάθεια, οι 75 (11%) την παρουσίασαν. Γενικά, επιδείνωση της ασθένειας παρατηρήθηκε στο 41% του πληθυσμού. Τα παραπάνω στοιχειά φανερώνουν μία σύνδεση μεταξύ της αμφιβληστροειδοπάθειας και της διάρκειας του διαβήτη. Δεν παρατηρήθηκε σημαντική διαφορά στην ανάπτυξη αμφιβληστροειδοπάθειας μεταξύ ανδρών και γυναικών. Οι γυναίκες είχαν ελαφρώς μεγαλύτερο ρυθμό εμφάνισης αμφιβληστροειδοπάθειας και οι άντρες μεγαλύτερο ρυθμό εμφάνισης παραγωγικής αμφιβληστροειδοπάθειας (proliferative retinopathy), χωρίς όμως οι διαφορές αυτές να είναι στατιστικά σημαντικές. Επομένως, το φύλο δε φαίνεται να συνδέεται με την εμφάνιση αμφιβληστροειδοπάθειας.

Η εμφάνιση του οιδήματος της ωχράς κηλίδας (macular edema) και οι παράγοντες που συνδέονται με αυτό εξετάζονται πιο πρόσφατα στην εργασία "The Wisconsin Epidemiologic Study of Diabetic Retinopathy XXIII" [KKL+09]. Σε αυτή εξετάστηκαν 955 ινσουλινοεξαρτώμενα άτομα με διαβήτη τύπου Ι και με διάγνωση του Διαβήτη πριν τα 30 χρόνια. Η εμφάνιση του οιδήματος της ωχράς κηλίδας συνδέθηκε με τα προχωρημένα στάδια της αμφιβληστροειδοπάθειας, με το αρσενικό φύλο, τα υψηλά επίπεδα της γλυκοζυλιωμένης αιμοσφαιρίνης, την πρωτεϊνουρία, τη υψηλή συστολική και διαστολική αρτηριακή πίεση και το κάπνισμα. Για πολυμεταβλητή ανάλυση χρησιμοποιήθηκε ένα γενικευμένο γραμμικό μοντέλο (Generalized Linear Model - GLM), το οποίο αποτελεί μία γενίκευση της μεθόδου των ελαχίστων τετραγώνων (least squares regression). Αποτελέσματά της δείχνουν τη σύνδεση του οιδήματος με υψηλά επίπεδα της γλυκοζυλιωμένης αιμοσφαιρίνης και συστολικής πίεσης.

Στη μελέτη EURODIAB Prospective Complications [PSC+01] μελετώνται οι παράγοντες κινδύνου για την ανάπτυξη παραγωγικής αμφιβληστροειδοπάθειας. Η μελέτη περιλαμβάνει 31 κέντρα σε 16 ευρωπαϊκές χώρες και βασίζεται σε 3250 ασθενείς με διαβήτη τύπου Ι και χρησιμοποιεί μεθόδους γραμμικής παρεμβολής (linear regression). Σύμφωνα με τα αποτελέσματά της, ο γλυκαιμικός έλεγχος και η χρονική διάρκεια του διαβήτη αποτελούν παράγοντες ισχυρά συνδεδεμένους με την

εμφάνιση αμφιβληστροειδοπάθειας. Παρατηρήθηκε, επίσης, αύξηση του κινδύνου εμφάνισής της για τιμές διαστολικής πίεσης πάνω από 84 mmHg. Αν και ο γλυκαιμικός έλεγχος αποδείχτηκε ο πιο σχετικός παράγοντας, τρέχουσες προσεγγίσεις εντατικής θεραπείας δεν μπορούν από μόνες τους να αποτρέψουν παντελώς την εμφάνιση της επιπλοκής. Συγκεκριμένα, δε φαίνεται να υπάρχει κάποιο γλυκαιμικό "κατώφλι", κάτω από το οποίο ένας ασθενής μπορεί να είναι σίγουρος ότι είναι προστατευμένος από την επιπλοκή, γεγονός που επιδεικνύει ότι πρέπει να αναζητηθούν άλλες μέθοδοι πρόληψης.

Η υπεργλυκαιμία σε συνδυασμό με την εμφάνιση αμφιβληστροειδοπάθειας, νεφροπάθειας και νευροπάθειας σε διαβητικά άτομα τύπου Ι μελετάται από την ερευνητική ομάδα του Diabetes Control and Complications Trial (DCCT) [Dia93]. Ένα σύνολο από 1441 ασθενείς που είτε δεν είχαν αμφιβληστροειδοπάθεια (ομάδα πρωτογενούς πρόληψης), είτε βρίσκονταν στα πολύ αρχικά της στάδια (ομάδα δευτερογενούς πρόληψης) ακολούθησαν είτε συμβατική θεραπεία (μία ή δύο ενέσεις ινσουλίνης καθημερινά) ή πιο εντατική θεραπεία (τρεις ή περισσότερες ενέσεις καθημερινά ή συνεχής παροχή ινσουλίνης μέσω αντλίας). Οι ασθενείς παρακολουθούνταν για 6,5 χρόνια, κατά τα οποία εξεταζόταν τακτικά το ενδεχόμενο εμφάνισης κάποιας επιπλοκές. Στην πρώτη ομάδα, η ανάπτυξη αμφιβληστροειδοπάθειας για τους 36 πρώτους μήνες ακολούθησε ίδιους ρυθμούς και για τα δύο είδη θεραπείας. Από το σημείο αυτό και μετά, παρουσιάστηκε σημαντική μείωση στην κατηγορία της εντατικής θεραπείας. Πιο συγκεκριμένα, τα αποτελέσματα έδειξαν ότι η εντατική θεραπεία μείωσε τον κίνδυνο εμφάνισης αμφιβληστροειδοπάθειας κατά 76% κατά μέσο όρο. Στην ομάδα δευτερογενούς πρόληψης, η κατηγορία εντατικής θεραπείας είχε υψηλότερους ρυθμούς επιδείνωσης κατά το πρώτο έτος. Μετά τους 36 μήνες, όμως, αποδείχτηκε ότι η εντατική θεραπεία μείωσε τον κίνδυνο επιδείνωσης της ασθένειας κατά 54%. Επιπλέον, και στις δύο ομάδες, η εντατική θεραπεία μείωσε την εμφάνιση μικρολευκωματινουρίας κατά 39%, της λευκωματινουρίας κατά 54% και της κλινικής νευροπάθειας κατά 60%. Η εντατική θεραπεία συνδέθηκε όμως και με συχνά επεισόδια υπογλυκαιμίας.

Η επίδραση της εντατικής θεραπείας μελετήθηκε και στη U.K. Prospective Diabetes Study (UKPDS) [UKP98], αλλά σε ασθενείς με διαβήτη τύπου ΙΙ αυτή φορά. Σύμφωνα με τη μελέτη αυτή, ο συχνός γλυκαιμικός έλεγχος μείωσε την πιθανότητα εμφάνισης αμφιβληστροειδοπάθειας και νευροπάθειας και, πιθανώς, και νεφροπάθειας. Ο ρυθμός εμφάνισης μικροαγγειακών επιπλοκών συνολικά μειώθηκε κατά 25% στους ασθενείς που ακολούθησαν εντατική θεραπεία σε σχέση με όσους ακολούθησαν την παραδοσιακή. Επιπλέον, σύμφωνα με τη UKPD Study προέκυψε ότι κάθε μείωση των επιπέδων γλυκοζυλιωμένης αιμοσφαιρίνης (π.χ. από 8 σε 7%), συνεπάγεται μείωση της πιθανότητας εμφάνισης μικροαγγειακών επιπλοκών κατά 35%.

Η UKPD Study μελέτησε επίσης και τη επίδραση του αυστηρού ελέγχου των επιπέδων αρτηριακής πίεσης στην εμφάνιση επιπλοκών [UKP98]. Ένα σύνολο από 1148 ασθενείς με διαβήτη τύπου ΙΙ χωρίστηκε τυχαία σε δύο ομάδες, μία ομάδα στην οποία εφαρμόστηκε λιγότερο αυστηρός έλεγχος

αρτηριακής πίεσης (180/105 mmHg) και μία στην οποία εφαρμόστηκε αυστηρός έλεγχος (150/85 mmHg) με τη χρήση Αναστολέων του μετατρεπτικού ενζύμου της αγγειοτασίνης (ACE inhibitor) ή β-αποκλειστών (b-blockers). Μετά από παρακολούθηση 8,4 ετών, οι ασθενείς της ομάδας αυστηρού ελέγχου είχαν κατά 10/5 mmHg χαμηλότερη αρτηριακή πίεση και παρουσίαζαν κατά 34% μείωση της εξέλιξης της αμφιβληστροειδοπάθειας και κατά 47% μικρότερο κίνδυνο μείωσης της οπτικής οξύτητας. Επίσης, εμφάνιζαν χαμηλότερη συχνότητα θανάτων σχετικών με το διαβήτη και εγκεφαλικών επεισοδίων.

Η συσχέτιση αρτηριακής πίεσης και επιπλοκών μελετάται και στην εργασία Appropriate Blood Pressure Control in Diabetes (ABCD) Trial [EJG+00]. Οι ασθενείς (470 στο σύνολο) χωρίστηκαν και πάλι σε δύο ομάδες, μία λιγότερο αυστηρού ελέγχου και μία αυστηρού. Η τιμή αρτηριακής πίεσης που επιτεύχθηκε κατά μέσο όρο ήταν 132/78 mmHg στην ομάδα αυστηρού ελέγχου και 138/86 mmHg στη λιγότερο αυστηρού ελέγχου. Αν και η εντατική θεραπεία συνδέθηκε με μικρότερα ποσοστά θανάτου, δεν υπήρξε μεγάλη διαφορά ανάμεσα στις ομάδες σχετικά με την εμφάνιση αμφιβληστροειδοπάθειας και νευροπάθειας.

Η μελέτη Hoorn [LDM+03] αποτελεί ακόμα μία εργασία που εξετάζει την εμφάνιση διαβητικής αμφιβληστροειδοπάθειας. Κρίνει σαν παράγοντες σχετικούς με την επιπλοκή τα επίπεδα γλυκοζυλιωμένης αιμοσφαιρίνης, τα επίπεδα αρτηριακής πίεσης και την παχυσαρκία. Βασίστηκε σε 233 άτομα με ηλικία από 50 έως 74 ετών, τα οποία παρακολουθήθηκαν για μία διάρκεια 9,4 ετών. Τα χαρακτηριστικά των ατόμων που είχαν αναπτύξει αμφιβληστροειδοπάθεια μέχρι το τέλος των 9,4 ετών συγκρίθηκαν με αυτά των ατόμων που δεν την εμφάνισαν, χρησιμοποιώντας t τεστ, χ^2 τεστ με διόρθωση συνέχειας ή ένα Mann-Whitney τεστ σε περίπτωση ασύμμετρης κατανομής. Λογιστική παλινδρόμηση (Logistic Regression) χρησιμοποιήθηκε για τον υπολογισμό της αναλογίας πιθανοτήτων (odds ratios - ORs) με διάστημα εμπιστοσύνης 95%.

Όλες οι εργασίες που αναφέραμε αποτελούν αξιόλογες μελέτες μεγάλης κλίμακας, που έχουν προσφέρει πολλά στην προσπάθεια αναζήτησης των παραγόντων που οδηγούν στην εμφάνιση της αμφιβληστροειδοπάθειας. Το γεγονός ότι κάποιες καταλήγουν σε εν μέρει αντικρουόμενα αποτελέσματα ή σε παρόμοια, αλλά όχι με την ίδια βεβαιότητα, αποδεικνύει έμπρακτα την πολύπλοκη φύση του προβλήματος και την ανάγκη περαιτέρω διερεύνησης. Επιπλέον, οι παραπάνω εργασίες υιοθετούν κυρίως στατιστικές και γραμμικές μεθόδους, που ενδέχεται να αδυνατούν να απεικονίσουν πλήρως ένα τέτοιο πολύπλοκο, μη-γραμμικό πρόβλημα. Η εφαρμογή πιο σύγχρονων, "έξυπνων" μεθόδων από το χώρο της Τεχνητής Νοημοσύνης, όπως αυτές που χρησιμοποιούνται στην παρούσα διπλωματική, ενδεχομένως να μπορούν να προσφέρουν περισσότερα στην κατεύθυνση αυτή.

3.2 Μοντέλα πρόβλεψης/εκτίμησης κινδύνου εμφάνισης επιπλοκών του διαβήτη.

Τα μοντέλα εκτίμησης της πιθανότητας εμφάνισης μακροπρόθεσμων επιπλοκών υπολογίζουν την πιθανότητα εμφάνισης διάφορων επιπλοκών μίας ασθένειας και την επίδραση διαφόρων μεθόδων θεραπείας στην μείωση της πιθανότητας εμφάνισης αυτών των επιπλοκών.

Αν και η λήψη διάφορων ιατρικών αποφάσεων βασίζεται κατά κύριο λόγω σε κλινικές δοκιμές, σε μελέτες από τους τομείς της στατιστικής και της φαρμακολογίας και σε ότι έχει δείξει η μέχρι τώρα εμπειρία, τα υπολογιστικά συστήματα πρόβλεψης και εκτίμησης κινδύνου εμφάνισης επιπλοκών έχουν αρχίσει να κερδίζουν ολοένα και μεγαλύτερο έδαφος, ειδικά στις περιπτώσεις που οι αποφάσεις επηρεάζουν την εξέλιξη του ασθενή σε βάθος χρόνου και που δεν έχει βρεθεί μία πάγια βέλτιστη τακτική. Μία τέτοια περίπτωση είναι και η επιλογή της καταλληλότερης για το διαβητικό άτομο θεραπείας προς αποφυγή της εμφάνισης επιπλοκών. Επομένως, ένα μοντέλο που έχει συγκεντρώσει τη γνώση που προκύπτει μέσα από πολλές περιπτώσεις ασθενών και μπορεί να "βλέπει" μακροπρόθεσμα, μπορεί να φανεί πολύ χρήσιμο στα χέρια ενός ειδικού, σαν "σύμβουλος" στην επιλογή της κατάλληλης θεραπείας και στη γενικότερη διαχείριση της ασθένειας του ατόμου (disease management, health planning) βάσει της τρέχουσας κλινικής του εικόνας.

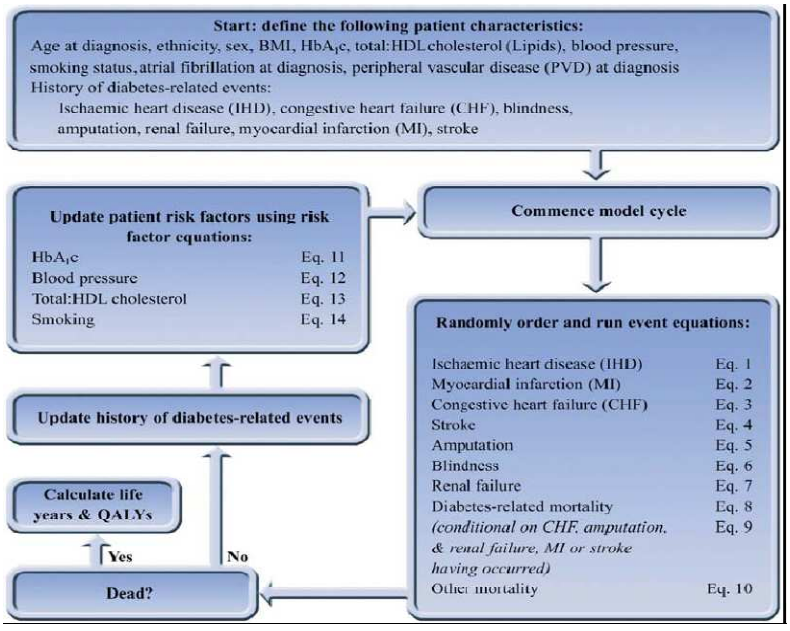
Ένα μοντέλο εκτίμησης της πιθανότητας εμφάνισης μίας επιπλοκής μπορεί να προσφέρει και στον τομέα της ενημέρωσης και της αφύπνισης του ατόμου σχετικά με τον κίνδυνο. Εφόσον μπορεί βάσει της τρέχουσας κατάστασης του ασθενή να εκτιμήσει τον κίνδυνο που διατρέχει μελλοντικά, θέτει έμμεσα το ζήτημα της ενδεχόμενης αλλαγής στη διατροφή, στις συνήθειες και στη θεραπεία που αυτός ακολουθεί και "κρούει τον κώδωνα του κινδύνου".

Έχουν αναπτυχθεί διάφορες προσεγγίσεις για την υλοποίηση μοντέλων εκτίμησης κινδύνου εμφάνισης των μακροπρόθεσμων επιπλοκών του διαβήτη. Στη συνέχεια αναφέρουμε μερικά από τα πιο διαδεδομένα μοντέλα που συναντήσαμε στη βιβλιογραφία.

Ένα από τα πιο γνωστά μοντέλα πρόβλεψης στο χώρο αυτό είναι το UKPDS Outcomes Model [CGB+04], το οποίο μελετήσαμε ιδιαίτερα. Αποτελεί ένα μοντέλο εκτίμησης της πιθανής εμφάνισης επιπλοκών του Σακχαρώδους Διαβήτη σε ασθενείς με διαβήτη τύπου ΙΙ, με στόχο τον υπολογισμό αποτελεσμάτων σχετικών με τον τομέα των Οικονομικών Υγείας (Health economics), όπως το προσδόκιμο ποιοτικής ζωής (quality-adjusted life expectancy). Πιο συγκεκριμένα, το UKPDS Outcomes Model περιλαμβάνει έναν αριθμό από παραμετρικές εξισώσεις για τον υπολογισμό του κινδύνου εμφάνισης για πρώτη φορά για κάθε μία από επτά επιπλοκές του διαβήτη (θανατηφόρο ή μη θανατηφόρο έμφραγμα του μυοκαρδίου, ισχαιμική καρδιοπάθεια, εγκεφαλικό επεισόδιο, καρδιακή ανεπάρκεια, ακρωτηριασμός, νεφροπάθεια και ασθένεια των ματιών μετρημένη με όρους τύφλωσης στο ένα μάτι) και πιθανότητα θανάτου. Στη συνέχεια εκτιμά την προσδοκώμενη διάρκεια

ζωής και την προσδοκώμενη διάρκεια "ποιοτικής" ζωής. Ένα πολύ σημαντικό προσόν της εργασίας αυτής είναι ότι είναι σχεδιασμένο έτσι ώστε να συμπεριλαμβάνει την εξάρτηση μεταξύ των διάφορων επιπλοκών σε κάθε ασθενή. Οι επιπλοκές μπορεί να σχετίζονται όχι μόνο επειδή μοιράζονται του ίδιους παράγοντες κινδύνου (για παράδειγμα τα υψηλά επίπεδα γλυκόζης), αλλά και λόγω ενδο-εξαρτήσεων, καθώς η εμφάνιση μίας επιπλοκής αυξάνει την πιθανότητα εμφάνισης και μίας άλλης [BRC+00]. Για παράδειγμα, η πιθανότητα εμφάνισης καρδιακής ανεπάρκειας ή εμφράγματος του μυοκαρδίου επηρεάζεται άμεσα από τα επίπεδα συστολικής αρτηριακής πίεσης, αλλά η πιθανότητα εμφράγματος του μυοκαρδίου είναι μεγαλύτερη σε ασθενείς με ιστορικό καρδιακής ανεπάρκειας. Η ανάπτυξη του μοντέλου βασίστηκε στα δεδομένα 3642 ασθενών μέσα από παρακολούθηση 10,3 ετών. Είσοδοι στο μοντέλο είναι τα χαρακτηριστικά του ασθενή (ηλικία, φύλο, εθνικότητα, κ.α.), χρονικά μεταβαλλόμενοι παράγοντες κινδύνου (γλυκοζυλιωμένη αιμοσφαιρίνη, συστολική αρτηριακή πίεση και ολική χοληστερόλη) και το ιστορικό του ασθενή σε επιπλοκές. Κάθε επιπλοκή μοντελοποιήθηκε μέσω μίας ή περισσοτέρων εξισώσεων με χρονικά μεταβαλλόμενες παραμέτρους, οι οποίες προσαρμόστηκαν στα δεδομένα των ασθενών. Το μοντέλο αποτελείται από έναν αριθμό κύκλων/καταστάσεων. Κάθε κύκλος αντιστοιχεί σε ένα έτος, μέσα στο οποίο ο ασθενής ενδέχεται να παρουσιάσει κάποια μη θανατηφόρα επιπλοκή ή/και να πεθάνει. Σε κάθε κύκλο υπολογίζονται οι πιθανότητες εμφάνισης των επιπλοκών και κάθε πιθανότητα που υπολογίζεται συγκρίνεται με έναν τυχαίο αριθμό από την ομοιόμορφη κατανομή στο διάστημα [0,1]. Μέσω της σύγκρισης αυτής αποφασίζεται το αν το μοντέλο από εκείνο το σημείο και μετά θεωρεί ότι συνέβη το αντίστοιχο περιστατικό ή όχι. Επίσης, σε κάθε κύκλο ανανεώνονται οι τιμές των παραγόντων κινδύνου και οι νέες τιμές μεταφέρονται στον επόμενο κύκλο του μοντέλου. Στο τέλος της προσομοίωσης, υπολογίζονται τα χρόνια ζωής και τα χρόνια ποιοτικής ζωής.

Παρακάτω φαίνεται σχηματικά ο αλγόριθμος του μοντέλου:



Σχήμα 3.1: Ο αλγόριθμος του UKPDS Outcomes Model.

Ένα επίσης γνωστό μοντέλο είναι το PROPHET (Prospective Population Health Event Tabulation). [CCJ92]. Πρόκειται για ένα ευέλικτο και δυνατό πρόγραμμα προσομοίωσης χρόνιων, μη ιάσιμων ασθενειών, που συνδυάζει στοιχεία από Δέντρα Αποφάσεων (Decision Trees), αλυσίδες Markov (Markov processes) και τεχνικές Monte Carlo. Προσομοιώνει ασθένειες όπως ο διαβήτης τύπου I, ο τύπου II, η στεφανιαία νόσος, η συμφορητική καρδιακή ανεπάρκεια, το άσθμα, το εγκεφαλικό επεισόδιο, η υπέρταση, και η παχυσαρκία. Περιλαμβάνει περισσότερες από 100 μεταβλητές οι οποίες αναπαριστούν βιολογικούς παράγοντες, συμπτώματα, αποτελέσματα, κ.α. Προσομοιώνει την επίδραση διάφορων ασθενειών και θεραπειών μέσω μεταβάσεων του ασθενή σε έναν αριθμό καταστάσεων. Ο προσομοιωμένος ασθενής μεταβαίνει από τη μία κατάσταση στην επόμενη σύμφωνα με τα χαρακτηριστικά του, τη θεραπεία που ακολουθεί και τις προκαθορισμένες πιθανότητες μετάβασης, που τα στοιχεία αυτά συνεπάγονται. Σε κάθε κύκλο υπολογίζονται τα οφέλη μίας θεραπείας ή μίας εξέτασης με όρους ποιότητας ζωής, καθώς επίσης και το οικονομικό τους κόστος.

Η επίδραση των διαφόρων ειδών θεραπείας σε άτομα με διαβήτη τύπου I και τύπου II στη μακροπρόθεσμη κατάσταση της υγείας, καθώς και οι οικονομικοί όροι της κάθε μίας, προσομοιώνονται στο CORE model [PRV+04]. Το μοντέλο αποτελείται από 14 αλληλοεξαρτώμενα υπο-μοντέλα που προσομοιώνουν τις επιπλοκές του διαβήτη και από ένα υπο-μοντέλο προσομοίωσης της θνησιμότητας. Κάθε υπο-μοντέλο είναι τύπου Markov με πιθανότητες που εξαρτώνται από το χρόνο, την κατάσταση του ασθενή και τον τύπο του διαβήτη. Για τον υπολογισμό των αποτελεσμάτων λαμβάνονται υπόψη τα χαρακτηριστικά του ασθενή, το ιστορικό του σε επιπλοκές, θεραπείες, εξετάσεις και αλλαγές στις φυσιολογικές παραμέτρους συναρτήσει του χρόνου. Εκτιμώνται η εμφάνιση και η εξέλιξη των διάφορων επιπλοκών, το προσδόκιμο ζωής, τα χρόνια ποιοτικής ζωής (quality-adjusted life-years – QALY) και το συνολικό κόστος.

Ακόμη ένα μοντέλο τύπου Markov για την πρόβλεψη της εξέλιξης του διαβήτη τύπου II σε κάποια μακροπρόθεσμη επιπλοκή είναι το Model of Complications NIDDM [EJH+97]. Αποτελεί το πρώτο μοντέλο αυτού του είδους για το διαβήτη τύπου II. Οι πιθανότητες εμφάνισης επιπλοκών είναι προκαθορισμένες και βασίζονται σε πληθυσμιακές μελέτες που έχουν πραγματοποιηθεί.

Αναφέρουμε, τέλος, το Παγκόσμιο Μοντέλο Διαβήτη (Global Diabetes Model – GDM) [BRC+00]. Πρόκειται για ένα συνεχές και στοχαστικό μοντέλο προσομοίωσης του διαβήτη τύπου II, κατάλληλο για προβλέψεις και σε ατομικό και σε πληθυσμιακό επίπεδο. Προβλέπει την εμφάνιση επιπλοκών, καθώς επίσης και τα οφέλη και το κόστος της κάθε θεραπείας. Σαν παράγοντες κινδύνου χρησιμοποιούνται η γλυκοζυλιωμένη αιμοσφαιρίνη, η συστολική αρτηριακή πίεση, οι λιποπρωτεΐνες χαμηλής πυκνότητας (LDL), οι λιποπρωτεΐνες υψηλής πυκνότητας (HDL), τα τριγλυκερίδια, το κάπνισμα και η λήψη ασπιρίνης. Οι συναρτήσεις πιθανότητας εμφάνισης των μικροαγγειακών και μακροαγγειακών επιπλοκών που χρησιμοποιεί είναι βασισμένες σε έναν συνδυασμό από δημοσιευμένες μελέτες και αναλύσεις δεδομένων διαβητικών ατόμων.

Τα παραπάνω μοντέλα έχουν αποτελέσει σταθμό στον τομέα της πρόβλεψης της εμφάνισης μακροπρόθεσμων επιπλοκών του διαβήτη. Το γεγονός ότι κάποια βασίζονται σε πληθυσμιακές μελέτες που έχουν προηγηθεί, ενδέχεται να αφαιρεί από την ακρίβειά τους. Επιπλέον, σε κάποια ελέγχθηκε μόνο η εσωτερική τους εγκυρότητα (internal validity). Δηλαδή, η αξιολόγησή τους πραγματοποιήθηκε μόνο με τα ίδια δεδομένα βάσει των οποίων αναπτύχθηκαν οι εξισώσεις με τις οποίες λειτουργεί. Αυτού του είδους η αξιολόγηση δεν είναι απόλυτα ενδεικτική της απόδοσης του μοντέλου, καθώς φανερώνει ότι το μοντέλο μπορεί να προβλέπει με ακρίβεια τα δεδομένα στα οποία βασίστηκε ή περιπτώσεις δεδομένων παρόμοιων με αυτά, αλλά είναι άγνωστο το πώς λειτουργεί σε άλλους πληθυσμούς.

3.3 Χρήση μεθόδων Τεχνητής Νοημοσύνης σε εφαρμογές επιλογής χαρακτηριστικών

Στην ενότητα αυτή παραθέτουμε μερικές εργασίες που κάνουν χρήση των εργαλείων που χρησιμοποιούνται και στην παρούσα διπλωματική.

Η εφαρμογή του κριτηρίου της Αμοιβαίας Πληροφορίας στην επιλογή χαρακτηριστικών προτάθηκε πρώτα από τον Battiti το 1994. Μία παραλλαγή αυτού του κριτηρίου, το κριτήριο της Μέγιστης – Σχετικότητας και Ελάχιστου – Πλεονασμού (Maximum-Relevance Minimum-Redundancy) προτείνεται το 2005 από τους Peng et al. [PLD05]. Τα κριτήριο αυτό βασίζεται στην ιδέα ότι το καλύτερο υποσύνολο θα είναι αυτό που περιέχει χαρακτηριστικά εξαρτημένα με την έξοδο, αλλά όχι εξαρτημένα μεταξύ τους. Έχουν διατυπωθεί και άλλες παραλλαγές του κριτηρίου από τους Yang and Moody (1999), και από τους Kwak and Choi (2002). Στις περισσότερες εργασίες που συναντήσαμε ([ADC03], [Bro09], [DBR+10], [KC02], [LSL+09], [SP10], [TFM+01]) τα κριτήρια αυτά χρησιμοποιούνται σε συνδυασμό με έναν αλγόριθμο σειριακής αναζήτησης που σταδιακά προσθέτει ή αφαιρεί χαρακτηριστικά. Οι "άπληστοι" (greedy) αυτοί σειριακοί αλγόριθμοι συχνά υποφέρουν από προσκόλληση σε τοπικά ελάχιστα και ενδέχεται να μην βρουν τη βέλτιστη δυνατή λύση. Επιλέγοντας κάθε φορά το χαρακτηριστικό με τη μέγιστη τιμή της συνάρτησης καταλληλότητας δεν εξασφαλίζεται ότι θα καταλήξουμε στο καλύτερο δυνατό υποσύνολο, καθώς ενδέχεται ο συνδυασμός δύο χαρακτηριστικών να προβλέπει με μεγαλύτερη ακρίβεια την έξοδο από ότι μόνο του το χαρακτηριστικό με τη μεγαλύτερη Αμοιβαία Πληροφορία. Επίσης, η μέθοδος σταδιακής πρόσθεσης χαρακτηριστικών δεν επιτρέπει την αφαίρεση χαρακτηριστικών που επιλέχθηκαν σε προηγούμενο βήμα.

Ένας πιο "προχωρημένος" αλγόριθμος προτείνεται στην εργασία "A novel information theoretic-interact algorithm (IT-IN) for feature selection" [DBR+10]. Το κριτήριο επιλογής βασίζεται πάλι σε έννοιες της Θεωρία Πληροφορίας, αλλά ελέγχει επίσης τη συνέπεια (consistency) μεταξύ των δεδομένων. Για την αξιολόγηση του αλγόριθμου χρησιμοποιούνται τρεις διαφορετικοί ταξινομητές

για δέκα διαφορετικά σύνολα δεδομένων και τα αποτελέσματά του είναι πολύ ικανοποιητικά και για τους τρεις ταξινομητές.

Η εργασία [TFM+01] αποτελεί μία συγκριτική μελέτη διαφόρων προσεγγίσεων επιλογής χαρακτηριστικών. Ο στόχος είναι η επιλογή του βέλτιστου υποσυνόλου χαρακτηριστικών για διάγνωση με τη βοήθεια υπολογιστή (computer-aided diagnosis – CAD) της οξείας πνευμονικής εμβολής (acute pulmonary embolism-PE). Η μελέτη βασίστηκε σε δεδομένα που προέρχονται από σπινθηρογράφημα αιμάτωσης των πνευμόνων (Perfusion lung scan). Υλοποιούνται οι εξής μέθοδοι επιλογής χαρακτηριστικών: τρεις μέθοδοι που κάνουν χρήση της έννοιας της Αμοιβαίας Πληροφορίας, η μέθοδος της Σταδιακής Γραμμικής Διαχωριστικής Ανάλυσης (Stepwise Linear Discriminant Analysis – Stepwise LDA) και ένας Γενετικός Αλγόριθμος. Στη συνέχεια υλοποιούνται δύο μοντέλα πρόβλεψης, ένα γραμμικό και ένα μη γραμμικό. Το γραμμικό μοντέλο βασίζεται στη Γραμμική Διαχωριστική Ανάλυση (LDA), ενώ το μη γραμμικό είναι ένα Τεχνητό Νευρωνικό Δίκτυο (ANN) τριών επιπέδων πρόσθιας τροφοδότησης. Τελικά, αποδεικνύεται η δύναμη του κριτηρίου της Αμοιβαίας Πληροφορίας σε σχέση με πιο παραδοσιακές τεχνικές, όπως η LDA. Ο γενετικός αλγόριθμος καταφέρνει να εξαλείψει ένα περιττό χαρακτηριστικό που η μέθοδος της Αμοιβαίας Πληροφορίας διατήρησε. Αν και αποτελεσματικός, ο τρόπος που υλοποιήθηκε περιλαμβάνει την επανεκπαίδευση εκατοντάδων Νευρωνικών Δικτύων, κάτι που κρίθηκε πολύ ακριβό υπολογιστικά.

Η δύναμη και η ευελιξία των Γενετικών Αλγορίθμων τους καθιστά πολύτιμο εργαλείο στη διαδικασία επιλογής χαρακτηριστικών. Η χρήση Γενετικών Αλγορίθμων μπορεί να υιοθετηθεί σε επιλογή χαρακτηριστικών και σε προσέγγιση-φίλτρο (filter) και σε προσέγγιση-περιτύλιγμα (wrapper). Μερικές εφαρμογές που υιοθετούν τη χρήση του είναι οι εργασίες [HRK+99], [IN00], [MDJ09], [YH97] και [SS89].

Συχνά οι Γενετικοί Αλγόριθμοι χρησιμοποιούνται, επίσης, σε υβριδικές προσεγγίσεις επιλογής χαρακτηριστικών, που αποτελούν συνδυασμό της filter και της wrapper προσέγγισης. Στην εργασία [HCX07] πραγματοποιούνται δύο στάδια βελτιστοποίησης, ένα "εξωτερικό" και ένα "εσωτερικό". Το "εξωτερικό" περιλαμβάνει έναν Γενετικό Αλγόριθμο με συνάρτηση καταλληλότητας την Αμοιβαία Πληροφορία μεταξύ της εξόδου του ταξινομητή και της πραγματικής εξόδου (wrapper προσέγγιση). Το "εσωτερικό" είναι ένα στάδιο τοπικής αναζήτησης, που κατατάσσει τα χαρακτηριστικά (feature ranking) του υποψήφιου υποσυνόλου βάσει της Αμοιβαίας Πληροφορίας των χαρακτηριστικών με την έξοδο και μεταξύ τους και εξαλείφει αυτά με το χαμηλότερο σκορ (filter προσέγγιση). Στην εργασία [Can04] πραγματοποιείται αρχικά η filter τεχνική, κατά την οποία επιλέγονται τα υποσύνολα των χαρακτηριστικών που συνεπάγονται καλύτερο διαχωρισμό μεταξύ των κλάσεων. Υλοποιείται στη συνέχεια ένας Γενετικός Αλγόριθμος που αρχικοποιείται με αυτά τα υποσύνολα και έχει σαν συνάρτηση καταλληλότητας την απόδοση του ταξινομητή (wrapper τεχνική).

Τα Τεχνητά Νευρωνικά Δίκτυα χρησιμοποιούνται ευρέως ως ταξινομητές, λόγω των ιδιοτήτων τους που, όπως έχουμε εξηγήσει στο προηγούμενο κεφάλαιο, τα καθιστούν ιδιαίτερα κατάλληλα για τέτοιου είδους εργασίες. Μία τέτοια εργασία είναι η [SR07], όπου χρησιμοποιείται η μεταευριστική μέθοδος βελτιστοποίησης με Αποικίες Μυρμηγκιών (Ant Colony Optimization), καθοδηγούμενη από την απόδοση ταξινόμησης ενός Νευρωνικού Δικτύου.

Έχουν προταθεί και μέθοδοι επιλογής χαρακτηριστικών με χρήση Νευρωνικών Δικτύων, στις οποίες η βασική ιδέα είναι η "αποσύνθεση" των βαρών. Εφαρμόζεται, δηλαδή, ένα είδος περιορισμού στις τιμές των βαρών που συνδέουν τις εισόδους με το κρυφό επίπεδο με σκοπό τα βάρη που συνδέονται με ασήμαντα χαρακτηριστικά να διατηρήσουν τιμές κοντά στο μηδέν. Επιλέγονται τα χαρακτηριστικά που αντιστοιχούν στα βάρη που θα "επιζήσουν". Εργασίες που υιοθετούν αυτή τη λογική είναι η [VB02] και η [CSG+96].

Ένα συνδυασμό Γενετικού Αλγόριθμου και Νευρωνικού Δικτύου αποτελεί η εργασία [GG10]. Συνάρτηση καταλληλότητας του Γενετικού Αλγόριθμου είναι το μέσο τετραγωνικό σφάλμα της ταξινόμησης του Νευρωνικού Δικτύου που εκπαιδεύεται βάσει του κάθε χρωμοσώματος.

Συνδυασμός των δύο αυτών υπολογιστικών εργαλείων υιοθετείται, τέλος, και στην εργασία [VZ07], αλλά με διαφορετική λογική. Η ψηφιακή μαστογραφία είναι από τις πλέον κατάλληλες μεθόδους για την έγκαιρη διάγνωση του καρκίνου του μαστού. Είναι όμως αρκετά δύσκολος ο διαχωρισμός μεταξύ καλοηθών και καοηθών μικροασβεστώσεων (οι μικρότερες δομές που εντοπίζονται σε μια μαστογραφία). Στην εργασία αυτή υλοποιείται είναι υπολογιστικό σύστημα επιλογής χαρακτηριστικών και ταξινόμησης που μπορεί να παρέχει μία δεύτερη γνώμη στον ακτινολόγο για την αξιολόγηση των μικροασβεστώσεων. Προτείνεται ένας αλγόριθμος που συνδυάζει έναν Γενετικό Αλγόριθμο και ένα Νευρωνικό Δίκτυο με σκοπό την επιλογή των καλύτερων χαρακτηριστικών και της κατάλληλης δομής του Νευρωνικού Δικτύου. Αρχικά εκπαιδεύεται ένα νευρωνικό δίκτυο με ένα τυχαίο υποσύνολο χαρακτηριστικών και τυχαία αρχικά βάρη. Τα βάρη ρυθμίζονται με έναν γενετικό αλγόριθμο με συνάρτηση καταλληλότητας την απόδοση της ταξινόμησης. Τα χαρακτηριστικά επιλέγονται, επίσης με έναν γενετικό αλγόριθμο με κριτήριο την απόδοση του ταξινομητή και η όλη διαδικασία επαναλαμβάνεται επαναληπτικά.

3.4 Επιλογή προδιαθεσικών παραγόντων για εκτίμηση κινδύνου εμφάνισης αμφιβληστροειδοπάθειας με μεθόδους Τεχνητής Νοημοσύνης

Όπως έχουμε αναφέρει, στόχος μας είναι η εύρεση των παραγόντων που έχουν προγνωστική αξία για την εμφάνιση της διαβητικής αμφιβληστροειδοπάθειας και η υλοποίηση ενός μοντέλου

εκτίμησης του κινδύνου εμφάνισής της χρησιμοποιώντας, και για τις δύο αυτές εργασίες, εργαλεία από το χώρο της Τεχνητής Νοημοσύνης.

Η εργασία της επιλογής των σημαντικών παραγόντων πριν την υλοποίηση του μοντέλου αποτελεί ένα πολύ σημαντικό στάδιο. Πρώτα απ' όλα, μπορεί να αντιμετωπιστεί σαν μία αυτόνομη μελέτη σχετικά με τους προδιαθεσιακούς παράγοντες που συνδέονται με την αμφιβληστροειδοπάθεια. Επιπλέον, είναι σημαντικό να διαπιστώσουμε σε ποια χαρακτηριστικά πρέπει να βασιστούμε για την υλοποίηση του μοντέλου πρόβλεψης, καθώς ενδέχεται κάποια από τα χαρακτηριστικά που διαθέτουμε να είναι "θορυβώδη", να παρεμποδίζουν δηλαδή την κατανόηση των εσωτερικών συσχετίσεων μεταξύ των δεδομένων. Ακόμα κι αν κάποια χαρακτηριστικά σχετίζονται άμεσα με το διαβήτη γενικά, την εκδήλωσή του ή με κάποια από τις μακροχρόνιες επιπλοκές του, ενδέχεται να μην σχετίζονται με την εμφάνιση διαβητικής αμφιβληστροειδοπάθειας, επομένως είναι σημαντικό να εξεταστεί ποια χαρακτηριστικά συνδέονται αποκλειστικά με την τελευταία. Τέλος, επειδή διαθέτουμε ένα αρκετά μεγάλο πλήθος χαρακτηριστικών, η διατήρηση όλων επιβαρύνει την υπολογιστική διαδικασία και αυξάνει τις απαιτήσεις για μνήμη και τον χρόνο επεξεργασίας και εκπαίδευσης του μοντέλου πρόβλεψης. Η επιλογή των σημαντικών μόνο χαρακτηριστικών βελτιώνει σημαντικά αυτές τις πλευρές του δεδομένου προβλήματος.

Για την επίτευξη του στόχου μας, προσπαθήσαμε να εφαρμόσουμε τη γνώση που αποκομίσαμε από τη βιβλιογραφική μας έρευνα με σκοπό την ανάπτυξη μίας, όσο το δυνατό, αποδοτικής μεθοδολογίας. Αποφασίσαμε να συνδυάσουμε τις δύο τεχνικές επιλογής χαρακτηριστικών filter και wrapper, έτσι ώστε να εκμεταλλευτούμε τα θετικά στοιχεία και των δύο προσεγγίσεων. Υιοθετήσαμε την έννοια της Αμοιβαίας Πληροφορίας για την επιλογή των χαρακτηριστικών, λόγω της διακριτικής της ικανότητας σε αλληλεξαρτήσεις. Αποφύγαμε "άπληστους" αλγόριθμους αναζήτησης χρησιμοποιώντας ένα Γενετικό Αλγόριθμο με συνάρτηση καταλληλότητας την Αμοιβαία Πληροφορία των χαρακτηριστικών με την έξοδο (δηλαδή με την εμφάνιση ή όχι της αμφιβληστροειδοπάθειας) και των χαρακτηριστικών μεταξύ τους, έτσι ώστε το κάθε υποψήφιο υποσύνολο-χρωμόσωμα να κρίνεται "συνολικά". Με τον τρόπο αυτό, η δύναμη της Αμοιβαίας Πληροφορίας στον εντοπισμό αλληλεξαρτήσεων και η ευελιξία του γενετικού αλγόριθμου στην αναζήτηση καλύτερων λύσεων συνδυάζονται σε ένα εργαλείο και αξιοποιούνται ταυτόχρονα. Η διαδικασία αυτή αποτελεί ένα "προπαρασκευαστικό" στάδιο και βασίζεται στη filter προσέγγιση της επιλογής χαρακτηριστικών, καθώς κρίνει τα χαρακτηριστικά ανεξαρτήτως ταξινομητή. Έξοδος του σταδίου είναι μερικά υποσύνολα που κρίθηκαν ως "καλύτερα" από το Γενετικό Αλγόριθμο.

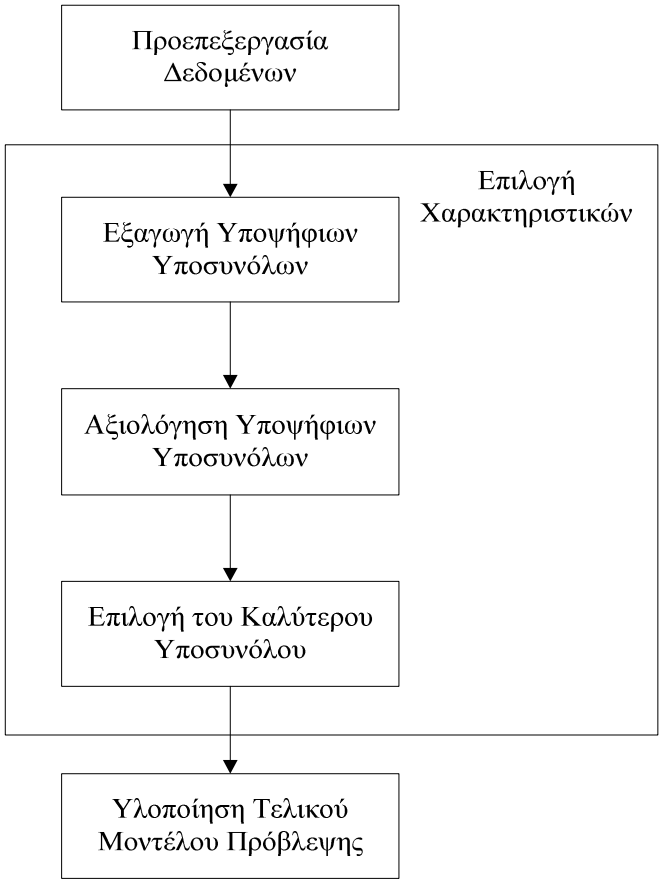
Στη συνέχεια υλοποιείται ένα Τεχνητό Νευρωνικό Δίκτυο για κάθε ένα από τα υποψήφια υποσύνολα που προέκυψαν από το Γενετικό Αλγόριθμο. Το Τεχνητό Νευρωνικό Δίκτυο αποτελεί το μοντέλο πρόβλεψης εμφάνισης αμφιβληστροειδοπάθειας της μεθοδολογίας μας. Το κάθε δίκτυο εκπαιδεύεται και αξιολογείται με τη μέθοδο της Διασταυρωμένης Επικύρωσης (Cross - Validation). Διερευνάται παράλληλα ο βέλτιστος αριθμός νευρώνων και οι βέλτιστες τιμές αρχικών βαρών του δικτύου. Με

κριτήριο την απόδοση που επιτυγχάνεται σε κάθε δίκτυο, επιλέγεται το τελικό υποσύνολο υποψήφιων παραγόντων, καθώς και οι νευρώνες και οι αρχικές τιμές βαρών και πολώσεων που δίνουν το καλύτερο αποτέλεσμα. Με τον τρόπο αυτό εφαρμόζεται και η wrapper λογική της επιλογής χαρακτηριστικών και ταυτόχρονα επιλέγονται οι βέλτιστες τιμές των παραμέτρων του δικτύου.

Η αξιολόγηση πραγματοποιείται με όρους ROC ανάλυσης, μίας μεθόδου που εφαρμόζεται ευρέως σε στατιστικές μελέτες και σε ιατρικές διαγνωστικές δοκιμασίες. Στη μεθοδολογία μας η διαδικασία της αξιολόγησης έχει σα στόχο και την επιλογή του βέλτιστου υποσυνόλου προδιαθεσικών παραγόντων και του αντίστοιχου μοντέλου, αλλά και την εκτίμηση της συνολικής αξίας του μοντέλου, δηλαδή, της ικανότητας γενίκευσής του και της αξιοπιστίας του, σαν εργαλείο πρόβλεψης κινδύνου.

Υλοποιείται, τέλος, το τελικό μοντέλο πρόβλεψης με το πλήθος των νευρώνων και τις τιμές των αρχικών βαρών που έχουν προκύψει από την προηγούμενη διαδικασία. Εκπαιδεύεται με το 70% των δεδομένων μας και δοκιμάζεται ακόμη μία φορά για επαλήθευση της απόδοσής του με το υπόλοιπο 30% των δεδομένων μας.

Παρακάτω φαίνεται το διάγραμμα ροής της μεθοδολογίας μας σε υψηλό αφαιρετικό επίπεδο, που δίνει μία γενική ιδέα για την λογική της και συνοψίζει όσα περιγράψαμε στην ενότητα αυτή. Σε περισσότερες λεπτομέρειες των επιμέρους σταδίων της μεθοδολογίας θα μπούμε στο επόμενο κεφάλαιο.



Σχήμα 3.2: Διάγραμμα Ροής Μεθοδολογίας.

4

Ανάπτυξη Μεθοδολογίας

Στο κεφάλαιο αυτό αναπτύσσονται η μεθοδολογία και οι αλγόριθμοι που ακολουθήθηκαν για την επίλυση του προβλήματος που πραγματεύεται η παρούσα διπλωματική, την εύρεση δηλαδή των προδιαθεσικών παραγόντων και την εκτίμηση του κινδύνου εμφάνισης διαβητικής αμφιβληστροειδοπάθειας. Περιγράφονται παράλληλα τα προβλήματα που συναντήσαμε κατά την υλοποίηση της μεθοδολογίας, μερικές τεχνικές λεπτομέρειες και εξηγούνται οι λόγοι για τους οποίους προτιμήθηκαν κάποιες μέθοδοι από άλλες υπάρχουσες.

Όλα τα στάδια της μεθοδολογίας και όλες οι επιμέρους διεργασίες υλοποιήθηκαν στο Matlab R2008a.

4.1 Περιγραφή των δεδομένων

Η ανάπτυξη του μοντέλου εκτίμησης του κινδύνου εμφάνισης διαβητικής αμφιβληστροειδοπάθειας, βασίστηκε σε πραγματικά δεδομένα ατόμων με διαβήτη τύπου ΙΙ. Για κάθε άτομο διαθέτουμε, για μία χρονική περίοδο (follow-up) 6 ετών, μετρήσεις διαφόρων χαρακτηριστικών και το αν εμφάνισε ή όχι διαβητική αμφιβληστροειδοπάθεια σε κάθε έτος. Όπως εξηγήσαμε στο 2^{ο} κεφάλαιο, με τον όρο “χαρακτηριστικό” (feature) εννοούμε ένα σύνολο από ανεξάρτητες και αμοιβαία αποκλειόμενες τιμές. Για παράδειγμα, η ηλικία του ατόμου αποτελεί ένα χαρακτηριστικό. Όλα τα χαρακτηριστικά μαζί συνιστούν το προφίλ του κάθε ατόμου, του κάθε δηλαδή στιγμιότυπου (instance) των δεδομένων μας. Για κάθε άτομο διαθέτουμε συνολικά 27 χαρακτηριστικά. Αυτά είναι τα εξής:

(1) Δείκτης Μάζας Σώματος - ΔΜΣ (Body Mass Index – BMI), ο οποίος δίνεται από τη σχέση: $BMI = \frac{mass\,(kg)}{(height\,(m))^2}$. Χρησιμοποιείται σαν ένα μέτρο εκτίμησης του πόσο "υγειές" είναι το βάρος του ατόμου σε σχέση με το ύψος του. Ένας Δείκτης Μάζας Σώματος μεταξύ 18,5 έως 25 φανερώνει ιδανικό βάρος. Κάτω από 18,5 δηλώνει ότι το άτομο μπορεί να είναι λιποβαρές, ενώ πάνω από 25 δείχνει υπέρβαρο άτομο.

(2) Περίμετρος μέσης (Waist Circumference – WC). Αποτελεί ένα δείκτης του λίπους που είναι συγκεντρωμένο στην κοιλιακή χώρα. Σε συνδυασμό με το ΔΜΣ προσφέρει ένα χρήσιμο στοιχείο που συνδέεται άμεσα με τον κίνδυνο υγείας που διατρέχει το άτομο εξαιτίας της παχυσαρκίας. Η περίμετρος της μέσης αποτελεί τη μικρότερη περίμετρο της κοιλιακής χώρας (στο ύψος του ομφαλού) και μετριέται σε εκατοστά. Όταν ξεπερνά τα 102 εκατοστά στους άνδρες και τα 88 εκατοστά στις γυναίκες, υπάρχουν αυξημένες πιθανότητες για προβλήματα υγείας.

(3) Περίμετρος ισχίων (Hips Circumference – HC). Η περίμετρος των ισχίων αποτελεί τη μεγαλύτερη περίμετρο στην περιοχή των γλουτών, αποτελεί ένα δείκτη του λίπους που είναι συγκεντρωμένο στην περιοχή και μετριέται επίσης σε εκατοστά.

(4) Ολική χοληστερόλη (Total Cholesterol – chl). Η χοληστερόλη είναι βασικό δομικό συστατικό της μεμβράνης των κυττάρων και μετριέται σε mg/dL. Υψηλά επίπεδα χοληστερόλης όμως επιβαρύνουν τον οργανισμό λόγω εναπόθεσής της στα τοιχώματα των αγγείων. Αν κάποιο άτομο έχει τιμή ολικής χοληστερόλης 200 mg/dL και πάνω, υπάρχουν αυξημένοι κίνδυνοι για έμφραγμα του μυοκαρδίου και εγκεφαλικό επεισόδιο.

(5) Τριγλυκερίδια (Triglyceride - tg). Τα τριγλυκερίδια είναι λιπαρές χημικές ενώσεις που παίζουν πολύ σημαντικό ρόλο στο μεταβολισμό. Χρησιμεύουν στη μεταφορά των λιπαρών οξέων στο αίμα. Τα λιπαρά οξέα χρησιμοποιούνται για την παραγωγή και την αποθήκευση ενέργειας. Όμως, υψηλά επίπεδα τριγλυκερίδιων στο αίμα, συνδέονται με την αθηροσκλήρυνση και, συνεπώς, με τον κίνδυνο εμφάνισης καρδιοαγγειακών παθήσεων. Τα φυσιολογικά επίπεδα τριγλυκεριδίων είναι κάτω από 150 mg/dL. Οριακά επίπεδα θεωρούνται οι τιμές 150 – 199 mg/dL, υψηλά επίπεδα οι τιμές 200 – 499 mg/dL και πολύ υψηλά όταν ξεπερνάνε τα 500 mg/dL.

(6) Λιποπρωτεΐνες υψηλής πυκνότητας (High Density Lipoproteins - HDL). Είναι η λεγόμενη "καλή" χοληστερίνη. Είναι ένα είδος λιποπρωτεΐνης που μεταφέρει λιπίδια στο συκώτι όπου καταστρέφονται και απεκκρίνονται από τον οργανισμό, προστατεύοντας με αυτόν τον τρόπο τα αγγεία από αθηροσκλήρυνση. Άτομα με χαμηλές τιμές HDL έχουν υψηλό κίνδυνο στεφανιαίας νόσου, έστω κι αν έχουν μειωμένη τιμή ολικής χοληστερίνης. Μετριέται σε mg/dL.

(7) Λιποπρωτεΐνες χαμηλής πυκνότητας (Low Density Lipoproteins - LDL). Είναι άλλο ένα είδος λιποπρωτεΐνης η υψηλή συγκέντρωση του οποίου σχετίζεται με αθηροσκλήρυνση,

και όλες τις συνέπειές της, που έχουν ήδη περιγραφεί. Για το λόγο αυτό της έχει δοθεί η ονομασία "κακή" χοληστερίνη. Μετριέται σε mg/dL.

(8) Κρεατινίνη (Creatinine). Η κρεατινίνη είναι άχρηστο αζωτούχο παραπροϊόν του μεταβολισμού, και συγκεκριμένα της κρεατίνης. Παράγεται και αποβάλλεται από τον οργανισμό σε καθημερινή βάση. Η άνοδος των επιπέδων της κρεατινίνης είναι δείγμα κακής κυκλοφορία αίματος και κακής λειτουργίας του ουροποιητικού συστήματος. Φυσιολογικές τιμές θεωρούνται οι 0,8 - 1,4 mg/dL.

(9) Ουρία (Urea). Η ουρία είναι το τελικό προϊόν του μεταβολισμού των πρωτεϊνών. Συντίθεται στο ήπαρ και από εκεί περνάει στο αίμα και απεκκρίνεται από τα νεφρά. Τα επίπεδά της στο αίμα αποτελούν ένδειξη της νεφρικής λειτουργίας. Φυσιολογικές τιμές θεωρούνται οι 7 - 20 mg/dL.

(10) Ουρικό οξύ (Uric acid). Το ουρικό οξύ είναι το τελικό μεταβολικό προϊόν των πουρινών (δομικών μονάδων του RNA και του DNA). Το μεγαλύτερο μέρος του παράγεται στο ήπαρ και στη συνέχεια απεκκρίνεται από τα νεφρά. Αυξημένα επίπεδα ουρικού οξέος στο αίμα μπορεί να προκύψουν όταν υπάρχει σημαντική καταστροφή κυττάρων που περιέχουν πουρίνες ή όταν υπάρχει πρόβλημα απέκκρισής του από τα νεφρά. Φυσιολογικές τιμές θεωρούνται οι 3,5 - 7,0 mg/dL.

(11) Γλυκοζυλιωμένη αιμοσφαιρίνη (Glycosylated Hemoglobin). Η γλυκοζυλιωμένη αιμοσφαιρίνη (HbA1c) παράγεται από τη συμπύκνωση μορίων γλυκόζης με αμινομάδες της αιμοσφαιρίνης, μία αντίδραση που ονομάζεται μη ενζυματική γλυκοζυλίωση (γιατί λαμβάνει χώρα χωρίς την παρουσία ενζύμου) και συμβαίνει σε όλο το χρόνο ζωής των ερυθρών αιμοσφαιρίων. Αποτελεί ένα δείκτη μακροπρόθεσμης αξιολόγησης της ρύθμισης της γλυκόζης και της πορείας του διαβήτη, καθώς αντανακλά τα επίπεδα της γλυκόζης που αντιστοιχούν στους τελευταίους μήνες. Συγκεκριμένα, η τιμή της εξαρτάται από τη μέση συγκέντρωση της γλυκόζης του αίματος τις τελευταίες 8 με 10 εβδομάδες. Ουσιαστικά είναι το ποσοστό της αιμοσφαιρίνης που έχει υποστεί γλυκοζυλίωση (επί τοις εκατό, %) και οι φυσιολογικές τιμές της είναι 5-8%.

(12) Γλυκόζη αίματος (Blood Glucose Levels). Είναι η συγκέντρωση της γλυκόζη που υπάρχει στο αίμα. Έχουμε ήδη αναφερθεί στη γλυκόζη, στη σημασία της και στις συνέπειες που έχει η διαταραχή των επιπέδων της στον ανθρώπινο οργανισμό. Τα επίπεδα της γλυκόζης έχουν αρκετές διακυμάνσεις κατά τη διάρκεια της ημέρας. Φυσιολογικά επίπεδα της γλυκόζης όταν το άτομο είναι νηστικό θεωρούνται οι 90–130 mg/dL, ενώ τα επίπεδα γλυκόζης μετά το γεύμα δε θα πρέπει να ξεπερνάνε τα 180 mg/dL.

(13) Συστολική αρτηριακή πίεση (Systolic Blood Pressure). Αρτηριακή πίεση είναι η πίεση που ασκεί το αίμα στα τοιχώματα των αγγείων κατά την κυκλοφορία του. Κατά τη διάρκεια κάθε καρδιακού χτύπου η αρτηριακή πίεση κυμαίνεται μεταξύ μίας μέγιστης τιμής (συστολική πίεση) και μίας ελάχιστης (διαστολική πίεση). Συστολική αρτηριακή πίεση

είναι η μέγιστη τιμή της αρτηριακής πίεσης, που συμβαίνει στο τέλος κάθε καρδιακού κύκλου, όταν οι κοιλίες της καρδιάς συστέλλονται. Υψηλές τιμές πίεσης συνδέονται με κίνδυνο εμφάνισης καρδιοπάθειας. Φυσιολογικές τιμές της συστολικής αρτηριακής πίεσης είναι οι 90 – 120 mmHg.

(14) Διαστολική αρτηριακή πίεση (Diastolic Blood Pressure). Διαστολική αρτηριακή πίεση είναι η ελάχιστη τιμή της αρτηριακής πίεσης, που συμβαίνει στην αρχή κάθε καρδιακού κύκλου, όταν οι κοιλίες της καρδιάς διαστέλλονται. Φυσιολογικές τιμές της διαστολικής αρτηριακής πίεσης είναι οι 60 – 80 mmHg.

(15) Το είδος της θεραπείας που ακολουθεί ο ασθενής (Treatment). Το χαρακτηριστικό αυτό παίρνει διάφορες διακριτές τιμές, κάθε μία από τις οποίες αντιστοιχεί σε ένα είδος θεραπευτικής αγωγής. Η αντιστοιχία είναι η εξής: 1=δίαιτα, 2=σουλφονυλουρίες, 3=μεγλιτινίδες, 4=γλιταζόνες, 5=ακαρβόζη, 6=διγουανίδια, 8=σουλφονυλουρίες+διγουανίδια, 9=ινσουλίνη, 10=ινσουλίνη+δισκία, 11=γλιταζόνες+σουλφονυλουρίες,12=γλιταζόνες+διγουανίδια, 13=γλιταζόνες+σουλφονυλουρίες+διγουανίδια.

(16) Αναστολείς του μετατρεπτικού ενζύμου της αγγειοτασίνης – AMEA (Angiotensin-converting Enzyme Inhibitors – ACE inhibitors). Τα φάρμακα αυτά αναστέλλουν την μετατροπή της αγγειοτασίνης Ι σε αγγειοτασίνη ΙΙ. Η αγγειοτασίνη ΙΙ προκαλεί σύσπαση των αρτηριών και αύξηση της έκκρισης μιας άλλης ουσίας, της αλδοστερόνης, η οποία αυξάνει την κατακράτηση νατρίου (αλατιού) και νερού από τα νεφρά [ΣΑΑ+07]. Είναι χρήσιμα φάρμακα σε αρρώστους με υπέρταση, καρδιακή ανεπάρκεια, μετά από έμφραγμα ή με διαβητική νεφροπάθεια. Η τιμή "1" συμβολίζει ότι το άτομο δεν λαμβάνει ΑΜΕΑ, ενώ η τιμή "2" αντιστοιχεί στη λήψη ΑΜΕΑ.

(17) Ανταγωνιστές των Υποδοχέων της Αγγειοτασίνης (Angiotensin II receptor antagonists – ΑΤ). Τα φάρμακα αυτά εμποδίζουν την πρόσδεση της αγγειοτασίνης στη θέση της δράσης της, δηλαδή στους ειδικούς υποδοχείς της αγγειοτασίνης Ι). Είναι χρήσιμα φάρμακα σε αρρώστους με υπέρταση, καρδιακή ανεπάρκεια, υπερτροφία της αριστερής κοιλίας της καρδιάς και διαβητική νεφροπάθεια [ΣΑΑ+07]. Χρησιμοποιείται η τιμή "1" όταν το άτομο δεν λαμβάνει ΑΤ και η τιμή "2" όταν λαμβάνει.

(18) Ανταγωνιστές του ασβεστίου (Calcium Antagonists/Calcium Channel Blockers –CCBs). Εμποδίζουν την είσοδο ασβεστίου στα μυϊκά κύτταρα του τοιχώματος των αρτηριών με αποτέλεσμα αγγειοδιαστολή. Είναι χρήσιμα φάρμακα στην υπέρταση και τη στηθάγχη (πόνος στο στήθος από ισχαιμία της καρδιάς) και σε αρρυθμίες. Χρησιμοποιείται η τιμή "1" όταν το άτομο δεν λαμβάνει CCBs και η τιμή "2" όταν λαμβάνει.

(19) Διουρητικά (Diuretics). Τα διουρητικά φάρμακα αρχικά αυξάνουν το ποσό των ούρων και μειώνουν την κατακράτηση αλατιού και νερού και κατά συνέπεια τον όγκο του αίματος. Σε μακροχρόνια καθημερινή χορήγηση δεν προκαλούν αυξημένη διούρηση, αλλά ελαττώνουν

την αντίσταση των αρτηριών. Είναι πολύ χρήσιμα σε περιπτώσεις όπου απαιτείται συνδυασμός φαρμάκων [ΣΑΑ+07]. Η τιμή "1" σημαίνει ότι το άτομο δεν λαμβάνει διουρητικά και η τιμή "2" ότι λαμβάνει.

(20) Βήτα-αποκλειστές (B-Blockers). Οι βήτα-αποκλειστές μειώνουν τη δράση μιας αγγειοσυσπαστικής ουσίας που λέγεται νοραδρεναλίνη, μπλοκάροντας τη θέση δράσης της (στους βήτα υποδοχείς). Τα φάρμακα αυτά μειώνουν τη συχνότητα των παλμών της καρδιάς (προκαλούν δηλαδή βραδυκαρδία) και την καρδιακή παροχή. Χρησιμοποιούνται στην υπέρταση, μετά από έμφραγμα, σε στηθάγχη, στην καρδιακή ανεπάρκεια και σε αρρυθμίες [ΣΑΑ+07]. Και πάλι, η τιμή "1" συμβολίζει ότι το άτομο δεν λαμβάνει βήτα-αποκλειστές και η τιμή "2" ότι λαμβάνει.

(21) Υπολιπιδαιμική αγωγή (Lipid-lowering therapy). Αναφέρεται στη θεραπεία που ακολουθείται για τη μείωση των υψηλών επιπέδων χοληστερόλης. Οι τιμές που μπορεί να πάρει το χαρακτηριστικό αυτό και η αντίστοιχη σημασία τους είναι οι εξής: 1=δεν ακολουθείται υπολιπιδαιμική αγωγή , 2=στατίνες, 3=φιμπράτες, 4=συνδυασμός στατίνης/εζετιμίμπης.

(22) Ασπιρίνη (Aspirin). Η ασπιρίνη είναι ένα αντιαιμοπεταλιακό φάρμακο που εμποδίζει τη συσσώρευση και τη συγκόλληση των αιμοπεταλίων στο αίμα και τη δημιουργία θρόμβων στις αρτηρίες. Η σημασία της στην πρόληψη εκδήλωσης ενός δεύτερου ισχαιμικού επεισοδίου ή εγκεφαλικού (δευτερογενής πρόληψη), έχει τεκμηριωθεί σε πολυάριθμες μελέτες [Στε10]. Όσον αφορά στην πρόληψη εκδήλωσης ενός πρώτου επεισοδίου στηθάγχης, εμφράγματος ή εγκεφαλικού (πρωτογενής πρόληψη), ο ρόλος της δεν έχει ξεκαθαριστεί πλήρως. Επίσης, εφόσον η ασπιρίνη μειώνει την πηκτικότητα του αίματος μπορεί να είναι μια αποτελεσματική πρόληψη για την αμφιβληστροειδοπάθεια των διαβητικών, αν η λήψη της εφαρμοστεί αρκετά νωρίς [BML01]. Χρησιμοποιούμε την τιμή "1" όταν το άτομο δε λαμβάνει καθόλου ασπιρίνη, την τιμή "2" για λήψη χαμηλής δόσης ασπιρίνης (100mg ημερησίως) και την τιμή "3" για υψηλή δόση ασπιρίνης (325mg ημερησίως).

(23) Φύλλο (Sex). Η τιμή "1" αντιστοιχεί σε άνδρα και η τιμή "2" σε γυναίκα.

(24) Οικογενειακό Ιστορικό (Parents). Χρησιμοποιούμε την τιμή "1" για τα άτομα των οποίων οι γονείς έπασχαν από Σακχαρώδη Διαβήτη και την τιμή "2" για την αντίθετη περίπτωση.

(25) Κάπνισμα (Smoking). Η τιμή "1" χρησιμοποιείται για τους μη καπνιστές, η τιμή "2" για τους καπνιστές και η τιμή "3" για άτομα που κάπνιζαν στο παρελθόν, αλλά δεν καπνίζουν πια.

(26) Διάρκεια του Διαβήτη (Duration of Diabetes Mellitus). Το χρονικό διάστημα από το έτος κατά το οποίο έγινε η διάγνωση μέχρι το έτος κατά το οποίο πραγματοποιούνται οι μετρήσεις.

(27) Ηλικία (Age). Η ηλικία του ατόμου κατά το έτος διεξαγωγής των μετρήσεων.

Για κάθε άτομο του συνόλου δεδομένων, εκτός από τιμές μετρήσεων των διαφόρων χαρακτηριστικών, διαθέτουμε και μία μεταβλητή εξόδου (output/class/label). Η τιμή "0" συμβολίζει ότι το άτομο δεν εμφάνισε αμφιβληστροειδοπάθεια κατά το έτος στο οποίο αντιστοιχούν οι μετρήσεις που διαθέτουμε και η τιμή "1" συμβολίζει την εμφάνιση της επιπλοκής. Με τον όρο "εμφάνισε αμφιβληστροειδοπάθεια" εννοούμε ότι του διαγνώστηκε από ειδικό, μέσω ιατρικής εξέτασης, κάποιο από τα γνωρίσματα της διαβητικής αμφιβληστροειδοπάθειας.

4.2 Προεπεξεργασία των δεδομένων

Το πρώτο βήμα της διαδικασίας επίλυσης του δεδομένου προβλήματος αφορά στην προεπεξεργασία των δεδομένων (Data Preprocessing). Η προεπεξεργασία μπορεί να έχει πολύ σημαντική επίδραση στην απόδοση μίας εφαρμογής μηχανικής μάθησης [KKP06]. Στην περίπτωσή μας, περιλαμβάνει διάφορες αποφάσεις σχετικά με τον "καθαρισμό" των δεδομένων (data cleaning), το χειρισμό τους και τη διακριτοποίηση των δεδομένων (discretization).

4.2.1 Καθαρισμός Δεδομένων και Αποφάσεις

Το κομμάτι αυτό της εργασίας περιλαμβάνει αποφάσεις που πήραμε σχετικά με το πώς θα αντιμετωπίσουμε τα ελλιπή δεδομένα (missing data), μία διαδικασία που ονομάζεται καθαρισμός των δεδομένων, και αποφάσεις σχετικά με το πώς θα χειριστούμε γενικότερα τα δεδομένα που διαθέτουμε, ως εισόδους στο επόμενο στάδιο της υλοποίησης. Η διαδικασία αυτή αποδείχθηκε αρκετά απαιτητική και χρονοβόρα. Αρχικά είχαμε ένα σχετικά χαώδες σύνολο ιατρικών δεδομένων αποτελούμενο από 658 διαβητικά άτομα με μετρήσεις 31 χαρακτηριστικών για 6 έτη παρακολούθησης. Για το κάθε άτομο ξέρουμε αν εμφάνισε ή όχι αμφιβληστροειδοπάθεια και το έτος που την εμφάνισε. Το αρχικό αυτό σύνολο περιείχε ελλείψεις σε πολλούς ασθενείς και σε πολλά χαρακτηριστικά. Ενδεικτικά αναφέρουμε ότι το ποσοστό των ατόμων στα οποία είχαμε έλλειψη σε πάνω από το 80% των μετρήσεων έφτανε το 97%. Για το λόγο αυτό, αλλά και επειδή έχουμε να κάνουμε με ιατρικά δεδομένα και η έλλειψη σχετικών γνώσεων δε μας επιτρέπει ευχέρεια κινήσεων, η διαδικασία της προεπεξεργασίας αποδείχτηκε αρκετά απαιτητική.

Σχετικά με τον καθαρισμό των δεδομένων, υπάρχει ένας αριθμός μεθόδων από τις οποίες μπορεί κανείς να επιλέξει πώς θα χειριστεί τις ελλείψεις των δεδομένων που διαθέτει. Μερικές είναι [KKP06]:

- Μέθοδος Αγνόησης Στιγμιότυπων με άγνωστες τιμές (Method of Ignoring Instances with Unknown Feature Values), η οποία είναι η απλούστερη, αλλά περιορίζει σημαντικά το σύνολο των δεδομένων.
- Μέθοδος "Πιο συνηθισμένη τιμή" (Most Common Feature Value). Η τιμή του χαρακτηριστικού που εμφανίζεται πιο συχνά επιλέγεται ως τιμή όλων των άγνωστων

δεδομένων του χαρακτηριστικού. Η μέθοδος αυτή έχει το μειονέκτημα ότι δίνει ακόμα περισσότερο προβάδισμα στην πιο συχνά εμφανιζόμενη περίπτωση.

- Μέθοδος "Έννοια με την πιο συνηθισμένη τιμή" (Concept Most Common Feature Value). Αυτή τη φορά η τιμή του χαρακτηριστικού που εμφανίζεται πιο συχνά επιλέγεται να συμπληρώσει τις άγνωστες τιμές μόνο των στιγμιότυπων που ανήκουν στην ίδια κλάση με το στιγμιότυπο που έχει την πιο συχνά εμφανιζόμενη τιμή. Με τη μέθοδο αυτή δίνεται προβάδισμα στην πιο συχνή τιμή μέσα στην ίδια κλάση, με αποτέλεσμα να "φαίνεται" πιο έντονα ότι η τιμή αυτή αντιπροσωπεύει την κλάση.
- Αντικατάσταση με το μέσο όρο (Mean substitution). Συμπλήρωση των ελλιπών δεδομένων με το μέσο όρο όσων των διαθέσιμων τιμών. Μία παραλλαγή της μεθόδου είναι αντί να χρησιμοποιηθεί ο "γενικός" μέσος όρος, να χρησιμοποιείται κάθε φορά ο μέσος όρος των στιγμιότυπων που ανήκουν στην ίδια κλάση. Η μέθοδος αυτή δίνει ένα προβάδισμα στη "μέση περίπτωση".
- Μέθοδοι παρεμβολής ή ταξινόμησης (Regression or classification methods): Ανάπτυξη ενός μοντέλου παρεμβολής ή ταξινόμησης, χρησιμοποιώντας όλα τα υπόλοιπα χαρακτηριστικά για την πρόβλεψη των ελλείψεων ενός χαρακτηριστικού. Αυτή η μέθοδος είναι αρκετά πιο περίπλοκη και προϋποθέτει ελάχιστες ελλείψεις στα υπόλοιπα χαρακτηριστικά.
- Hot deck imputation. Αναγνώριση της πιο "όμοιας" περίπτωσης με αυτή με το άγνωστο δεδομένο (όσον αφορά στις τιμές των υπόλοιπων χαρακτηριστικών) και αντικατάσταση της άγνωστης τιμής με την τιμή της "όμοιας" περίπτωσης. Η μέθοδος αυτή προσθέτει μια "πλασματική" αλληλεξάρτηση μεταξύ των χαρακτηριστικών.
- Μέθοδος χειρισμού της άγνωστης τιμής ως "ιδιαίτερης" (Method of Treating Missing Feature Values as Special). Αντιμετώπιση του "άγνωστο" ("unknown") σαν μία νέα τιμή του χαρακτηριστικού.

Στη συνέχεια, αναφέρουμε τις αποφάσεις που λάβαμε σχετικά με τον καθαρισμό των δεδομένων, αλλά και με το χειρισμό τους γενικά:

Για τη Γλυκοζυλιωμένη αιμοσφαιρίνη, τη Γλυκόζη του αίματος, τη Συστολική Αρτηριακή Πίεση και τη Διαστολική Αρτηριακή Πίεση είχαμε 5 μετρήσεις για κάθε έτος. Υπολογίσαμε το μέσο όρο των 5 αυτών μετρήσεων έτσι ώστε να έχουμε και για αυτά τα χαρακτηριστικά μία μέτρηση για κάθε έτος.

Απορρίψαμε τα χαρακτηριστικά για τα οποία είχαμε μέτρηση σε τουλάχιστον 3 στα 6 έτη για λιγότερους από 100 στους 658 ασθενείς. Αυτά ήταν η μικρολευκωματουρία, η συγκέντρωση απολιποπρωτεΐνης Α1, η συγκέντρωση απολιποπρωτεΐνης Β και η συγκέντρωση λιποπρωτεΐνης Α. Θεωρήσαμε ότι ο μεγάλος αριθμός ελλείψεων των χαρακτηριστικών αυτών τα καθιστά όχι μόνο περιττά, αλλά και ενδεχόμενη πηγή αλλοίωσης των αποτελεσμάτων μας. Έχουμε επομένως πλέον 27 χαρακτηριστικά, αυτά που περιγράψαμε στο προηγούμενο κεφάλαιο.

Στη συνέχεια, πάνω στη προσπάθεια διατήρησης ενός όσο το δυνατό πιο αξιόπιστου συνόλου δεδομένων, εξετάσαμε το ενδεχόμενο να έχουμε έναν πληθυσμό τουλάχιστον 100 ατόμων με ελάχιστες ελλείψεις. Στην τιμή της εξόδου (αν εμφάνισε ή όχι αμφιβληστροειδοπάθεια) δεν είχαμε καμία έλλειψη, γεγονός πολύ σημαντικό, καθώς δεν χρειάστηκε να απορρίψουμε δεδομένα για το λόγο αυτό, ούτε να αναρωτηθούμε πώς μπορούμε να συμπληρώσουμε τις ελλείψεις τους. Επίσης, δεν είχαμε καμία έλλειψη στα χαρακτηριστικά: Φύλλο, Οικογενειακό Ιστορικό, Κάπνισμα, Διάρκεια του Διαβήτη και Ηλικία. Από τα υπόλοιπα 22 χαρακτηριστικά εξετάσαμε για ένα πλήθος χαρακτηριστικών από 16 έως 22 σε πόσα άτομα διαθέτουμε τουλάχιστον 3 μετρήσεις στα 6 έτη. Προέκυψαν οι παρακάτω τιμές:

ΠΙΝΑΚΑΣ 4.1: Πλήθος ατόμων με ελλείψεις στο σύνολο των δεδομένων.

Πλήθος χαρακτηριστικών	**Πλήθος ατόμων που έχουν μέτρηση σε τουλάχιστον 3 έτη στο αντίστοιχο πλήθος χαρακτηριστικών**
16	213
17	204
18	201
19	186
20	173
21	150
22	125

Παρατηρούμε ότι για 125 άτομα διαθέτουμε μετρήσεις σε τουλάχιστον 3 από τα 6 έτη για όλα τα χαρακτηριστικά, ενώ για 173 άτομα αυτό συμβαίνει για 20 στα 22 χαρακτηριστικά. Αποφασίσαμε να διατηρήσουμε το σύνολο των 173 ατόμων, καθώς είναι αρκετά μεγαλύτερο και οι ελλείψεις που περιλαμβάνει είναι πολύ λίγες. Τα υπόλοιπα άτομα απορρίφθηκαν (Μέθοδος Αγνόησης Στιγμιότυπων με άγνωστες τιμές).

Έπειτα, προέκυψε ένα ζήτημα σχετικό με το γενικότερο χειρισμό των δεδομένων και τον τρόπο που αυτά θα αποτελέσουν είσοδο για το επόμενο στάδιο της υλοποίησης. Επειδή ο διαβήτης, γενικά, και η διαβητική αμφιβληστροειδοπάθεια αποτελούν καταστάσεις που εξελίσσονται με αργούς ρυθμούς σε βάθος χρόνου, το αποτέλεσμα της εξόδου δεν εξαρτάται μόνο από τις μετρήσεις του τρέχοντος έτους, αλλά από την όλη εξέλιξη του προφίλ του ατόμου. Για παράδειγμα, αν η ρύθμιση των επιπέδων της γλυκόζης ενός ατόμου ήταν ανεπαρκής για κάποια έτη, το γεγονός αυτό ενδέχεται να επηρεάζει την πιθανότητα να εμφανίσει αμφιβληστροειδοπάθεια, ακόμα κι αν το τρέχον έτος η γλυκόζη του βρίσκεται σε φυσιολογικά επίπεδα. Το πρόβλημα που έπρεπε να επιλύσουμε ήταν το πώς θα "απεικονιστεί" αυτή η χρονική εξέλιξη στο σύνολο των δεδομένων μας και θα αξιοποιηθεί από το μοντέλο πρόβλεψης. Για αυτό το λόγο, αποφασίσαμε η τιμή του κάθε χαρακτηριστικού για κάθε έτος να προκύπτει από το μέσο όρο της τιμής του τρέχοντος έτους και των προηγούμενων

ετών. Προφανώς, η Διάρκεια του Διαβήτη και η Ηλικία αυξάνονται κατά 1 κάθε έτος. Έτσι, το τελικό μας σύνολο δεδομένων αποτελείται από 173 άτομα × 6 έτη = 1038 στιγμιότυπα (instances).

Κάθε στιγμιότυπο αποτελείται από τις μετρήσεις των 27 χαρακτηριστικών ενός ατόμου για ένα έτος και την αντίστοιχη έξοδο, το αν δηλαδή εμφάνισε ή όχι την επιπλοκή το συγκεκριμένο έτος.

Οι ελλείψεις των δεδομένων αντιμετωπίστηκαν ως εξής:

- Στα χαρακτηριστικά που παίρνουν συνεχείς τιμές, δηλαδή στα: Δείκτης Μάζας Σώματος, Περίμετρος μέσης, Περίμετρος ισχύων, Ολική χοληστερόλη, Τριγλυκερίδια, Λιποπρωτεΐνες υψηλής πυκνότητας, Λιποπρωτεΐνες χαμηλής πυκνότητας, Κρεατινίνη, Ουρία, Ουρικό οξύ, Γλυκοζυλιωμένη αιμοσφαιρίνη, Γλυκόζη αίματος, Συστολική αρτηριακή πίεση, Διαστολική αρτηριακή πίεση, η τιμή όποιου έτους δεν είχαμε μέτρηση συμπληρώθηκε με το μέσο όρο των τιμών των προηγούμενων ετών του ατόμου στο συγκεκριμένο χαρακτηριστικό. Στις ελάχιστες περιπτώσεις που είχαμε έλλειψη μέτρησης στο πρώτο έτος, συμπληρώσαμε με την τιμή του επόμενου έτους.
- Στα χαρακτηριστικά που παίρνουν διακριτές τιμές, δηλαδή στα: Είδος της θεραπείας που ακολουθείται (Treatment), Λήψη Αναστολέων του μετατρεπτικού ενζύμου της αγγειοτασίνης (ΑΜΕΑ), Λήψη Ανταγωνιστών των Υποδοχέων της Αγγειοτασίνης (ΑΤ), Λήψη Ανταγωνιστών ασβεστίου, Λήψη Διουρητικών, Λήψη Βήτα-αποκλειστών (B-Blockers), Υπολιπιδαιμική αγωγή, Λήψη Ασπιρίνης (Aspirin), η συμπλήρωση με το μέσο όρο δεν έχει νόημα. Για παράδειγμα, ας θεωρήσουμε κάποιο άτομο που για μία χρονιά ακολουθούσε μία συγκεκριμένη δίαιτα για τη ρύθμιση των επιπέδων της γλυκόζης του (treat=1) και τη δεύτερη χρονιά του χορηγήθηκε ινσουλίνη (treat=9). Αν η θεραπεία που ακολούθησε την επόμενη χρονιά μας είναι άγνωστη, θα ήταν λάθος να αντικαταστήσουμε αυτή την έλλειψη με το μέσο όρο των προηγούμενων ετών, γιατί κάτι τέτοιο μας δίνει (1+9)/2=5, που αντιστοιχεί σε λήψη ακαρβόζης, σε ένα άσχετο δηλαδή αποτέλεσμα. Αν παρατηρήσουμε τις τιμές που παίρνουν τα χαρακτηριστικά αυτά, θα προσέξουμε ότι υπάρχει μία διαβάθμιση από το πιο "ελαφρύ" στο πιο "βαρύ". Στα περισσότερα χαρακτηριστικά, δηλαδή, η τιμή "1" αντιστοιχεί σε "μη λήψη" και η τιμή "2" σε "λήψη". Με το σκεπτικό ότι κάθε ενέργεια που ακολουθείται έχει μακροπρόθεσμη επίπτωση, θεωρήσαμε σωστό να αντικαταστήσουμε τις ελλείψεις αυτών των χαρακτηριστικών με την τιμή του προηγούμενου έτους του χαρακτηριστικού. "Συμπληρώνουμε" δηλαδή κάθε φορά την τακτική που ακολούθησε το άτομο την προηγούμενη χρονιά.

  Επισημαίνουμε στο σημείο αυτό ότι μετά το φιλτράρισμα που είχε προηγηθεί οι ελλείψεις που είχαμε ήταν ελάχιστες, επομένως οι αντικαταστήσεις που πραγματοποιήσαμε, με τον καλύτερο κάθε φορά δυνατό τρόπο, θεωρούμε ότι δεν επηρεάζουν πολύ το τελικό αποτέλεσμα της μελέτης μας.

- Στο χαρακτηριστικό “Λήψη Ανταγωνιστών των Υποδοχέων της Αγγειοτασίνης (ΑΤ)” για 12 άτομα στα 173 δεν είχαμε καμία μέτρηση για κανένα έτος. Στις περιπτώσεις αυτές συμπληρώσαμε το κάθε έτος με το μέσο όρο των τιμών των υπόλοιπων ατόμων για το συγκεκριμένο έτος (Mean substitution). Επειδή πρόκειται για διακριτό χαρακτηριστικό, αποτελέσματα μικρότερα από “1,5” τα θεωρήσαμε ίσα με “1”, και ίσα με “2” στην αντίθετη περίπτωση.

4.2.2 Διακριτοποίηση (Discretization)

Διακριτοποίηση ονομάζεται η διαδικασία μετατροπής συνεχών τιμών σε διακριτές, ή αλλιώς η κατάταξη συνεχών τιμών σε κάποια διαστήματα τιμών (bins). Ένα παράδειγμα είναι η κατάταξη της ηλικίας 53 ετών στο διάστημα 50-55, και η αντιμετώπιση όλων των ηλικιών που το διάστημα αυτό περιλαμβάνει σαν μία τιμή. Η διακριτοποίηση θεωρείται ότι βελτιώνει την απόδοση μίας εφαρμογής μηχανικής μάθησης, καθώς μεγάλο πλήθος διαφορετικών τιμών καθιστά πιο χρονοβόρα την υπολογιστική διαδικασία και ενδέχεται να παρεμποδίζει τη σύλληψη της πληροφορίας που εμπεριέχουν τα δεδομένα εκπαίδευσης, μειώνοντας έτσι την αποτελεσματικότητα της εφαρμογής. Ειδικά στην περίπτωσή μας η διακριτοποίηση κρίθηκε απαραίτητη, καθώς η διαδικασία του υπολογισμού της Αμοιβαίας Πληροφορίας συνεχών μεταβλητών είναι όχι μόνο πολύ πιο περίπλοκη και πρακτικά αδύνατη [KC02], καθώς περιλαμβάνει την εκτίμηση της συνάρτησης πυκνότητας πιθανότητας για κάθε χαρακτηριστικό, αλλά συνοδεύεται και από απώλεια πληροφορίας, καθώς οι μέθοδοι που χρησιμοποιούνται για τη συνάρτηση πυκνότητας πιθανότητας είναι προσεγγιστικοί. Για όλους τους παραπάνω λόγους, η διακριτοποίηση θεωρήθηκε σκόπιμη.

Το πρόβλημα της επιλογής των σωστών ορίων των διαστημάτων της διακριτοποίησης και του σωστού πλήθους τους παραμένει ένα ανοιχτό ζήτημα. Οι αλγόριθμοι διακριτοποίησης μπορούν να χωριστούν σε χωρίς επίβλεψη (unsupervised algorithms) και με επίβλεψη (supervised algorithms) [R97]. Οι αλγόριθμοι χωρίς επίβλεψη διακριτοποιούν τα χαρακτηριστικά χωρίς να λαμβάνουν υπόψη τις κλάσεις, ενώ το αντίθετο συμβαίνει στους αλγόριθμους με επίβλεψη, όπου η κλάση στην οποία αντιστοιχεί η κάθε τιμή επηρεάζει την επιλογή του διαστήματος στο οποίο θα καταταχθεί.

Η απλούστερη μέθοδος διακριτοποίησης ονομάζεται “διακριτοποίηση σταθερού μεγέθους” (equal size discretization) και ανήκει στην κατηγορία των μεθόδων χωρίς επίβλεψη. Υπολογίζεται αρχικά η μέγιστη και η ελάχιστη τιμή του προς διακριτοποίηση χαρακτηριστικού και το εύρος αυτό χωρίζεται σε n ισομεγέθη διαστήματα. Η επιλογή του k σχετίζεται με την επιθυμητή ακρίβεια. Μία άλλη μέθοδος χωρίς επίβλεψη είναι η “διακριτοποίηση σταθερής συχνότητας” (equal frequency discretization). Σύμφωνα με αυτή, χωρίζουμε τις τιμές του χαρακτηριστικού σε διαστήματα που περιέχουν ίσο πλήθος τιμών.

Η μέθοδος που ακολουθήθηκε στην παρούσα διπλωματική είναι η διακριτοποίηση σταθερού μεγέθους, αλλά με έναν ιδιαίτερο τρόπο εύρεσης του πλήθους n των διαστημάτων, και όχι

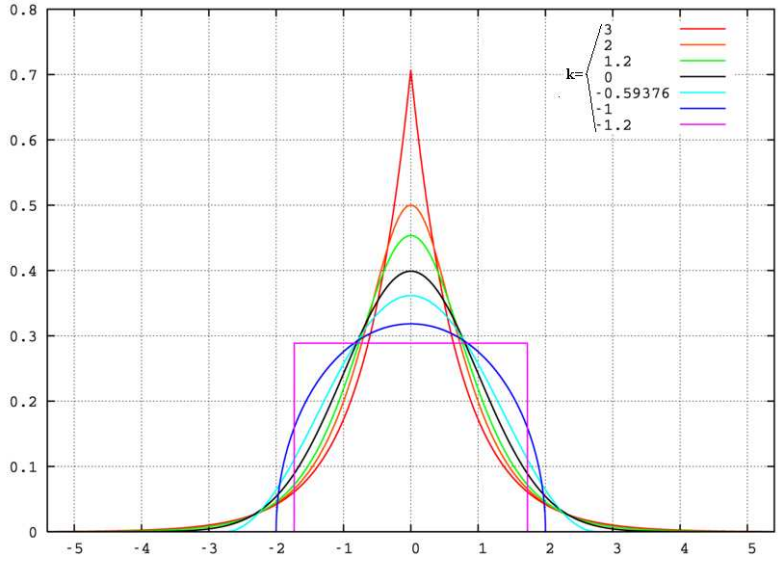
εμπειρικά, όπως είθισται. Οι μέθοδοι με επίβλεψη απορρίφθηκαν, καθώς θα κατεύθυναν την ταξινόμηση, η οποία αποτελεί τον στόχο μας. Για τον υπολογισμό του n ακολουθήθηκε η μέθοδος που βασίζεται στον κανόνα του Doane (Doane's rule) έτσι όπως περιγράφεται στην εργασία των G. D. Tourassi, E. D. Frederick, M. Markey και G. E. Floyd, Jr "Application of the mutual information criterion for feature selection in computer-aided diagnosis" [TFM+01] και που συχνά χρησιμοποιείται σε εφαρμογές αναγνώρισης φωνής. Έχει προταθεί ότι για δεδομένα που δεν ακολουθούν την κατανομή Gauss (κανονική κατανομή), το κατάλληλο πλήθος των διαστημάτων διακριτοποίησης δίνεται από τη σχέση:

$$n = \log_2 N + 1 + \log_2(1 + k\sqrt{N/6})$$

, όπου k είναι η κύρτωση (kurtosis) και N το πλήθος των δειγμάτων (στιγμιότυπων). Η κύρτωση αποτελεί ένα μέτρο της "αιχμηρότητας" μίας κατανομής. Η κύρτωση υπολογίζεται από τον τύπο:

$$k = \frac{E(x-\mu)^2}{\sigma^4} - 3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2)^2} - 3$$

Στο παρακάτω σχήμα φαίνεται η τιμή της κύρτωσης μερικών κατανομών:



Σχήμα 4.1: Γραφική αναπαράσταση γνωστών κατανομών και η αντίστοιχη τιμή της κύρτωσης.

Για δεδομένα που ακολουθούν την κατανομή Gauss έχουμε k=0, επομένως ο κατάλληλος αριθμός των διαστημάτων είναι:

$$n = \log_2 N + 1$$

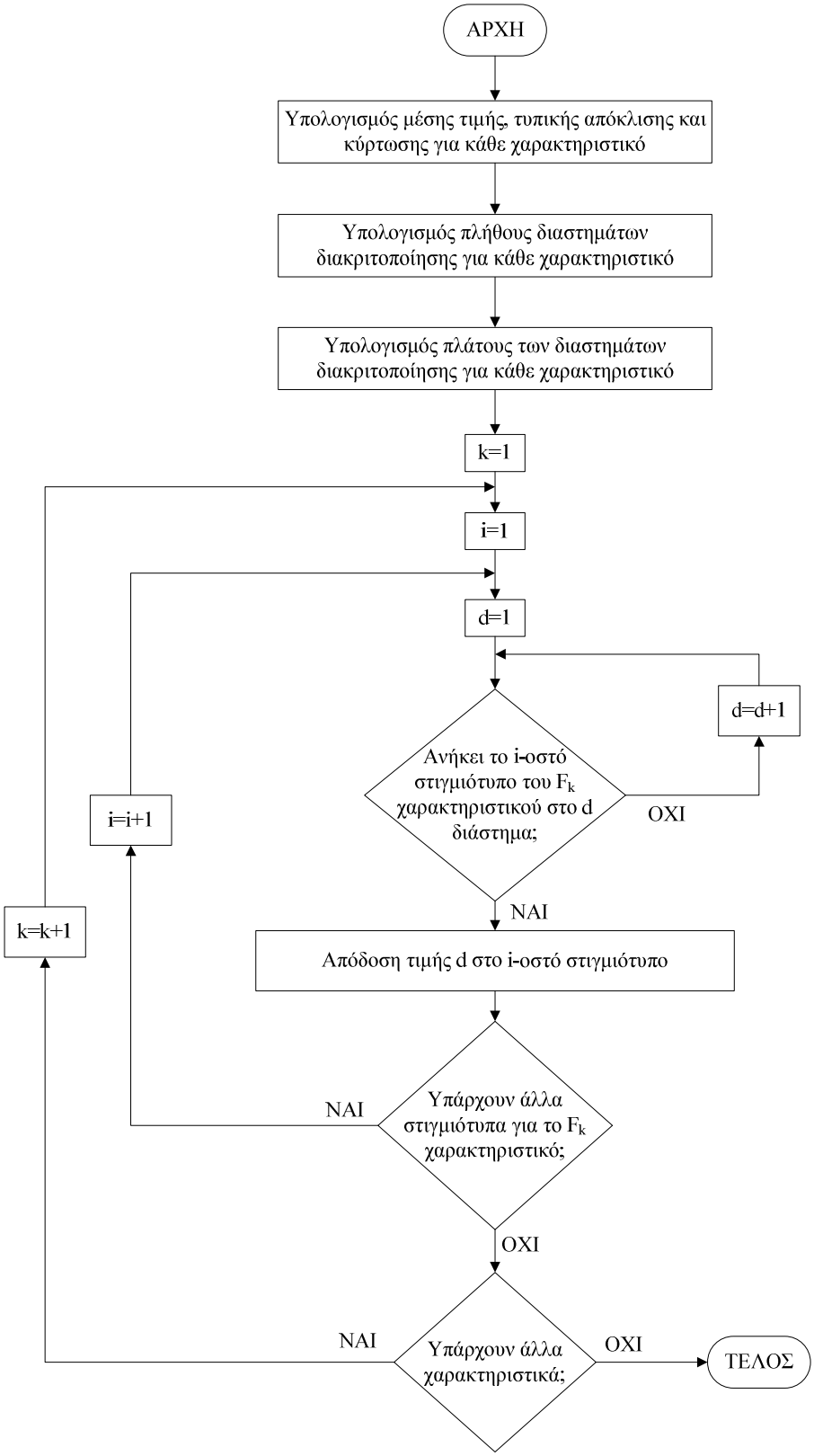
Η μέθοδος που ακολουθήθηκε περιλαμβάνει των υπολογισμό της μέσης τιμής (mean – μ) και της τυπικής απόκλισης (standard deviation – σ) κάθε χαρακτηριστικού. Στη συνέχεια, το διάστημα [μ-2σ, μ+2σ] (στο οποίο στις κανονικές τιμές περιλαμβάνεται περίπου το 95% του πληθυσμού) χωρίζεται σε n ίσα μέρη, όπου το n υπολογίζεται ξεχωριστά για κάθε χαρακτηριστικό βάσει των παραπάνω σχέσεων. Κάθε τιμή κατατάσσεται μέσα στο διάστημα στο οποίο ανήκει. Υπάρχουν δύο

επιπλέον διαστήματα για τυχόν ακραίες τιμές που βρίσκονται έξω από το διάστημα [μ-2σ, μ+2σ]. Με τη μέθοδο αυτή, η "αιχμηρότητα" της κατανομής των τιμών κάθε χαρακτηριστικού χρησιμοποιείται σαν ένα μέτρο εκτίμησης της "ακρίβειας" που απαιτείται για τη διακριτοποίησή του.

Η παραπάνω λογική ακολουθήθηκε για όλα τα δεδομένα που παίρνουν συνεχείς τιμές, στο κάθε ένα ξεχωριστά. Στα δεδομένα που είναι από τη φύση τους διακριτά (λήψη/μη λήψη) και στην έξοδο, προφανώς δεν ακολουθήθηκε αυτή η διαδικασία. Επίσης, από τη διαδικασία αυτή εξαιρέθηκε η Διάρκεια του Διαβήτη, γιατί οι δυνατές τιμές αυτού του χαρακτηριστικού δεν ήταν πολλές και μπορεί να θεωρηθεί ήδη διακριτό, αλλά και γιατί, θεωρώντας ότι ακόμα κι ένα έτος παραπάνω ίσως να έχει σημασία για την έξοδο, η εφαρμογή της παραπάνω τακτικής θα μας οδηγούσε σε απώλεια ακρίβειας.

Με την εφαρμογή αυτής της μεθόδου διακριτοποίησης πετύχαμε το εξής: τα χαρακτηριστικά των οποίων η κατανομή είναι πιο "αιχμηρή", που παρουσιάζει δηλαδή πιο έντονες κορυφές και διακυμάνσεις, και που επομένως χρειάζεται μεγαλύτερη ακρίβεια για την περιγραφή της, το πλήθος n των διαστημάτων ήταν μεγαλύτερο και το μέγεθος των διαστημάτων μικρότερο από ότι σε άλλα χαρακτηριστικά. Επίσης, κάθε χαρακτηριστικό διακριτοποιήθηκε όπως ταιριάζει στη "φύση" του. Για παράδειγμα, το χαρακτηριστικό "Τριγλυκερίδια" έχει μεγάλο εύρος τιμών (ελάχιστη τιμή 36 mg/dL και μέγιστη 927 mg/dL) και βάσει όσων ειπώθηκαν στο προηγούμενο κεφάλαιο σχετικά με το διαχωρισμό των τιμών σε φυσιολογικές και μη, δεν χρειάζεται μεγάλη ακρίβεια κατά τη διακριτοποίηση τους. Στο χαρακτηριστικό αυτό χρησιμοποιήθηκαν 19 μεγάλα διαστήματα. Αντίθετα, στη "Γλυκοζιωμένη Αιμοσφαιρίνη", παρόλο που το εύρος των τιμών είναι πολύ μικρότερο (ελάχιστη τιμή 4,30% και μέγιστη 11,47%) χρησιμοποιήθηκαν 15 διαστήματα, η διακριτοποίηση δηλαδή πραγματοποιήθηκε με πολύ μεγαλύτερη ακρίβεια, ακριβώς όπως έπρεπε λόγω της ίδιας της "φύσης" του χαρακτηριστικού (οι φυσιολογικές τιμές του είναι 4,8-6,1%).

Ακολουθεί ένα σχηματικό διάγραμμα που συνοψίζει τη διαδικασία της διακριτοποίησης που ακολουθήσαμε:



Σχήμα 4.2: Διάγραμμα ροής διακριτοποίησης.

4.3 Επιλογή Χαρακτηριστικών (1^{ο} Στάδιο): Γενετικός Αλγόριθμος

Το επόμενο βήμα αφορά στην υλοποίηση της επιλογής χαρακτηριστικών (feature selection). Για την επίλυση του προβλήματος ακολουθήθηκε η ιδέα του συνδυασμού της τεχνικής-φίλτρο (filter approach/technique) και της τεχνικής-περιτύλιγμα (wrapper approach/technique), με σκοπό την εκμετάλλευση των πλεονεκτημάτων και της πρώτης και της δεύτερης. Για αυτό το λόγο η επιλογή των χαρακτηριστικών υλοποιείται σε δύο στάδια. Πραγματοποιείται αρχικά ένα "προπαρασκευαστικό" στάδιο που ακολουθεί τη filter λογική και που αναλύεται σε αυτή την ενότητα. Η όλη διαδικασία ολοκληρώνεται στο δεύτερο στάδιο, που ακολουθεί τη wrapper λογική και με το οποίο θα ασχοληθούμε στην επόμενη ενότητα.

Για τη υλοποίηση της filter λογικής χρησιμοποιήθηκε ένας Γενετικός Αλγόριθμος. Μετά την ολοκλήρωση του σταδίου της Προεπεξεργασίας, έχουμε πλέον ένα σύνολο δεδομένων αποτελούμενο από 1038 στιγμιότυπα, 27 διακριτά χαρακτηριστικά και 1, επίσης διακριτή, μεταβλητή εξόδου. Το σύνολο αυτό αποτελεί την είσοδο του γενετικού αλγορίθμου. Κάθε χρωμόσωμα του γενετικού αλγορίθμου αντιστοιχεί σε ένα υποψήφιο υποσύνολο χαρακτηριστικών και αποτελείται από 27 μεταβλητές-ψηφία, όσα είναι και τα χαρακτηριστικά που διαθέτουμε. Κάθε μεταβλητή μπορεί να πάρει δύο τιμές: την τιμή "0" που συμβολίζει την απουσία του χαρακτηριστικού που αντιπροσωπεύει η μεταβλητή από το υποσύνολο και την τιμή "1" που συμβολίζει την παρουσία του χαρακτηριστικού.

4.3.1 Συνάρτηση καταλληλότητας

Η συνάρτηση καταλληλότητας (fitness function) του γενετικού αλγορίθμου, το κριτήριο δηλαδή για την αξιολόγηση του κάθε χρωμοσώματος – υποψήφιου υποσυνόλου περιλαμβάνει την έννοια της Αμοιβαίας Πληροφορίας και στηρίζεται σε μία αντιμετώπιση που έχει ακολουθηθεί από πολλούς ερευνητές ([ADC03], [Bro09], [DBR+10], [HCX07], [KC02], [LSL+09], [PLD05], [SP10], [TFM+01]). Σύμφωνα με αυτή, ένα "καλό" υποσύνολο χαρακτηριστικών είναι αυτό που περιλαμβάνει χαρακτηριστικά πολύ "συσχετισμένα" (related/predictive of) με την έξοδο και "ασυσχέτιστα" (unrelated/not predictive of) μεταξύ τους. Η έννοια της "συσχέτισης" μετράται με όρους αμοιβαίας πληροφορίας. Όταν η αμοιβαία πληροφορία μεταξύ δύο μεταβλητών παίρνει υψηλές τιμές σημαίνει ότι οι δύο αυτές μεταβλητές συνδέονται στενά και ότι η μία παρέχει πληροφορία για την άλλη. Αντίθετα, όταν η αμοιβαία πληροφορία τους είναι μηδέν, σημαίνει ότι οι δύο αυτές μεταβλητές δε συνδέονται ή είναι ανεξάρτητες η μία από την άλλη. Οι δύο αυτές συνθήκες για το χαρακτηρισμό ενός υποσυνόλου ως "καλό" συμπεριλαμβάνονται στη συνάρτηση καταλληλότητας ως εξής:

Έστω ένα υποψήφιο υποσύνολο S που αποτελείται από n χαρακτηριστικά. Για κάθε χαρακτηριστικό f_i, i=1,…,n, του υπό εξέταση υποσυνόλου S υπολογίζεται η αμοιβαία πληροφορία του με την έξοδο

Ο, δηλαδή η ποσότητα $I(f_i;O)$. Στη συνέχεια, υπολογίζεται η αμοιβαία πληροφορία του με καθένα από τα υπόλοιπα f_j, $j=1,\ldots,n$ και $j \neq i$ χαρακτηριστικά που ανήκουν στο υποσύνολο, δηλαδή οι ποσότητες $I(f_i;f_j)$. Επισημαίνουμε στο σημείο αυτό ότι, όπως εξηγήσαμε στο 2ο κεφάλαιο, ισχύει ότι $I(f_i;f_j) = I(f_j;f_i)$, επομένως υπολογίζεται κάθε φορά η αμοιβαία πληροφορία κάθε χαρακτηριστικού με τα "επόμενά" του μόνο. Οι παραπάνω ποσότητες αθροίζονται, υπολογίζονται δηλαδή τα αθροίσματα: $\sum_{i=1}^{n} I(f_i; O)$ και $\sum_{i=1}^{n}(\sum_{j=i+1}^{n} I(f_i; f_j))$ και η τελική μορφή της συνάρτησης καταλληλότητας έχει ως εξής:

$$F(S) = \sum_{i=1}^{n} I(f_i; O) - \beta \sum_{i=1}^{n}(\sum_{j=i+1}^{n} I(f_i; f_j))$$

Όπου β μία παράμετρος που χρησιμοποιείται για να ρυθμίσει τη βαρύτητα που δίνεται στην αμοιβαία πληροφορία μεταξύ των χαρακτηριστικών, ή αλλιώς στον πλεονασμό (redundancy) χαρακτηριστικών, σε σχέση με την πληροφορία που παρέχουν για την έξοδο. Μπορεί να πάρει τιμές από 0 μέχρι 1. Η ανάθεση β=0 είναι ισοδύναμη με την επιλογή χαρακτηριστικών ανεξάρτητα, ενώ η χρήση μεγαλύτερης τιμής δίνει την ανάλογη έμφαση στην επιθυμία μείωσης των ενδο-χαρακτηριστικών εξαρτήσεων [Bro09]. Δεν μπορούμε να ξέρουμε ποια είναι η "ιδανική" τιμή για το β [HCX07], και αυτή πρέπει να οριστεί εμπειρικά.

Στην περίπτωσή μας, θεωρήσαμε ότι πρέπει να δοθεί πολύ μεγαλύτερη σημασία στην πληροφορία που παρέχει το υποσύνολο για την έξοδο, ακόμα κι αν αυτό συνεπάγεται τη συμπερίληψη μερικών αλληλεξαρτήσεων. Μετά από διάφορες δοκιμές και με κριτήριο διάφορα επιπλέον ζητήματα, όπως το πόσο απότομες ήταν κάθε φορά οι διακυμάνσεις της συνάρτησης, πόσο επιβαρύνονται τα μεγάλα υποσύνολα και πως θα εξομαλυνθεί η διαφορά στην τάξη μεγέθους των δύο όρων, καταλήξαμε στην τιμή β=0,03.

Καλά υποσύνολα είναι αυτά που πετυχαίνουν όσο το δυνατό μεγαλύτερη τιμή στην παραπάνω συνάρτηση. Στην υλοποίησή μας στόχος ήταν η ελαχιστοποίηση της –F(S), κάτι που αποτελεί ισοδύναμη ενέργεια.

Ο λόγος που χρησιμοποιήθηκε αυτή η συνάρτηση καταλληλότητας είναι το γεγονός ότι αποτελεί ένα αρκετά αξιόπιστο κριτήριο για την επιλογή των σημαντικών χαρακτηριστικών. Ένας ακόμη λόγος είναι ότι άλλα, πιο συνηθισμένα στατιστικά μέτρα (συσχέτισης, συμμεταβλητότητας, κ.α.) υποθέτουν γραμμικές εξαρτήσεις μεταξύ χαρακτηριστικών και εξόδου και χαρακτηριστικών μεταξύ τους, κάτι που δεν μπορούμε να υποθέσουμε για το πρόβλημα που θέλουμε να επιλύσουμε [BM09]. Επίσης, με τον τρόπο που τη χρησιμοποιούμε, ξεφεύγουμε από τους περιορισμούς της σειριακής πρόσθεσης ή αφαίρεσης χαρακτηριστικών και επιτρέπουμε μία πιο "σφαιρική" αξιολόγηση, καθώς εξετάζουμε "συνολικά" το κάθε υποσύνολο και όχι το ενδεχόμενο προσθαφαίρεσης του κάθε χαρακτηριστικού ξεχωριστά.

Η συνάρτηση καταλληλότητας δέχεται ως είσοδο ένα χρωμόσωμα και επιστρέφει έναν αριθμό που υποδηλώνει το βαθμό καταλληλότητας του. Η διαδικασία που περιγράψαμε επαναλαμβάνεται για κάθε χρωμόσωμα του πληθυσμού, σε κάθε γενεά. Η ποσότητες $I(f_i;O)$ και $I(f_i;f_j)$ αποθηκεύονται μόνιμα σε μεταβλητές, έτσι ώστε να μην χρειάζεται κάθε φορά ο επανυπολογισμός τους. Η πολυπλοκότητα της συνάρτησης καταλληλότητας είναι $O(n^2)$, όπου n το πλήθος των χαρακτηριστικών που αποτελούν είσοδο της συνάρτησης.

4.3.2 Υλοποίηση Γενετικού Αλγορίθμου

Στην υλοποίησή μας ο γενετικός αλγόριθμος δέχεται σαν είσοδο έναν πληθυσμό αποτελούμενο από 20 χρωμοσώματα. Πραγματοποιούνται συνολικά 100 επαναλήψεις – γενιές. Σε κάθε επανάληψη υπολογίζεται η συνάρτηση καταλληλότητας για το κάθε χρωμόσωμα και η επιλογή των χρωμοσωμάτων που θα περάσουν κατευθείαν στην επόμενη γενιά (elite) και αυτών που θα γίνουν γονείς (parents) γίνεται βάσει του σκορ που πετυχαίνει το κάθε χρωμόσωμα. Καλύτερα θεωρούνται αυτά που έχουν χαμηλότερη τιμή συνάρτησης καταλληλότητας (–F(S)).

Η επιλογή της χρήσης γενετικού αλγορίθμου στη μεθοδολογία μας οφείλεται στο γεγονός ότι οι γενετικοί αλγόριθμοι αποτελούν ένα πολύ εύρωστο εργαλείο αναζήτησης σε περίπλοκα προβλήματα, γεγονός που έχει αποδειχθεί και θεωρητικά και εμπειρικά. Είναι υπολογιστικά απλοί και παρ' όλα αυτά ισχυροί στην αναζήτησή τους για βελτίωση. Ο παράγοντας "τύχη" που περιλαμβάνουν τους δίνει προβάδισμα σε σχέση με άλλους αλγόριθμους αναζήτησης και τους καθιστά λιγότερο ευάλωτους στα τοπικά ελάχιστα.

Μετά από βιβλιογραφική έρευνα, βασιζόμενοι σε εργασίες που υιοθετούν τη χρήση γενετικών αλγορίθμων ([HRK+99], [IML+01], [IN00], [G86], [KO08], [SS89], [SR06], [TFM+01], [YH97]), αλλά και σε όσα περιγράψαμε στο 2^o κεφάλαιο σχετικά με την επιλογή των βέλτιστων παραμέτρων, ορίσαμε στην υλοποίησή μας τις παραμέτρους του αλγορίθμου ως εξής:

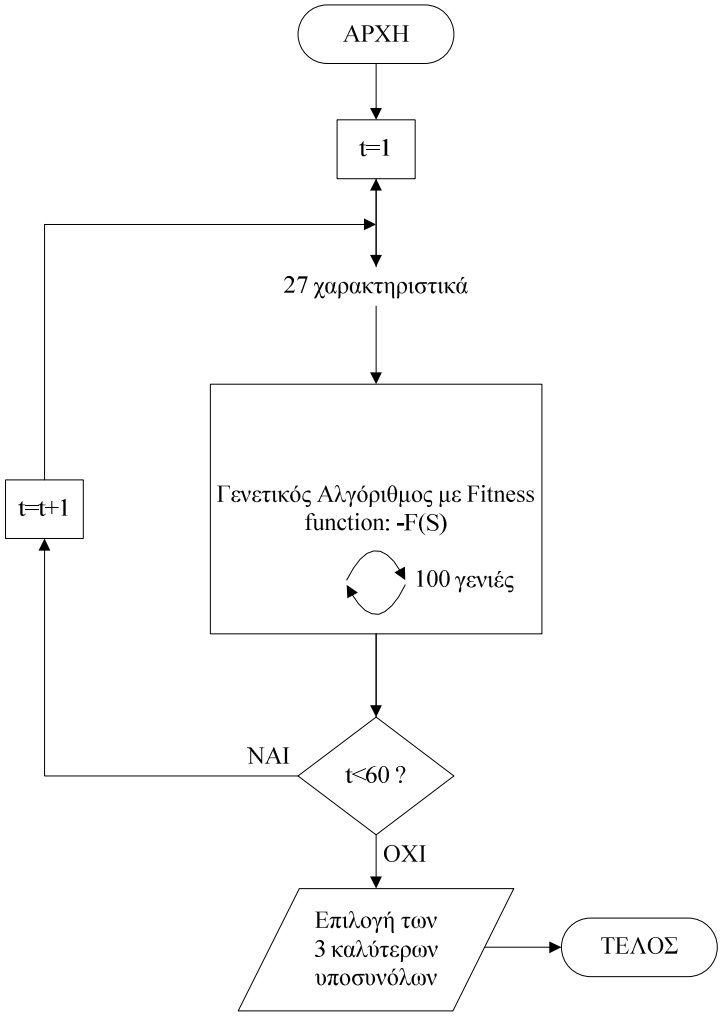
- Μέγεθος Πληθυσμού (Population size) = 20
- Πλήθος γενεών (Number of generations) = 100
- Αρχικοποίηση (Creation) : Τυχαία ομοιόμορφη (Random Uniform)
- Επιλογή γονέων (Selection) : Στοχαστική Ομοιόμορφη (Stochastic Uniform Selection)
- Πλήθος των ελιτιστικών χρωμοσωμάτων (Elitism) = 2
- Διασταύρωση (Crossover) : Διασταύρωση ενός σημείου (Single-point crossover)
- Πιθανότητα διασταύρωσης (Crossover rate) = 0,8
- Μετάλλαξη (Mutation) : Τυχαία ομοιόμορφη (Random Uniform)
- Πιθανότητα μετάλλαξη (Mutation rate) = 0,01
- Κριτήρια Τερματισμού (Stopping Criteria) :
  - Μέγιστο πλήθος γενεών = 100

- Χρονικό όριο (time limit) = +∞
- Όριο στην τιμή της συνάρτησης καταλληλότητας (fitness function limit) = -∞
- Ανοχή στη μεταβολή της τιμής της συνάρτησης καταλληλότητας (fitness function tolerance) = 0
- Πλήθος γενεών με μέση βελτίωση μικρότερη από την ανοχή (Stall generations) = 50
- Χρονικό διάστημα κατά το οποίο δεν υπάρχει βελτίωση (Stall Time limit) = +∞

Θέσαμε την τιμή 0 στην ανοχή στη μεταβολή της συνάρτησης καταλληλότητας, δηλαδή ουσιαστικά απενεργοποιήσαμε το συγκεκριμένο κριτήριο τερματισμού, γιατί μετά από μερικά "τρεξίματα" του αλγόριθμου παρατηρήθηκε το εξής φαινόμενο: όταν ορίζαμε μία τιμή ανοχής, έστω και πολύ μικρή (όπως 10^{-6}), υπήρχαν περιπτώσεις που ο αλγόριθμος τερμάτιζε στις 50 γενεές έχοντας βρει μία ικανοποιητική λύση. Αν όμως, ξανατρέχαμε τον αλγόριθμο με τις ίδιες αρχικές συνθήκες, αλλά με απενεργοποιημένο το κριτήριο ανοχής της μεταβολής, ο αλγόριθμος τερμάτιζε στις 100 γενεές (που είναι ο μέγιστος αριθμός γενεών που έχουμε ορίσει) και έχοντας βρει μία καλύτερη λύση. Μάλιστα, με παρατήρηση των λύσεων που έβρισκε σε κάθε επανάληψη, διαπιστώσαμε ότι η καλύτερη αυτή λύση εμφανίστηκε στη $90^{η}$ γενεά! Αυτό συνέβη γιατί στην πρώτη περίπτωση, οι λύσεις που έβρισκε ο αλγόριθμος δε βελτιώνονταν πάνω από το καθορισμένο όριο ανοχής για αρκετές γενιές, επομένως τερμάτιζε, ενώ στη δεύτερη περίπτωση, "απελευθερωμένος" πια από αυτή τη συνθήκη τερματισμού, του δινόταν η ευκαιρία, μέσω της επίδρασης μίας τυχαίας μετάλλαξης, να "ξεκολλήσει" από τη λύση αυτή και να καταλήξει σε μία ακόμα καλύτερη. Με τον τρόπο αυτό εκμεταλλευόμαστε πλήρως τη φιλοσοφία της "εξέλιξης" του γενετικού αλγορίθμου και την τυχαιότητα που αυτός περιλαμβάνει, για την εύρεση της καλύτερης λύσης.

Όπως έχουμε περιγράψει, ο γενετικός αλγόριθμος αποτελεί μία στοχαστική διαδικασία, δηλαδή σε κάθε τρέξιμό του δεν καταλήγει στα ίδια αποτελέσματα. Ξεκινάει από τυχαίες αρχικές συνθήκες και η εφαρμογή των διάφορων τελεστών του δεν είναι προκαθορισμένη σε κάθε βήμα, αλλά συμβαίνει με τυχαίο τρόπο. Στη μεθοδολογία μας, ο γενετικός αλγόριθμος αποτελεί το πρώτο "προπαρασκευαστικό" στάδιο της επιλογής χαρακτηριστικών, που υλοποιεί την filter προσέγγιση της. Πραγματοποιήθηκαν 60 τρεξίματα του αλγορίθμου και πολλές φορές παρατηρούταν ταύτιση αποτελεσμάτων. Από τα αποτελέσματα αυτά κρατήθηκαν τα 3 καλύτερα υποψήφια υποσύνολα. Με τον όρο "καλύτερα" εννοούμε τα υποσύνολα στα οποία κατέληξε ο γενετικός, με τη μικρότερη τιμή στη συνάρτηση καταλληλότητας. Τα υποσύνολα αυτά αποτελούν την έξοδο αυτού του σταδίου και είσοδο προς το επόμενο στάδιο της μεθόδου.

Το διάγραμμα ροής του γενετικού αλγόριθμου αυτού καθ' αυτού έχει παρατεθεί στο $2^{ο}$ κεφάλαιο της εργασίας. Παρακάτω φαίνεται το διάγραμμα ροής του σταδίου της υλοποίησής μας που μόλις περιγράψαμε:



Σχήμα 4.3: Διάγραμμα ροής filter προσέγγισης επιλογής χαρακτηριστικών.

4.4 Επιλογή χαρακτηριστικών (2^o Στάδιο): Τεχνητά Νευρωνικά Δίκτυα

Το στάδιο αυτό περιλαμβάνει την ολοκλήρωση της επιλογής χαρακτηριστικών και την παράλληλη υλοποίηση της ταξινόμησης. Εφαρμόζεται η wrapper προσέγγιση της επιλογής χαρακτηριστικών και ταυτόχρονα υλοποιείται το ίδιο το μοντέλο εκτίμησης του κινδύνου εμφάνισης αμφιβληστροειδοπάθειας.

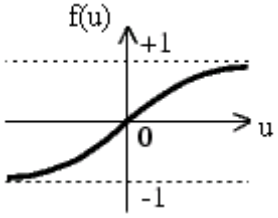
Το υπολογιστικό εργαλείο που χρησιμοποιούμε σε αυτό το στάδιο της μεθοδολογίας μας είναι τα Τεχνητά Νευρωνικά Δίκτυα. Για κάθε ένα από τα υποσύνολα που προέκυψαν από το προηγούμενο βήμα ακολουθούμε μία σχετικά σύνθετη διαδικασία που περιλαμβάνει την εκπαίδευση και την αξιολόγηση ενός συνόλου Νευρωνικών Δικτύων. Έξοδος του σταδίου αυτού είναι το υποσύνολο που μας οδηγεί στο καλύτερο μοντέλο ταξινόμησης, καθώς και το πλήθος των νευρώνων του εσωτερικού επιπέδου και των βαρών του προς υλοποίηση μοντέλου.

Ο λόγος που χρησιμοποιήθηκαν τεχνητά νευρωνικά δίκτυα στη μεθοδολογία μας είναι το γεγονός ότι τα νευρωνικά δίκτυα πρόσθιας τροφοδότησης πολλών στρωμάτων που ενσωματώνουν τη σιγμοειδή συνάρτηση έχουν πολλές δυνατότητες αναπαράστασης συναρτήσεων. Σύμφωνα με βασικό θεώρημα, τα δίκτυα αυτής της μορφής, με μόλις δύο μάλιστα επίπεδα (ένα κρυφό και ένα εξόδου), μπορούν να προσεγγίσουν μία οποιαδήποτε ομαλή συνάρτηση, όσο πιο κοντά επιθυμούμε [Δ07]. Για αυτό το λόγο, καλούνται και "Καθολικοί Προσεγγιστές" ("Universal Approximators"). Η χρήση των νευρωνικών δικτύων αποδεικνύεται ιδιαίτερα κατάλληλη για εργασίες ταξινόμησης, όπως η δική μας, λόγω των πλεονεκτικών χαρακτηριστικών τους (μη γραμμικότητα, ικανότητα μάθησης με παραδείγματα, ανοχή σε σφάλματα και σε θόρυβο, κ.α.).

4.4.1 Υλοποίηση Τεχνητών Νευρωνικών Δικτύων

Στη μεθοδολογία μας χρησιμοποιούμε πολυστρωματικά νευρωνικά δίκτυα πρόσθιας τροφοδότησης (Multi-layer Feedforward Neural Networks). Κάθε νευρωνικό δίκτυο δέχεται σαν εισόδους τις τιμές των χαρακτηριστικών που συγκροτούν το αντίστοιχο υποσύνολο.

Οι αποφάσεις που είχαμε να πάρουμε στο στάδιο αυτό αφορούσαν στην τοπολογία του δικτύων και στη διαδικασία εκπαίδευσης. Κάθε νευρωνικό δίκτυο έχει τόσες εισόδους όσα και τα χαρακτηριστικά του αντίστοιχου υποσυνόλου. Επίσης, όλα έχουν ένα κρυφό επίπεδο και ένα νευρώνα εξόδου. Για κάθε νευρωνικό δίκτυο ακολουθήθηκε ο ίδιος αλγόριθμος εκπαίδευσης, η Ανάστροφη Μετάδοση Λάθους (Back propagation), την οποία έχουμε περιγράψει αναλυτικά στο 2ο κεφάλαιο. Τα βάρη των νευρώνων ανανεώνονται βάσει του μέσου τετραγωνικού σφάλματος (Mean Square Error) μεταξύ της εξόδου του μοντέλου και της πραγματικής εξόδου. Όλοι οι νευρώνες έχουν ίδια συνάρτηση ενεργοποίησης, $f(u) = \frac{2}{(1+e^{-2u})^{-1}}$, η οποία είναι μαθηματικά ισοδύναμη με την tanh(u), με αμελητέες διαφορές στα αριθμητικά αποτελέσματα, και εκτελείται πιο γρήγορα από την tanh(u) στο Matlab, επομένως επιταχύνονται οι υπολογιστικές διαδικασίες και ενδείκνυται σε εργασίες ταξινόμησης. Τα οφέλη των σιγμοειδών συναρτήσεων γενικά έχουν περιγραφεί στο 2ο κεφάλαιο. Η γραφική παράσταση της συνάρτησης ενεργοποίησης έχει την παρακάτω μορφή:



Σχήμα 4.4: Η γραφική παράσταση της συνάρτησης ενεργοποίησης.

Όπως βλέπουμε η έξοδος του δικτύου βρίσκεται στο διάστημα [-1,1].

Πριν την είσοδο των δεδομένων στο δίκτυο εφαρμόζονται σε αυτά δύο συναρτήσεις επεξεργασίας που τα καθιστούν πιο "κατανοητά" για το δίκτυο και διευκολύνουν τη διαδικασία της εκπαίδευσής

του. Η μία συνάρτηση μετασχηματίζει τα δεδομένα εισόδου, έτσι ώστε όλες οι τιμές τους να περιλαμβάνονται στο διάστημα [-1,1] (“mapminmax”). Ο περιορισμός των εισόδων στο διάστημα αυτό είναι κοινή πρακτική που επιταχύνει τη διαδικασία της εκπαίδευσης για τα περισσότερα δίκτυα. Ο λόγος για τον οποίο συχνά κρίνεται αναγκαία η εφαρμογή της είναι ότι οι σιγμοειδείς συναρτήσεις, που χρησιμοποιούνται σαν συναρτήσεις ενεργοποίησης στους νευρώνες, “έρχονται σε κορεσμό” όταν η είσοδος του νευρώνα είναι πάνω από 3 ($e^{-3} \cong 0{,}05$). Όταν συμβαίνει αυτό, τα διαφορικά που υπολογίζονται κατά την Ανάστροφη Μετάδοση Λάθους είναι πολύ μικρά, με αποτέλεσμα η εκπαίδευση να πραγματοποιείται με πολύ αργούς ρυθμούς.

Μία ακόμα συνάρτηση που εφαρμόζουμε αφαιρεί τα χαρακτηριστικά που έχουν πάντα την ίδια τιμή (“removeconstantrows”), καθώς δεν προσφέρουν τίποτα στον υπολογισμό της εξόδου. Στην περίπτωσή μας δεν έχουμε τέτοιου είδους χαρακτηριστικά. Η συμπερίληψη τέτοιων χαρακτηριστικών αποτρέπεται από την εφαρμογή του προηγούμενου σταδίου, αυτού της επιλογής χαρακτηριστικών (ένα χαρακτηριστικό με συνεχώς σταθερές τιμές δεν προσφέρει καμία πληροφορία για την έξοδο και η αμοιβαία πληροφορία αυτού και της εξόδου είναι μηδέν, επομένως, χάρις στο προηγούμενο βήμα, δεν περιλαμβάνεται στο υποψήφιο υποσύνολο). Θεωρήσαμε όμως καλό να συμπεριλάβουμε μία τέτοια συνάρτηση, με στόχο τη γενικότητα του μοντέλου μας, την ενδεχόμενη μελλοντική χρήση του ανεξάρτητα από το προηγούμενο στάδιο και, στην περίπτωση αυτή, την όσο το δυνατό αποτελεσματικότερη αντιμετώπιση και άλλων μελλοντικών συνόλων δεδομένων (επαναχρησιμοποιησιμότητα - reusability). Οι ίδιες συναρτήσεις εφαρμόζονται και στην έξοδο, έτσι ώστε να είναι εφικτή η σύγκριση των εξόδων του δικτύου με τις πραγματικές εξόδους, καθώς και ο υπολογισμός των μέσων τετραγωνικών σφαλμάτων. Η έξοδος του δικτύου δηλαδή μετασχηματίζεται ώστε να ανήκει στο διάστημα [0,1], στο οποίο ανήκουν οι πραγματικές έξοδοι.

Η είσοδος κάθε νευρώνα του κρυφού επιπέδου προκύπτει από την άθροιση των εισόδων πολλαπλασιασμένων με τα αντίστοιχα βάρη που συνδέουν την κάθε είσοδο με το νευρώνα συν την πόλωση (bias) του νευρώνα. Ομοίως προκύπτει και η είσοδος του νευρώνα εξόδου από τις εξόδους των νευρώνων του κρυφού επιπέδου. Η εκπαίδευση περιλαμβάνει 80 εποχές (epochs), 80 δηλαδή παρουσιάσεις των δεδομένων εκπαίδευσης, τιμή που προέκυψε μετά από πειραματισμό και παρατήρηση της απόδοσης του μοντέλου με σκοπό την αποφυγή φαινομένων υπερπροσαρμογής. Τα βάρη και οι πολώσεις ανανεώνονται αφού δοθούν στο δίκτυο όλα τα δεδομένα εκπαίδευσης, στο τέλος δηλαδή της κάθε εποχής.

Η μόνη διαφορά μεταξύ των δικτύων που υλοποιήθηκαν αφορά στην τοπολογία τους και είναι το πλήθος των νευρώνων που συγκροτούν το κάθε ένα. Ο αριθμός των νευρώνων του κρυφού επιπέδου είναι για κάθε δίκτυο αυτός που συνεπάγεται την καλύτερη απόδοση. Ακολουθήθηκε μία επαναληπτική διαδικασία, κατά την οποία ελέγχθηκε η απόδοση του μοντέλου από 8 μέχρι 20 νευρώνες του κρυφού επιπέδου. Επίσης, επειδή η απόδοση είναι στενά εξαρτημένη από τις αρχικές

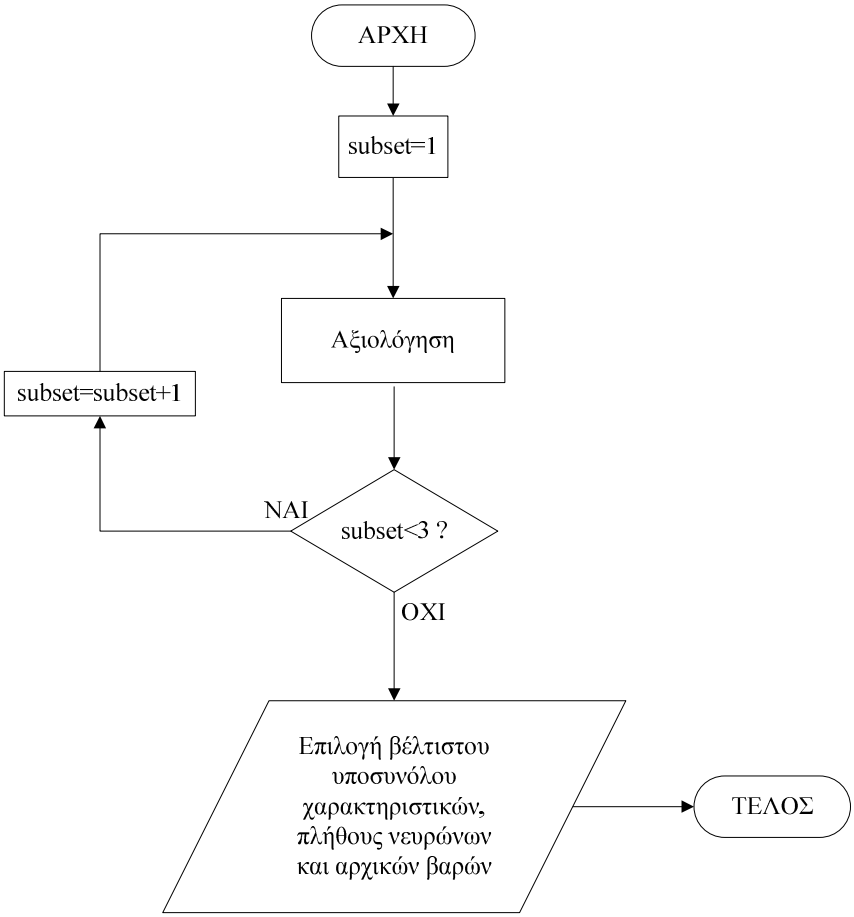
τιμές των βαρών και των πολώσεων του δικτύου, ακολουθήθηκε άλλη μία επαναληπτική διαδικασία για την εύρεση των αρχικών τιμών που μας οδηγούν στο βέλτιστο αποτέλεσμα.

Μετά την εκπαίδευση ακολουθεί η αξιολόγηση του μοντέλου, η οποία στην περίπτωσή μας έχει τριπλό ρόλο. Πρώτον, αποτελεί κομμάτι της ίδιας της μεθοδολογίας μας, καθώς μέσω της αξιολόγησης οδηγούμαστε στη επιλογή του τελικού υποσυνόλου προδιαθεσικών παραγόντων. Μέσω της αξιολόγησης υλοποιείται η wrapper λογική της επιλογής χαρακτηριστικών, σύμφωνα με την οποία "καλύτερο" θα είναι το υποσύνολο βάσει του οποίου έχουμε το ακριβέστερο μοντέλο ταξινόμησης. Επομένως η αξιολόγηση είναι απαραίτητη για την ίδια την επίλυση του προβλήματος. Ο δεύτερος ρόλος της αφορά στην εύρεση των αρχικών τιμών των βαρών και των πολώσεων του τελικού μας μοντέλου. Ουσιαστικά, δηλαδή, μέσω της διαδικασίας της αξιολόγησης βρίσκονται οι βέλτιστες παράμετροι του τελικού μοντέλου πρόβλεψης. Τέλος, ο τρίτος της ρόλος αφορά στην εκτίμηση της ποιότητας και της αξιοπιστίας του μοντέλου ως εργαλείο εκτίμησης κινδύνου εμφάνισης αμφιβληστροειδοπάθειας και στο γενικότερο έλεγχο του κατά πόσο επιτεύχθηκε ο στόχος της διπλωματικής και επιλύθηκε το δεδομένο πρόβλημα.

Αν και η αξιολόγηση αποτελεί μέρος της μεθοδολογίας μας και απαραίτητο στάδιο για την ολοκλήρωση του στόχου μας, δε θα συνεχίσουμε με την περιγραφή της στο σημείο αυτό. Θεωρήσαμε σωστό να την παρουσιάσουμε αναλυτικά, μαζί με την πρακτική που ακολουθήθηκε για την τελική απόφανση της αξιολόγησης (κριτήρια αξιολόγησης, κλπ.) καθώς και τα ίδια τα αποτελέσματα της, στο επόμενο κεφάλαιο που αφορά εξ ολοκλήρου στην αξιολόγηση των τεχνικών μας. Προς το παρόν θα την αντιμετωπίσουμε σαν ένα μαύρο κουτί.

Η διαδικασία της αξιολόγησης εφαρμόζεται σε όλα νευρωνικά δίκτυα. Μετά την ολοκλήρωση της αξιολόγησης διαθέτουμε διάφορα μέτρα εκτίμησης της απόδοσης του κάθε νευρωνικού. Επιλέγουμε, τέλος, το υποσύνολο χαρακτηριστικών για το οποίο έχουμε την καλύτερη απόδοση και τις αρχικές τιμές των βαρών και των πολώσεων για τις οποίες αυτή προέκυψε.

Το διάγραμμα ροής της υλοποίησης ενός νευρωνικού δικτύου έχει παραταθεί στο $2^{ο}$ κεφάλαιο. Ακολουθεί το διάγραμμα ροής της διαδικασίας που περιγράφηκε:



Σχήμα 4.5: Διάγραμμα ροής wrapper προσέγγισης επιλογής χαρακτηριστικών και υλοποίησης μοντέλου ταξινόμησης.

4.5 Υλοποίηση τελικού μοντέλου πρόβλεψης

Το τελικό μας μοντέλο εκτίμησης κινδύνου εμφάνισης αμφιβληστροειδοπάθειας έχει τον αριθμό των νευρώνων κρυφού επιπέδου και τις αρχικές τιμές βαρών και πολώσεων που προέκυψαν από το προηγούμενο βήμα και δέχεται ως εισόδους τις τιμές των χαρακτηριστικών του υποσυνόλου που επιλέχθηκε βάσει της διαδικασίας της αξιολόγησης. Εκπαιδεύεται με το 70% των δεδομένων μας και ελέγχεται με το 30%, για επαλήθευση της τελικής απόδοσής του. Εφόσον τα δεδομένα που διαθέτουμε περιλαμβάνουν μία χρονική περίοδο 6 ετών, το προς υλοποίηση μοντέλο πρέπει να μπορεί να προβλέψει την εξέλιξη ενός ατόμου, όσον αφορά στην αμφιβληστροειδοπάθεια, σε βάθος χρόνου 6 ετών. Με άλλα λόγια, αν διαθέτουμε τις μετρήσεις ενός διαβητικού ατόμου στα χαρακτηριστικά που χρησιμοποιούνται ως είσοδοι στο μοντέλο, εφαρμόζοντας 6 φορές τις μετρήσεις που διαθέτουμε και αυξάνοντας κάθε φορά τη μεταβλητή που αντιπροσωπεύει το έτος, το μοντέλο πρέπει να είναι σε θέση να εκτιμήσει την πιθανότητα εμφάνισης αμφιβληστροειδοπάθειας για κάθε έτος.

Η πιθανότητα το άτομο να εμφανίσει ή να μην εμφανίσει αμφιβληστροειδοπάθεια, προκύπτει από τα αποτελέσματα του ίδιου του μοντέλου ταξινόμησης. Αν, για παράδειγμα, η έξοδος του μοντέλου για ένα διάνυσμα εισόδων ενός ατόμου είναι 0,7, το άτομο έχει 70% πιθανότητα να εμφανίσει αμφιβληστροειδοπάθεια.

5

Αξιολόγηση

Στο κεφάλαιο αυτό παρουσιάζεται αναλυτικά η διαδικασία της αξιολόγησης των τεχνικών μας και τα αποτελέσματά που προέκυψαν.

5.1 Κριτήρια αξιολόγησης

Η αξία και η ποιότητα μίας διαγνωστικής δοκιμασίας (diagnostic test) μίας ασθένειας προσδιορίζεται κατά κύριο λόγο από τη διακριτική του ικανότητα, δηλαδή από την ικανότητά του να διακρίνει τα ασθενή άτομα από τα υγιή. Ένα μοντέλο εκτίμησης του κινδύνου εμφάνισης μίας ασθένειας αποτελεί ουσιαστικά μία "μακροπρόθεσμη" διαγνωστική δοκιμασία, η ποιότητα της οποίας αναφέρεται στην ικανότητά του να αποτιμά υψηλή τιμή πιθανότητας εμφάνισης της ασθένειας σε άτομα που όντως διατρέχουν υψηλότερο κίνδυνο να την εμφανίσουν σε σχέση με άλλα. Η αξιολόγησή του αντιμετωπίζεται συνήθως με τους ίδιους όρους με αυτούς μίας διαγνωστικής δοκιμασίας. Επιπλέον, οι όροι αξιολόγησης μίας διαγνωστικής δοκιμασίας έχουν πρόσφατα εισαχθεί στην αξιολόγηση πολλών εφαρμογών της μηχανική μάθησης γενικότερα. Για τους λόγους αυτούς, η πρακτική αυτή θα ακολουθηθεί και στην παρούσα διπλωματική.

Οι συχνότερα χρησιμοποιούμενες παράμετροι για την αξιολόγηση της διαγνωστικής ποιότητας είναι το Ποσοστό των Αληθώς Θετικών Αποτελεσμάτων (True Positive Rate) και το Ποσοστό των Αληθώς Αρνητικών Αποτελεσμάτων (True Negative Rate). Τα ποσοστά αυτά, καθώς και τα συμπληρωματικά τους (Ποσοστό των Ψευδώς Αρνητικών Αποτελεσμάτων και Ποσοστό των Ψευδώς Θετικών Αποτελεσμάτων αντίστοιχα), ονομάζονται λειτουργικά χαρακτηριστικά της

διαγνωστικής δοκιμασίας (Operating Characteristics). Οι παράμετροι αυτοί εξηγούνται στη συνέχεια [ΠΣ04].

Τα δύο ακραία αποτελέσματα του μοντέλου μας είναι η τιμή "1" που ονομάζεται θετικό αποτέλεσμα όσον αφορά στην υπόθεση που εξετάζεται, που στην περίπτωσή μας είναι η εμφάνιση της αμφιβληστροειδοπάθειας και αντιστοιχεί σε παθολογική κατάσταση και η τιμή "0" που ονομάζεται αρνητικό αποτέλεσμα και αντιστοιχεί σε φυσιολογική κατάσταση. Όσον αφορά στα αποτελέσματα της πρόβλεψης στα δεδομένα ελέγχου, δηλαδή στα δεδομένα για τα οποία ξέρουμε εκ των προτέρων την έξοδο και τα εισάγουμε σαν είσοδο στο ήδη εκπαιδευμένο μοντέλο μας, υπάρχουν τέσσερα πιθανά σενάρια:

Πρώτον, η πρόβλεψη να είναι θετική όσον αφορά στην εμφάνιση της αμφιβληστροειδοπάθειας και η τιμή της εξόδου να είναι πράγματι θετική. Το αποτέλεσμα αυτό ονομάζεται Αληθώς Θετικό (True Positive – TP). Αντίστοιχα, όταν και η πρόβλεψη και η πραγματική έξοδος είναι αρνητικές (δεν εμφάνισε αμφιβληστροειδοπάθεια), έχουμε Αληθώς Αρνητικό αποτέλεσμα (True Negative – TN). Όταν όμως η πραγματική τιμή της εξόδου είναι αρνητική και η πρόβλεψη είναι θετική, όταν δηλαδή το μοντέλο λανθασμένα προβλέπει την εμφάνιση της αμφιβληστροειδοπάθειας, έχουμε Ψευδώς Θετικό αποτέλεσμα (False Positive – FP). Στη στατιστική το λάθος αυτό ονομάζεται σφάλμα Τύπου I. Ονομάζεται και "ψευδές σήμα συναγερμού" (false alarm), καθώς αποδίδει κίνδυνο όταν αυτός δεν υπάρχει. Όταν πάλι η πραγματική τιμή της εξόδου είναι θετική και η πρόβλεψη είναι αρνητική, όταν δηλαδή το μοντέλο λανθασμένα δεν προβλέπει την εμφάνιση της αμφιβληστροειδοπάθειας, έχουμε ψευδώς αρνητικό αποτέλεσμα (False Negative – FN), ή αλλιώς σφάλμα Τύπου II. Γενικά το σφάλμα Τύπου II θεωρείται χειρότερο από το Τύπου I. Ειδικά στην ιατρική διάγνωση και στην εκτίμηση του κινδύνου εμφάνισης μίας ασθένειας, όπως καταλαβαίνουμε και με τη λογική, είναι χειρότερο μία διαγνωστική δοκιμασία ή ένα σύστημα πρόβλεψης να μην αντιλαμβάνεται καν μία ασθένεια ενώ το άτομο πάσχει από αυτή, από το να υποδεικνύει την ύπαρξή της ενώ το άτομο είναι υγιές. Για παράδειγμα, σε μία διαγνωστική δοκιμασία για τον ιό HIV είναι πολύ πιο επιζήμια η αρνητική ένδειξη όταν το άτομο έχει τον ιό, από ότι η θετική ένδειξη όταν το άτομο δεν τον έχει. Στην πρώτη περίπτωση λειτουργεί καθησυχαστικά, ενώ δε θα έπρεπε, και αποτρέπει την περαιτέρω εξέταση του ζητήματος, κάτι που μπορεί να είναι πολύ επικίνδυνο για την υγεία του ατόμου, ενώ στη δεύτερη περίπτωση επιφέρει μεν έναν "πανικό" χωρίς λόγο (αποτελεί δηλαδή ένα "ψευδές σήμα συναγερμού"), που όμως μπορεί να εξαλειφθεί με περαιτέρω εξέταση, χωρίς να θέτει το άτομο σε κίνδυνο. Σίγουρα, δεν πρέπει να παραβλέπουμε την επίδραση του ψυχολογικού παράγοντα σε τέτοιες περιπτώσεις, αλλά κατά γενική ομολογία, τα σφάλματα Τύπου II είναι περισσότερο επικίνδυνα για το άτομο.

Η Ευαισθησία ή αλλιώς το Ποσοστό των Αληθώς Θετικών Αποτελεσμάτων (Sensitivity ή True Positive Rate – TPR) του μοντέλου είναι ο λόγος των αληθώς θετικών αποτελεσμάτων προς το

συνολικό πλήθος των πραγματικά θετικών αποτελεσμάτων, δηλαδή, το ποσοστό των ασθενών που έχουν στην πραγματικότητα θετική έξοδο και ταξινομήθηκαν σωστά. Δίνεται από τη σχέση:

TPR = TP / P = TP / (TP + FN).

Η Ειδικότητα (Specificity – SPC ή True Negative Rate – TNR) ή το Ποσοστό των Αληθώς Αρνητικών Αποτελεσμάτων του μοντέλου είναι η αναλογία των αληθώς αρνητικών αποτελεσμάτων προς τα πραγματικά αρνητικά, δηλαδή το ποσοστό των ασθενών που έχουν στην πραγματικότητα αρνητική έξοδο και ταξινομήθηκαν σωστά. Δίνεται από τη σχέση:

SPC = TN / N = TN / (FP + TN).

Ο όρος "Ευαισθησία" χρησιμοποιείται επίσης για να ορίσει τη μικρότερη συγκέντρωση μίας ουσίας που μπορεί μία μέθοδος να ανιχνεύσει και να μετρήσει (Αναλυτική Ευαισθησία). Αντίστοιχα, ο όρος "Ειδικότητα" χρησιμοποιείται για να ορίσει την ιδιότητα μίας μεθόδου να μετράει μόνο την ουσία που επιδιώκεται να μετρηθεί, ανάμεσα σε άλλες. Σε αυτό ερευνητικό πεδίο, όμως, η Ευαισθησία και η Ειδικότητα αναφέρονται στο αποτέλεσμα της δοκιμασίας και όχι στην ίδια τη δοκιμασία.

Άλλο μέτρο αξιολόγησης είναι η Προγνωστική Αξία των Θετικών αποτελεσμάτων (Positive Predictive Value – PPV ή Precision), δηλαδή το ποσοστό των θετικών αποτελεσμάτων του μοντέλου που ήταν πράγματι θετικά. Δίνεται από τη σχέση: PPV = TP / (TP + FP).

Ουσιαστικά εκφράζει την ακρίβεια με την οποία ένα θετικό αποτέλεσμα συνεπάγεται την εμφάνιση της επιπλοκής.

Αντίστοιχα ορίζεται η Προγνωστική Αξία των Αρνητικών αποτελεσμάτων (Negative Predictive Value – NPV), η οποία δίνεται από τη σχέση: NPV = TN / (TN + FN) και εκφράζει την ακρίβεια με την οποία ένα αρνητικό αποτέλεσμα προβλέπει την απουσία της επιπλοκής.

Τα παραπάνω μεγέθη φαίνονται σχηματικά παρακάτω:

		Πραγματικό Αποτέλεσμα		
		Ασθενής	Υγιής	
Αποτέλεσμα Μοντέλου	Ασθενής	**True Positive**	**False Positive** (Type I error)	→ Positive predictive value = TP / (TP + FP)
	Υγιής	**False Negative** (Type II error)	**True Negative**	→ Negative predictive value = TN / (FN + TN)
		↓ Sensitivity = TP / (TP + FN)	↓ Specificity = TN / (FP + TN)	

Σχήμα 5.1: Λειτουργικά Χαρακτηριστικά.

Ο πίνακας των λειτουργικών χαρακτηριστικών ονομάζεται και Μήτρα Σύγχυσης (Confusion Matrix).

Ορίζεται, επίσης, το συμπληρωματικό μέγεθος της Ευαισθησίας, το Ποσοστό των Ψευδώς Αρνητικών Αποτελεσμάτων (False Negative Rate – FNR):

FNR = FN / P = FN / (TP + FN),

και το Ποσοστό των Ψευδώς Θετικών Αποτελεσμάτων (False Positive Rate – FPR):

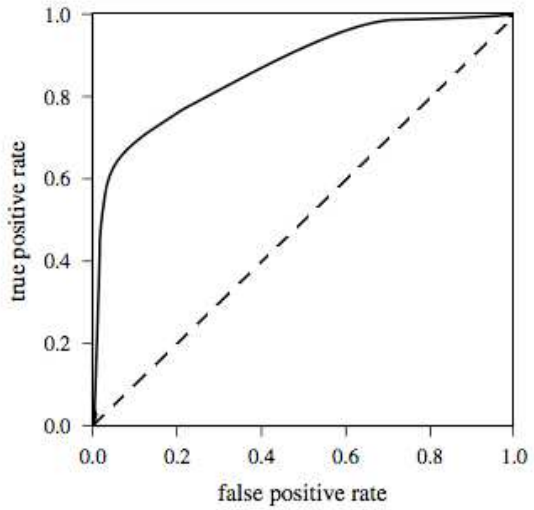
FP / N = FP / (FP + TN).

Η Ακρίβεια (Accuracy) ορίζεται ως ο λόγος των σωστά ταξινομημένων αποτελεσμάτων προς τα συνολικά. Δίνεται από τη σχέση: ACC = (TP + TN) / (P+N).

Κάθε διαγνωστική δοκιμασία έχει δύο χαρακτηριστικά: τη διαχωρίζουσα μεταβλητή (separator variable) και το διαχωριστικό όριο ή σημείο απόφασης (cut-off point ή dividing line). Η διαχωρίζουσα μεταβλητή είναι μία "μετρήσιμη" ιδιότητα που σχετίζεται με μία συγκεκριμένη ασθένεια. Το διαχωριστικό όριο είναι μία συγκεκριμένη τιμή της διαχωρίζουσας μεταβλητής, πέραν της οποίας οι τιμές της μεταβλητής, ή μάλλον τα αποτελέσματα της δοκιμασίας χαρακτηρίζονται ως "θετικά", δηλαδή "παθολογικά" και κάτω της οποίας "αρνητικά", δηλαδή "φυσιολογικά". Είναι δηλαδή η τιμή της μεταβλητής που διακρίνει τα άτομα σε ασθενή και υγιή στο υπό διερεύνηση νόσημα. Η επιλογή του διαχωριστικού είναι μεγάλης σημασίας, καθώς επηρεάζει τη διακριτική ικανότητα της δοκιμασίας. Οι βιολογικές μεταβλητές εμφανίζουν συνήθως μεγάλη διασπορά των τιμών σε ασθενείς και μη πληθυσμούς. Επίσης, σε πολλές περιπτώσεις παρατηρείται επικάλυψη των κατανομών των τιμών στις δύο αυτές ομάδες του πληθυσμού. Για το λόγο αυτό, είναι αδύνατος ο καθορισμός ενός διαχωριστικού ορίου που να ξεχωρίζει πλήρως και απόλυτα όλους τους ασθενείς από τους μη, συνεπώς η τέλεια διακριτική ικανότητα είναι ανέφικτος στόχος. Στόχος είναι η επιλογή του όσο το δυνατό καλύτερου διαχωριστικού ορίου, κάτι που προϋποθέτει τη γνώση του τρόπου με το οποίο αυτό επιδρά στη απόδοση της δοκιμασίας ή του μοντέλου.

Με αυτούς τους όρους, μία μέθοδος αξιολόγησης μίας διαγνωστικής δοκιμασίας ή ενός μοντέλου πρόβλεψης είναι η χάραξη καμπύλων λειτουργικών χαρακτηριστικών (Receiver Operating Characteristics curves - ROC curves). Η χρήση διαφορετικών τιμών διαχωριστικού ορίου, πάνω από τις οποίες το αποτέλεσμα χαρακτηρίζεται ως θετικό, έχει σαν αποτέλεσμα την ταυτόχρονη μεταβολή του Ποσοστού των Αληθώς Θετικών, Ψευδών Θετικών, Αληθώς Αρνητικών και Ψευδώς Αρνητικών αποτελεσμάτων της δοκιμασίας. Η ROC καμπύλη είναι το συνεχές γράφημα που ορίζουν τα ζεύγη σημείων (Ποσοστό των Ψευδώς Θετικών, Ποσοστό Αληθώς Θετικών) για όλες τις δυνατές τιμές του διαχωριστικού σημείου. Περιλαμβάνονται και οι δύο ακραίες περιπτώσεις, να είναι όλες οι τιμές της διαχωρίζουσας μεταβλητής ενδεικτικές ασθένειας και να μην είναι καμία τιμή ενδεικτική ασθένειας. Η ROC καμπύλη εκφράζει δηλαδή τη σχέση του Ποσοστού των Αληθώς Θετικών και Ψευδώς

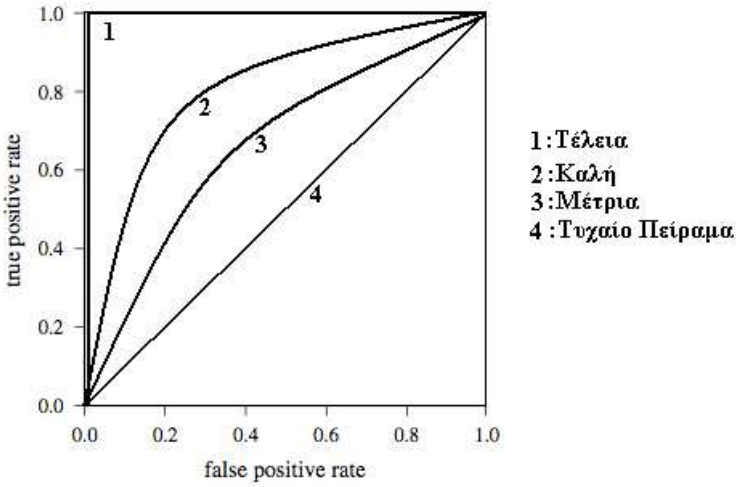
Θετικών, ή αλλιώς τη διακύμανση της ποιότητας της δοκιμασίας, καθώς μεταβάλλεται σταδιακά το διαχωριστικό όριο. Η ROC καμπύλη έχει την παρακάτω μορφή:



Σχήμα 5.2: Τυπική μορφή ROC καμπύλης.

Το σημείο της καμπύλης ROC που βρίσκεται πλησιέστερα στην πάνω αριστερή γωνία του τετραγώνου (όπου το Ποσοστό των Αληθών Θετικών είναι ίσο με 1 (100% Ευαισθησία) και το Ποσοστό των Ψευδώς Θετικών είναι ίσο με 0 (100% Ειδικότητα)) αντιστοιχεί στο βέλτιστο διαχωριστικό σημείο, στο σημείο δηλαδή με τη βέλτιστη αναλογία αληθώς θετικών και ψευδώς θετικών αποτελεσμάτων. Το σημείο (0,1) ονομάζεται και "τέλεια ταξινόμηση". Μία εντελώς τυχαία εικασία (random guess) για την έξοδο, για παράδειγμα μία απόφαση που να βασίζεται στο αποτέλεσμα της ρίψης ενός νομίσματος, θα έδινε ένα σημείο πάνω στη διαγώνια γραμμή. Σημεία πάνω από τη διαγώνιο αντιπροσωπεύουν καλά αποτελέσματα, ενώ όσο πιο κοντά στη διαγώνιο βρισκόμαστε τόσο χειρότερη είναι η απόδοση της δοκιμασίας. Σημεία κάτω από τη διαγώνιο αντιστοιχούν σε μοντέλο με χαμηλή απόδοση, του οποίου τα αποτελέσματα πρέπει να αντιστραφούν για να επιτευχθεί ένα αξιοπρεπές μοντέλο.

Εκτός από την επιλογή του βέλτιστου διαχωριστικού σημείου, με τη βοήθεια της καμπύλης ROC παρέχεται η δυνατότητα οπτικής και ποσοτικής εκτίμησης της ποιότητας και της αξιοπιστίας μίας διαγνωστικής δοκιμασίας ή ενός μοντέλου ανεξάρτητα από το διαχωριστικό όριο. Έτσι, είναι δυνατή η εκτίμηση της διακριτικής ικανότητας μίας διαγνωστικής δοκιμασίας ή μοντέλου, αλλά και η σύγκριση της διακριτικής ικανότητας δύο ή περισσότερων δοκιμασιών ή μοντέλων. Στο παρακάτω σχήμα φαίνεται μία συγκριτική αξιολόγηση μερικών δοκιμασιών βάσει των καμπυλών ROC:



Σχήμα 5.3: Συγκριτική Αξιολόγηση δοκιμασιών βάσει καμπυλών ROC.

Οι παράμετροι αξιολόγησης που αναλύθηκαν παραπάνω θα χρησιμοποιηθούν και στην παρούσα διπλωματική λόγω της ευρείας χρήσης τους στους τομείς της ιατρικής διάγνωσης, αλλά και της μηχανικής μάθησης, και λόγω της δυνατότητας που παρέχουν για μία αξιόπιστη εκτίμηση της απόδοσης.

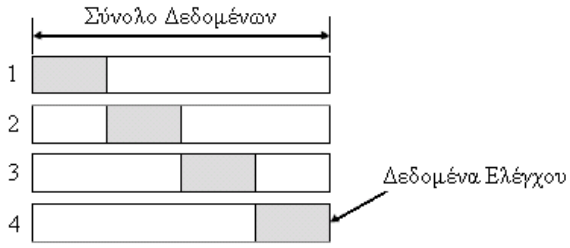
5.2 Σύστημα αξιολόγησης

Στην ενότητα αυτή περιγράφεται η διαδικασία που ακολουθήθηκε για την αξιολόγηση της υλοποίησής μας.

Η υψηλή απόδοση ενός ταξινομητή στα δεδομένα εκπαίδευσης δεν σημαίνει κάτι από μόνη της, καθώς ενδέχεται ο ταξινομητής να έχει "μάθει" πολύ καλά τα δεδομένα εκπαίδευσης, αλλά να αποτυγχάνει να ταξινομήσει σωστά δεδομένα που δεν έχει δει ακόμα. Επομένως, η αξιολόγηση μίας εργασίας ταξινόμησης περιλαμβάνει σίγουρα την εφαρμογή στο σύστημα ενός άλλου συνόλου δεδομένων, των δεδομένων ελέγχου. Μία ευρέως γνωστή και χρησιμοποιημένη στατιστική μέθοδος για την εκτίμηση της απόδοσης μίας εφαρμογής μηχανικής μάθησης, όπως είναι η ταξινόμηση, είναι η μέθοδος της Διασταυρωμένης Επικύρωσης (Cross-validation) [RTL08]. Το πρόβλημα με την αξιολόγηση ενός ταξινομητή είναι ότι η απόδοση του μπορεί να είναι διαφορετική ανάλογα με το ποια δεδομένα χρησιμοποιούνται για δεδομένα εκπαίδευσης. Ο στόχος της Διασταυρωμένης Επικύρωσης είναι να αποδεσμεύσει την απόδοση από αυτόν τον παράγοντα και να εκτιμήσει την ικανότητα γενίκευσης του ταξινομητή.

Υπάρχουν διάφορες τεχνικές Διασταυρωμένης Επικύρωσης. Η πιο βασική και ευρέως αποδεκτή μορφή της είναι η Διασταυρωμένη Επικύρωση σε k μέρη (k-fold Cross-Validation). Σύμφωνα με αυτή, το σύνολο των δεδομένων χωρίζεται σε k ίσα, ή σχεδόν ίσα μέρη. Πραγματοποιούνται k επαναλήψεις εκπαίδευσης και αξιολόγησης. Σε κάθε επανάληψη, 1 μέρος του συνόλου, διαφορετικό κάθε φορά, χρησιμοποιείται για την αξιολόγηση (δεδομένα ελέγχου) και τα υπόλοιπα k-1 μέρη για

την εκπαίδευση του ταξινομητή (δεδομένα εκπαίδευσης). Σε κάθε επανάληψη ο ταξινομητής αξιολογείται βάσει των δεδομένων ελέγχου, με χρήση διάφορων μέτρων αξιολόγησης, όπως η ακρίβεια ή το μέσο τετραγωνικό σφάλμα. Μετά την ολοκλήρωση της παραπάνω διαδικασίας έχουμε k μέτρα αξιολόγησης. Χρησιμοποιείται, τέλος, κάποιο στατιστικό μέτρο (όπως ο μέσος όρος) με το οποίο συνοψίζονται τα k μέτρα και προκύπτει η τελική εκτίμηση της απόδοσης του μοντέλου. Πρέπει να μην υπάρχει επικάλυψη μεταξύ των διαμερίσεων, να μην υπάρχουν δηλαδή κοινά στιγμιότυπα, καθώς επίσης κάθε διαμέριση να είναι αντιπροσωπευτική του συνόλου, να έχει δηλαδή ίδια αναλογία "1" και "0" στην έξοδο με το σύνολο. Παρακάτω φαίνεται σχηματικά η ιδέα της 4-fold Cross-validation. Το σκιασμένο μέρος είναι αυτό που χρησιμοποιείται κάθε φορά για τη αξιολόγηση του μοντέλου, ενώ όλα τα υπόλοιπα χρησιμοποιούνται για την εκπαίδευση.



Σχήμα 5.4: 4-fold Cross-validation.

Ένα σημαντικό ζήτημα είναι η επιλογή της κατάλληλης τιμής του k. Γενικά, η χρήση μεγάλης τιμής για το k θεωρείται καλή πρακτική, καθώς με μεγαλύτερο k η απόδοση εκτιμάται περισσότερες φορές και το σύνολο δεδομένων είναι πιο κοντά στο πλήρες σύνολο δεδομένων, αυξάνοντας έτσι την πιθανότητα κάθε συμπέρασμα σχετικό με την απόδοση του ταξινομητή να μην απέχει πολύ από την περίπτωση που χρησιμοποιείται όλο το σύνολο των δεδομένων για εκπαίδευση, δηλαδή από το τελικό μοντέλο. Όσο όμως αυξάνει το k, τόσο αυξάνει και η επικάλυψη μεταξύ των δεδομένων εκπαίδευσης σε κάθε επανάληψη. Επίσης, μειώνεται το σύνολο των δεδομένων ελέγχου, οδηγώντας σε λιγότερο ακριβείς και αντιπροσωπευτικές εκτιμήσεις της απόδοσης σε νέα δεδομένα. Οι παράγοντες αυτοί έχουν ληφθεί υπόψη και επικρατεί η άποψη ότι μία καλή τιμή είναι k=10, καθώς λειτουργεί με το 90% των δεδομένων δίνοντας υψηλές πιθανότητες σωστής γενίκευσης του συμπεράσματος σε ολόκληρο το σύνολο δεδομένων. Έτσι, η Διασταυρωμένη Επικύρωση σε 10 μέρη (10-fold Cross-Validation) θεωρείται από πολλούς ερευνητές η καλύτερη εκδοχή αυτής της μεθόδου [Koh95].

5.3 Οργάνωση διαδικασίας αξιολόγησης

5.3.1 Αξιολόγηση Συνάρτησης καταλληλότητας του Γενετικού Αλγορίθμου

Αρχικά ελέγξαμε την αξιοπιστία της συνάρτησης καταλληλότητας του γενετικού αλγόριθμου με τη χρήση ενός άλλου συνόλου δεδομένων. Το σύνολο δεδομένων που χρησιμοποιήσαμε είναι το

Wisconsin Breast Cancer Database, το οποίο αποκτήθηκε από το University of Wisconsin Hospitals, Madison από τον Dr. William H. Wolberg και είναι διαθέσιμο στο UCI Machine Learning Repository στην ηλεκτρονική διεύθυνση: http://archive.ics.uci.edu/ml/. Στην "αποθήκη" αυτή περιλαμβάνονται πολλά τεχνητά αλλά και αληθινά υποσύνολα δεδομένων που χρησιμοποιούνται ευρέως από την κοινότητα μηχανικής μάθησης, σε διάφορες εργασίες του χώρου.

Το Wisconsin Breast Cancer σύνολο δεδομένων περιλαμβάνει 699 δείγματα βιοψίας (στιγμιότυπα). Η ταξινόμηση των ατόμων αυτών αφορά στο χαρακτηρισμό του όγκου ως κακοήθη ή καλοήθη. Τα 458 δείγματα είναι καλοήθη και τα 241 είναι κακοήθη. Κάθε δείγμα περιγράφεται από 9 χαρακτηριστικά: (1) Πάχος Όγκου, (2) Ομοιομορφία στο μέγεθος των κυττάρων, (3) Ομοιομορφία στο σχήμα των κυττάρων, (4) Οριακή μεσοθαλάμια σύνδεση (Marginal Adhesion), (5) Μέγεθος Επιθηλιακών κυττάρων, (6) Παρουσία γυμνών πυρήνων, (7) Δίκτυο Χρωματίνης, (8) Φυσιολογικοί πυρηνίσκοι, (9) Μίτωση. Όλα τα χαρακτηριστικά είναι διακριτά και παίρνουν τιμές από 1 έως 10.

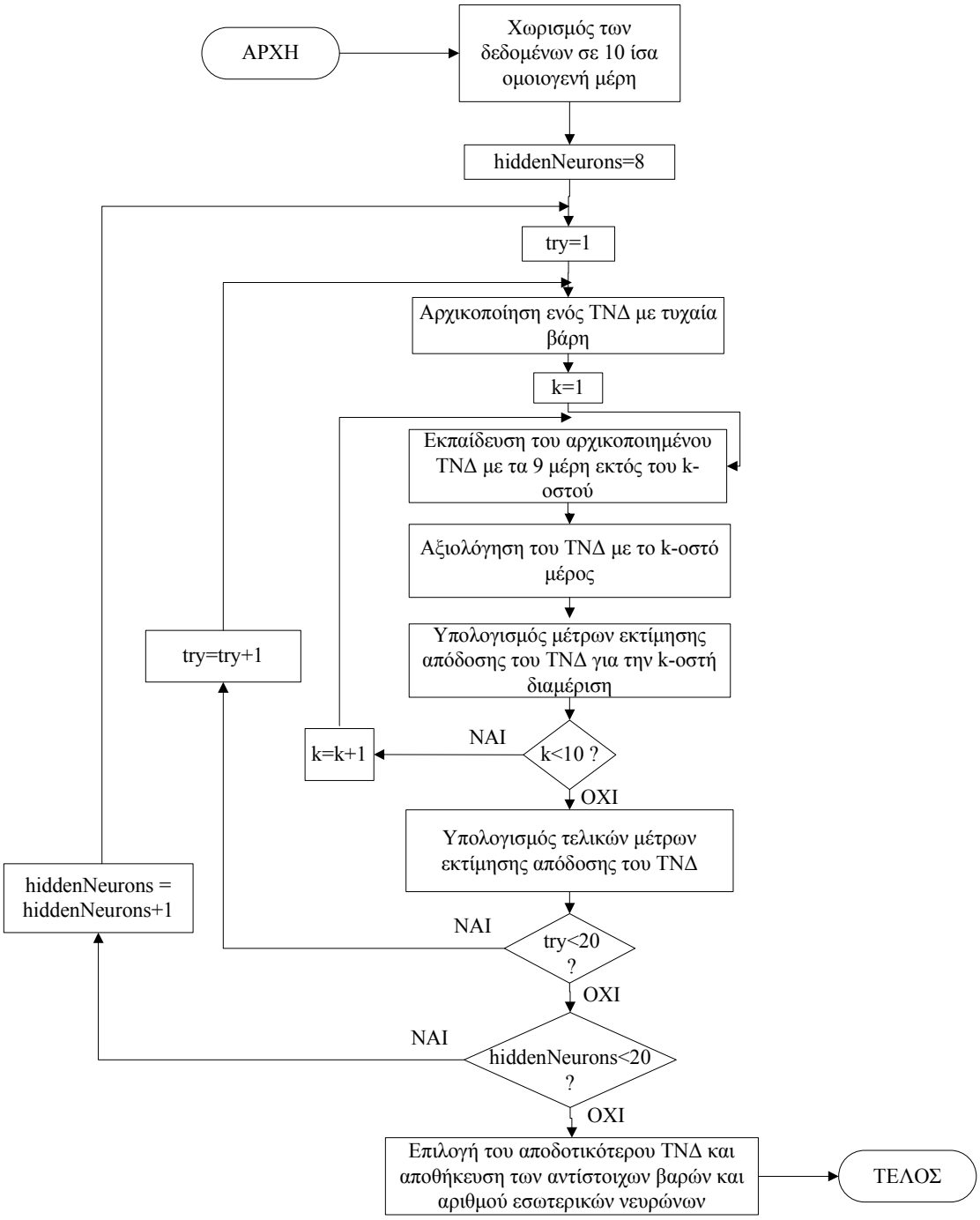
Συγκρίναμε τα αποτελέσματα της συνάρτησης καταλληλότητας με αυτά της εργασίας [VB02]. Στην εργασία αυτή προτείνεται μία μέθοδος επιλογής χαρακτηριστικών με χρήση νευρωνικών δικτύων που βασίζεται στο σφάλμα ταξινόμησης (wrapper approach). Ένα από τα σύνολα δεδομένων που χρησιμοποιεί είναι το Wisconsin Breast Cancer. Η σύγκριση των αποτελεσμάτων μας με αυτά της εργασίας, στην οποία τα χαρακτηριστικά επιλέγονται βάσει της απόδοσης της ταξινόμησης, παρουσιάζει μεγάλο ενδιαφέρον, καθώς αποτελεί ένα μέτρο της γενικότητας της συνάρτησης καταλληλότητας μας. Επίσης, αποτελεί μία ένδειξη του κατά πόσο αυτή πλησιάζει στην επίτευξη του στόχου της επιλογής χαρακτηριστικών, που είναι η ταξινόμηση.

5.3.2 Αξιολόγηση υποσυνόλων και Εύρεση βέλτιστων Παραμέτρων

Όσον αφορά στην αξιολόγηση των υποψήφιων υποσυνόλων του Γενετικού Αλγόριθμου και την εύρεση των βέλτιστων παραμέτρων του μοντέλου εκτίμησης κινδύνου, ακολουθήσαμε τη μέθοδο της Διασταυρωμένης Επικύρωσης σε 10 μέρη 10-fold Cross-validation (k=10). Το σύνολο των δεδομένων που διαθέτουμε αποτελείται από 1038 στιγμιότυπα, με 780 "0" και 258 "1" και χωρίστηκε σε 10 περίπου ίσα ομοιογενή μέρη. Κάθε μέρος περιλαμβάνει περίπου 104 στιγμιότυπα με 78 "0" και 26 "1" (υπάρχουν και δύο μέρη που περιλαμβάνουν 103 στιγμιότυπα με αναλογία "0"/"1" 78/25). Για κάθε υποψήφιο υποσύνολο ακολουθήθηκε η εξής διαδικασία:

Μεταβάλλουμε τον αριθμό των νευρώνων του εσωτερικού επιπέδου (hiddenNeurons) από 8 έως 20. Για κάθε πλήθος νευρώνων υλοποιούμε 20 δοκιμές (tries). Σε κάθε δοκιμή αρχικοποιούνται τυχαία τα βάρη και οι πολώσεις του νευρωνικού δικτύου και πραγματοποιείται ένας πλήρης κύκλος Διασταυρωμένης Επικύρωσης. Δηλαδή, σε κάθε δοκιμή εκπαιδεύονται 10 νευρωνικά δίκτυα ίδιας δομής και ίδιων παραμέτρων, ένα για το κάθε μέρος (k) της Διασταυρωμένης Επικύρωσης. Για κάθε μέρος της υπολογίζονται 6 κριτήρια αξιολόγησης. Τα κριτήρια που υπολογίζονται είναι το Ποσοστό των Αληθώς Θετικών (TPR), το Ποσοστό των Ψευδώς Αρνητικών (FNR), το Ποσοστό των Αληθώς

Αρνητικών (TNR), το Ποσοστό των Ψευδώς Θετικών (FPR), η Προγνωστική Αξία των Θετικών αποτελεσμάτων (PPV) και η Προγνωστική Αξία των Αρνητικών αποτελεσμάτων (NPV). Ο μέσος όρος της κάθε 10άδας κριτηρίων δίνει την τελική 6άδα παραμέτρων {TPR, FNR, TNR, FPR, PPV, NPV} της κάθε δοκιμής. Όλη η διαδικασία πραγματοποιείται επαναληπτικά. Στο τέλος επιλέγεται το νευρωνικό δίκτυο με τα καλύτερα αποτελέσματα στα κριτήρια αξιολόγησης. Επιλέγεται δηλαδή ο βέλτιστος αριθμός νευρώνων και οι βέλτιστες αρχικές τιμές των παραμέτρων. Παρακάτω φαίνεται το διάγραμμα ροής του συστήματος αξιολόγησης:



Σχήμα 5.5.: Διάγραμμα Ροής Συστήματος Αξιολόγησης.

Η παραπάνω διαδικασία πραγματοποιείται 3 φορές, μία για κάθε υποψήφιο υποσύνολο χαρακτηριστικών και επιλέγεται τελικά αυτό που πετυχαίνει την καλύτερη απόδοση, μαζί με το πλήθος νευρώνων και τις αρχικές τιμές που το συνοδεύουν.

5.4 Αποτελέσματα

Στην ενότητα αυτή παραθέτουμε αναλυτικά τα αποτελέσματα της μεθοδολογίας μας και της αξιολόγησής της.

5.4.1 Συνάρτηση καταλληλότητας

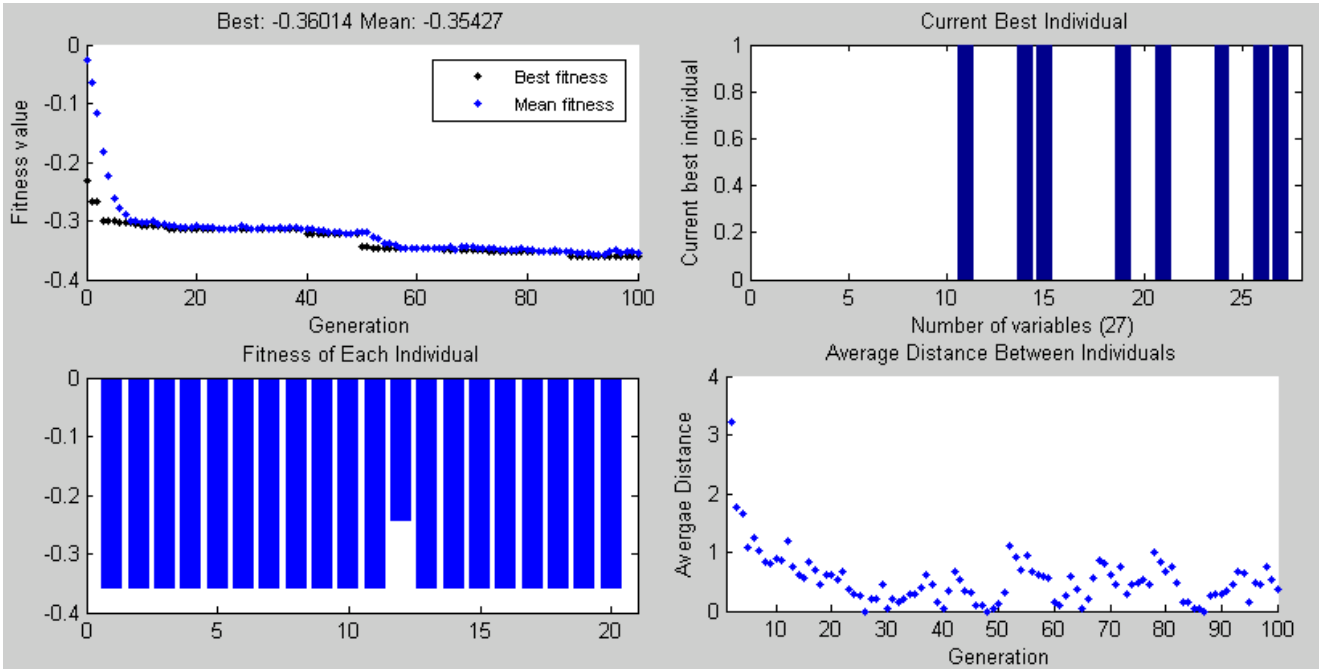
Πρώτα διεξαγάγαμε τον έλεγχο της αξιοπιστίας της συνάρτησης καταλληλότητας που χρησιμοποιούμε, με τον τρόπο που περιγράψαμε στην προηγούμενη ενότητα. Χρειάστηκε να αλλάξουμε την τιμή του βάρους β, καθώς έχουμε λιγότερα χαρακτηριστικά. Σύμφωνα με την εργασία των A. Verikas και M. Bacauskiene τα χαρακτηριστικά του συνόλου δεδομένων Wisconsin Breast Cancer κατατάσσονται από το πιο σημαντικό στο λιγότερο σημαντικό με την εξής σειρά: 6, 3, 1, 2, 7, 8, 4, 9, 5. Η συνάρτηση καταλληλότητας που χρησιμοποιούμε το καλύτερο σκορ πετυχαίνει το υποσύνολο {1, 2, 3, 6}. Σημειώνουμε στο σημείο αυτό, ότι η σειρά αυτή δεν εκφράζει τη σειρά με την οποία επιλέχθηκαν τα χαρακτηριστικά από τη συνάρτηση καταλληλότητας που χρησιμοποιούμε. Ο σκοπός μας δεν είναι η κατάταξη των χαρακτηριστικών από πλευράς σημασίας (feature ranking), όπως είναι στην εργασία [VB02], αλλά η επιλογή ενός "καλού" υποσυνόλου, και όπως φαίνεται από όσα εξηγήσαμε στο 5^{ο} κεφάλαιο, μας είναι άγνωστη η σειρά με την οποία επιλέγονται τα χαρακτηριστικά και δεν παίζει κανένα ρόλο στην εκπλήρωση του στόχου μας. Τα αμέσως επόμενα "καλά" υποσύνολα σύμφωνα με τη συνάρτηση καταλληλότητας μας είναι το {1, 3, 6} και το {1, 2, 3, 6, 7}. Όπως παρατηρούμε υπάρχει ταύτιση των αποτελεσμάτων, γεγονός που αποτελεί μία ένδειξη αξιοπιστίας της συνάρτησης καταλληλότητας.

5.4.2 Υποψήφια Υποσύνολα Χαρακτηριστικών

Στη συνέχεια εκτελέσαμε το στάδιο του γενετικού αλγορίθμου. Παραθέτουμε τα 3 καλύτερα υποσύνολα στα οποία καταλήξαμε μετά την ολοκλήρωση των επαναλήψεων. Κάθε ένα από αυτά συνοδεύεται από 4 γραφικές παραστάσεις. Σε αυτές αναπαρίστανται γραφικά: Η μέση και η βέλτιστη τιμή της συνάρτησης καταλληλότητας συναρτήσει των γενεών, η τιμή της συνάρτησης καταλληλότητας του κάθε χρωμοσώματος-ατόμου (individual) στην τελευταία γενιά, τα χαρακτηριστικά του καλύτερου χρωμοσώματος στην τελευταία γενιά και η μέση απόσταση μεταξύ των χρωμοσωμάτων σε κάθε γενιά.

- 1ο Υποσύνολο:

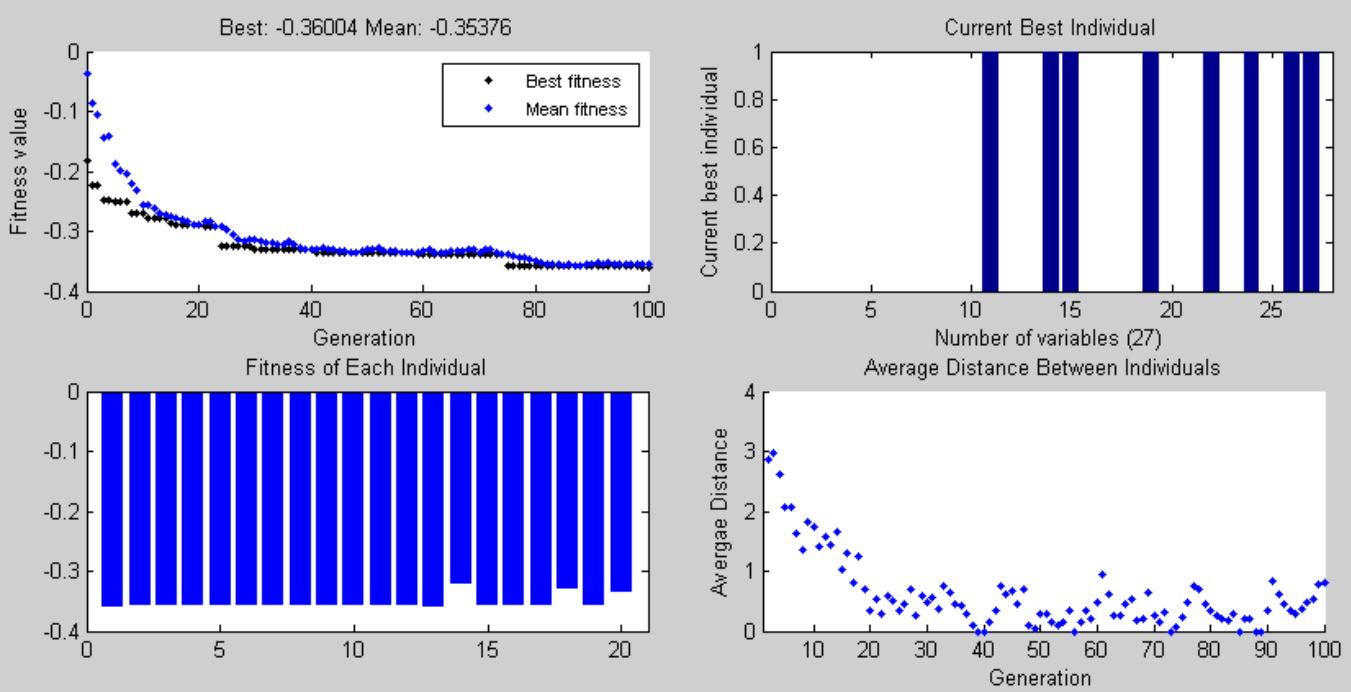
  {Γλυκοζυλιωμένη αιμοσφαιρίνη, Διαστολική αρτηριακή πίεση, Είδος θεραπείας, Λήψη διουρητικών, Υπολιπιδαιμική αγωγή, Οικογενειακό Ιστορικό Διαβήτη, Διάρκεια του Διαβήτη, Ηλικία}



Σχήμα 5.6: Γραφικές Παραστάσεις 1ου Υποψήφιου Υποσυνόλου.

- 2ο Υποσύνολο:

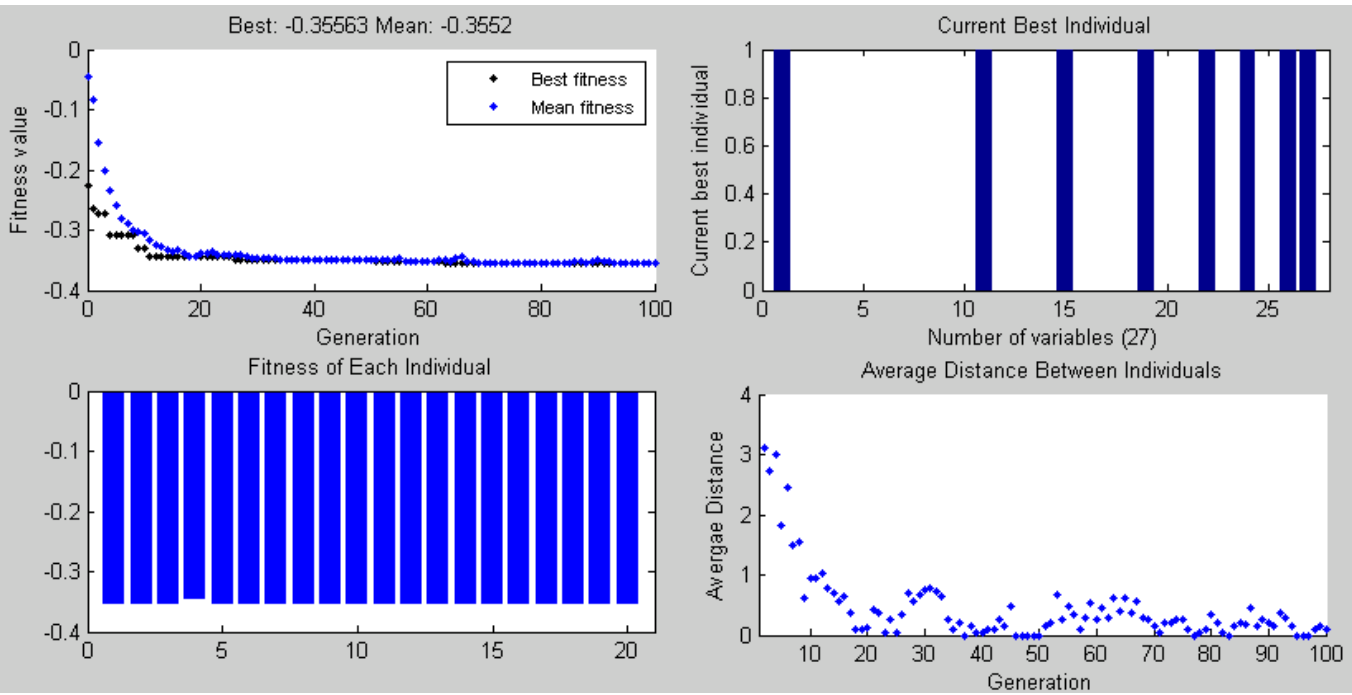
  {Γλυκοζυλιωμένη αιμοσφαιρίνη, Διαστολική αρτηριακή πίεση, Είδος θεραπείας, Λήψη διουρητικών, Λήψη ασπιρίνης, Οικογενειακό Ιστορικό Διαβήτη, Διάρκεια του Διαβήτη, Ηλικία}



Σχήμα 5.7: Γραφικές Παραστάσεις 2ου Υποψήφιου Υποσυνόλου.

- 3ο Υποσύνολο:

  {Δείκτης μάζας σώματος, Γλυκοζυλιωμένη αιμοσφαιρίνη, Είδος θεραπείας, Λήψη διουρητικών, Λήψη ασπιρίνης, Οικογενειακό Ιστορικό Διαβήτη, Διάρκεια του Διαβήτη, Ηλικία}



Σχήμα 5.8: Γραφικές Παραστάσεις 3ου Υποψήφιου Υποσυνόλου.

Παρατηρούμε ότι η συνάρτηση καταλληλότητας όλο και μειώνεται, καθώς από γενιά σε γενιά επιζούν τα καλύτερα χρωμοσώματα-λύσεις και μέσω διασταυρώσεων και μεταλλάξεων οδηγούν σε καλύτερες λύσεις. Σταδιακή μείωση παρουσιάζει επίσης η απόσταση μεταξύ των χρωμοσωμάτων, η οποία είναι ένδειξη της "ποικιλίας" του πληθυσμού, καθώς με το πέρασμα των γενεών ο αλγόριθμος συγκλίνει σε μία λύση. Η σύγκλιση αυτή φαίνεται και από την παρατήρηση του τελευταίου πληθυσμού του αλγόριθμου, όπου όλα τα χρωμοσώματα έχουν ίδια τιμή συνάρτησης καταλληλότητας, εκτός από ένα ή δύο. Στην πάνω δεξιά γραφική παράσταση φαίνονται οι αύξοντες αριθμοί των χαρακτηριστικών (έχουν παραταθεί στο 4^{ο} κεφάλαιο) που περιλαμβάνει το καλύτερο χρωμόσωμα, το οποίο αποτελεί έξοδο του γενετικού αλγορίθμου.

Στο 2^{ο} και στο 3^{ο} υποσύνολο δεν γίνεται αισθητή η αξία του μεγάλου αριθμού γενεών, καθώς ο αλγόριθμος καταλήγει στην έξοδό του από την 77^{η} και 63^{η} κιόλας γενιά, αντίστοιχα. Στο 1^{ο} όμως γίνεται οπτικά αντιληπτό το φαινόμενο που εξηγήσαμε και σε προηγούμενα κεφάλαια και που μας οδήγησε σε αυτή την απόφαση, καθώς ο αλγόριθμος βρίσκει μία καλύτερη λύση από αυτή που για αρκετές γενιές διατηρούσε, στην 90^{η} σχεδόν γενιά.

Αξίζει να αναφέρουμε ότι όσες φορές "έτρεξε" ο αλγόριθμος, το υποσύνολο στο οποίο κατέληγε περιλάμβανε πάντα, χωρίς ούτε μία εξαίρεση, τα χαρακτηριστικά: Γλυκοζυλιωμένη αιμοσφαιρίνη, Είδος θεραπείας, Οικογενειακό Ιστορικό Διαβήτη, Διάρκεια του Διαβήτη και Ηλικία. Τα αποτελέσματα στα οποία καταλήξαμε σε αυτό το στάδιο βρίσκονται σε συμφωνία με διάφορες επιδημιολογικές μελέτες που έχουν διεξαχθεί όσον αφορά στους παράγοντες που σχετίζονται με την εμφάνιση διαβητικής αμφιβληστροειδοπάθειας.

5.4.3 Τελικό Υποσύνολο Χαρακτηριστικών και Τελικό Μοντέλο Εκτίμησης Κινδύνου

Μετά την ολοκλήρωση της διαδικασίας της αξιολόγησης καταλήξαμε στο βέλτιστο πλήθος νευρώνων για το δίκτυο που αντιστοιχεί σε κάθε υποσύνολο χαρακτηριστικών. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα:

Πίνακας 5.1: Πλήθος Νευρώνων για το δίκτυο κάθε υποσυνόλου.

Υποσύνολο	Πλήθος Νευρώνων
1ο	14
2ο	18
3ο	17

Επίσης, αποθηκεύσαμε τις αρχικές τιμές των βαρών που αντιστοιχούν στο μοντέλο με την καλύτερη απόδοση για το κάθε υποσύνολο.

Όσον αφορά στην αξιολόγηση των υποψήφιων υποσυνόλων, είχαμε τα εξής αποτελέσματα:

Πίνακας 5.2: Αποτελέσματα Αξιολόγησης Υποψήφιων Υποσυνόλων.

	1ο Υποσύνολο			2ο Υποσύνολο			3ο Υποσύνολο		
	Test	**Train**	**All**	**Test**	**Train**	**All**	**Test**	**Train**	**All**
TPR(%)	69.02	76.40	75.66	72.09	78.47	77.83	**73.69**	78.04	77.60
FPR(%)	6.41	3.16	3.49	6.03	3.38	3.64	5.64	2.85	3.13
TNR(%)	93.59	96.84	96.51	93.97	96.62	96.36	94.36	97.15	96.87
FNR(%)	30.99	23.60	24.34	27.91	21.53	22.17	26.31	21.96	22.40
PPV(%)	78.30	88.90	87.81	80.36	88.49	87.63	81.29	90.07	89.16
NPV(%)	90.22	92.54	92.30	91.11	93.14	92.93	91.62	93.05	92.90

Στον παραπάνω πίνακα φαίνεται η τιμή του κάθε κριτηρίου αξιολόγησης του κάθε υποσυνόλου στα δεδομένα ελέγχου (Test), στα δεδομένα εκπαίδευσης (Train) και σε ολόκληρο το σύνολο των δεδομένων. Αυτό που μας ενδιαφέρει περισσότερο είναι η απόδοση στα δεδομένα ελέγχου.

Όπως φαίνεται από τον παραπάνω πίνακα, το υποσύνολο που πετυχαίνει την καλύτερη απόδοση είναι το 3^{ο}, που περιλαμβάνει τα χαρακτηριστικά: Δείκτης μάζας σώματος, Γλυκοζυλιωμένη αιμοσφαιρίνη, Είδος θεραπείας, Λήψη διουρητικών, Λήψη ασπιρίνης, Οικογενειακό Ιστορικό Διαβήτη, Διάρκεια του Διαβήτη, Ηλικία. Το Ποσοστό των Αληθώς Θετικών Αποτελεσμάτων στα δεδομένα ελέγχου είναι 73,69%, το Ποσοστό των Ψευδώς Θετικών είναι 5,64% και το Ποσοστό των Ψευδώς Αρνητικών είναι 26,31%.

Αρκετά ικανοποιητικά αποτελέσματα και κοντινές τιμές στα κριτήρια αξιολόγησης επιτυγχάνονται και με το 2^{ο} υποσύνολο, που αποτελείται από τα χαρακτηριστικά: Γλυκοζυλιωμένη αιμοσφαιρίνη, Διαστολική αρτηριακή πίεση, Είδος θεραπείας, Λήψη διουρητικών, Λήψη ασπιρίνης, Οικογενειακό Ιστορικό Διαβήτη, Διάρκεια Διαβήτη και Ηλικία. Υπενθυμίζουμε ότι τα βάρη και οι πολώσεις αρχικοποιούνται με τυχαίο τρόπο και έγιναν πολλές δοκιμές για την εύρεση των βέλτιστων αρχικών τιμών τους. Εφόσον η απόδοση του 2^{ου} και του 3^{ου} υποσυνόλου έχει τόσο κοντινές τιμές, δεν αποκλείεται με κάποια αρχική τιμή βαρών που δεν συναντήθηκε κατά τη διερεύνησή μας να επιτυγχάνεται καλύτερη απόδοση με το 2^{ο} υποσύνολο.

Γενικά, η απόδοση που επιτυγχάνεται δεν είναι πολύ υψηλή, αλλά είναι ικανοποιητική. Σημαντικό είναι να λάβουμε υπόψη άλλωστε, ότι η μέθοδος της Διασταυρωμένης Επικύρωσης που ακολουθήσαμε αποτελεί μία αρκετά αυστηρή διαδικασία αξιολόγησης. Όπως είπαμε, μετά το χωρισμό του συνόλου των δεδομένων σε 10 ίσα μέρη, το κάθε μέρος αποτελείται από περίπου 104 στιγμιότυπα με 78 “0” και 26 “1”. Επομένως στα δεδομένα ελέγχου, τα οποία σε κάθε επανάληψη είναι ένα από τα 10 μέρη, για κάθε Θετικό Αποτέλεσμα που ταξινομείται από το μοντέλο ως Αρνητικό (Ψευδώς Αρνητικό), το Ποσοστό των Αληθώς Θετικών Αποτελεσμάτων (TPR) και το Ποσοστό των Ψευδώς Αρνητικών Αποτελεσμάτων (FNR) μειώνεται και αυξάνεται αντίστοιχα κατά 1/26 ≈ 3,8%. Κάθε λάθος δηλαδή, τιμωρείται πολύ αυστηρά. Επιπλέον, ενδέχεται κάποια ζευγάρια δεδομένων εκπαίδευσης και ελέγχου να είναι πολύ καλύτερα από άλλα, όσον αφορά στην

“κατανόησή” τους από το δίκτυο, με αποτέλεσμα, κάποια ζευγάρια να χαλάνε τους μέσους όρους των κριτηρίων απόδοσης.

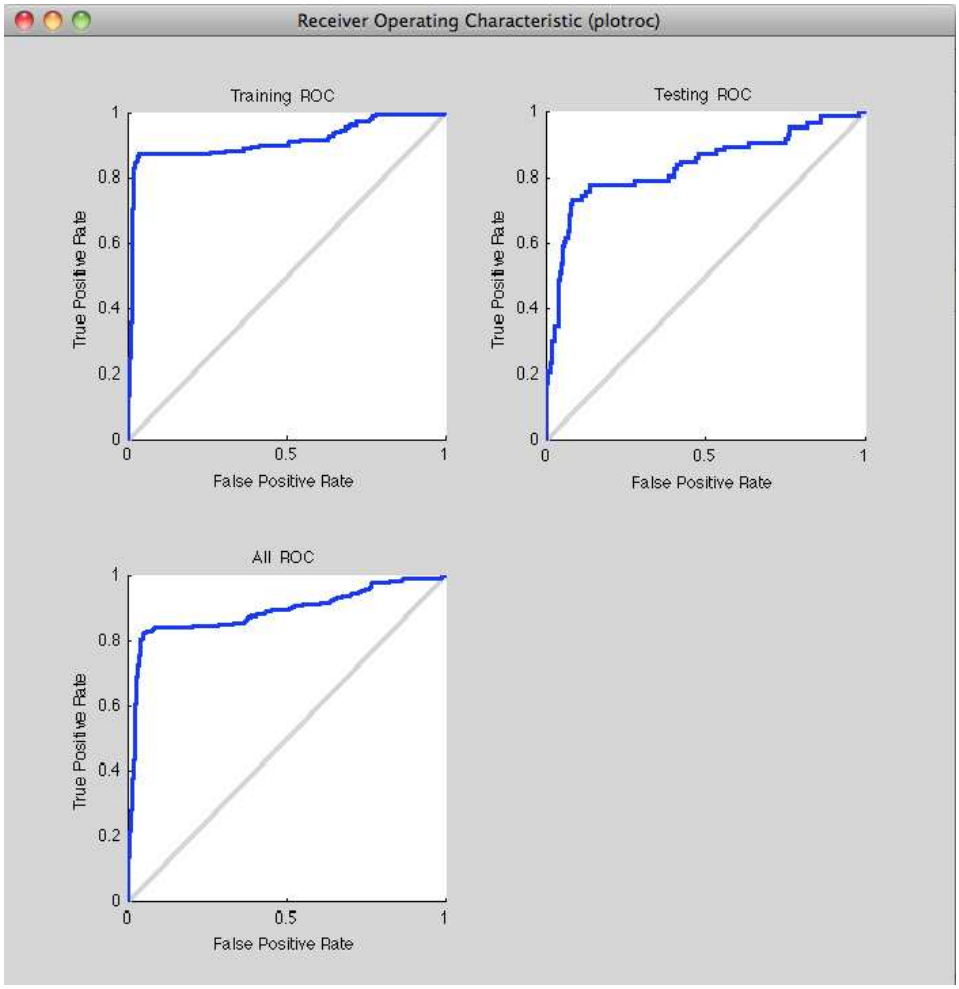
Υλοποιήσαμε στη συνέχεια το τελικό μοντέλο εκτίμησης κινδύνου εμφάνισης αμφιβληστροειδοπάθειας. Διατηρώντας το πλήθος των νευρώνων και τις αρχικές τιμές των παραμέτρων που προέκυψαν από το προηγούμενο βήμα ως βέλτιστες, εκπαιδεύουμε το Νευρωνικό Δίκτυο με τα 2/3 των δεδομένων και ελέγχουμε την απόδοσή του στο 1/3. Και πάλι έχει προηγηθεί ο διαχωρισμός των δεδομένων σε 3 ομοιογενή μέρη. Τα αποτελέσματα των κριτηρίων αξιολόγησης φαίνονται στον παρακάτω πίνακα:

Πίνακας 5.3: Αποτελέσματα Αξιολόγησης Τελικού Μοντέλου.

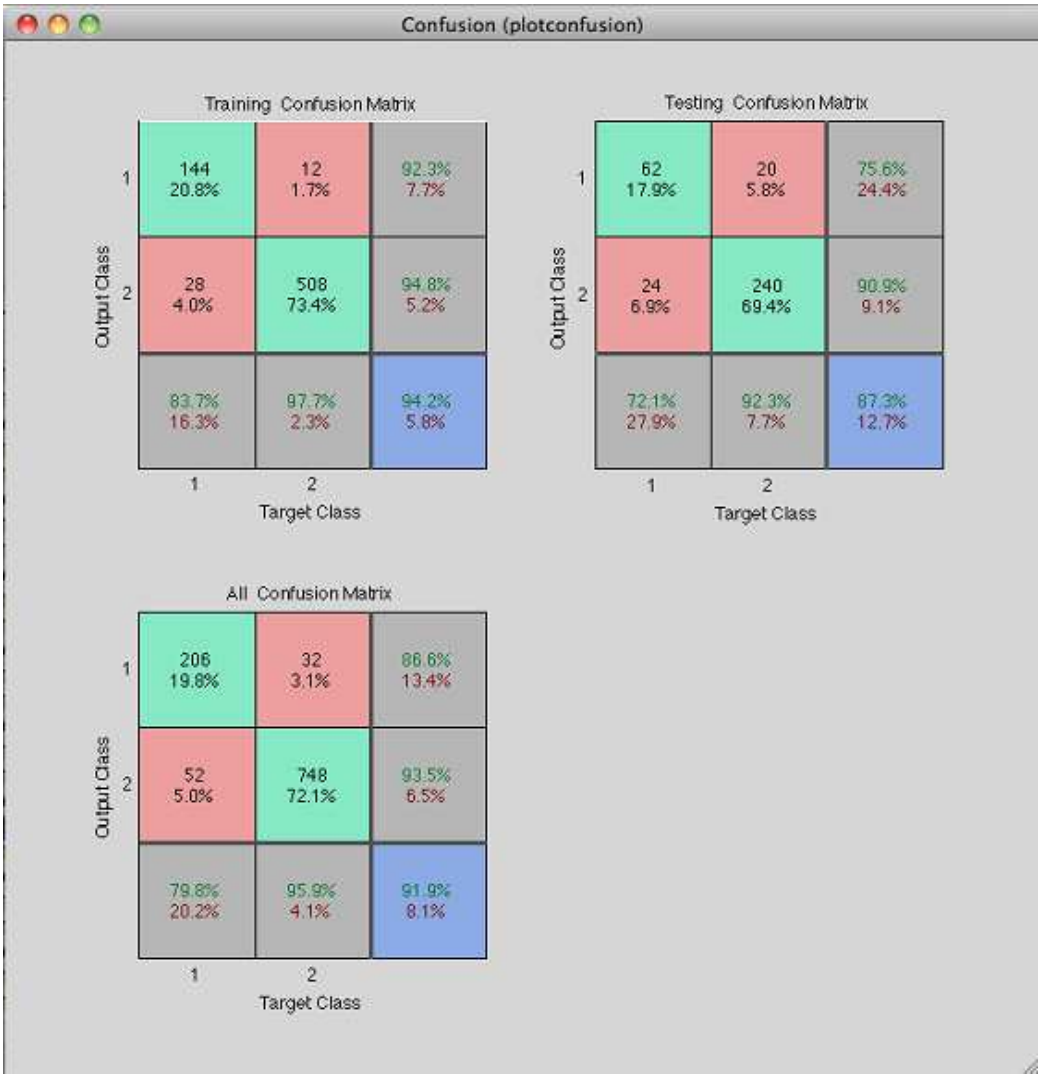
	Test	**Train**	**All**
TPR(%)	72.09	83.72	79.84
FPR(%)	7.69	2.31	4.10
TNR(%)	92.31	97.69	95.90
FNR(%)	27.91	16.28	20.16
PPV(%)	75.61	92.31	86.55
NPV(%)	90.91	94.78	93.50

Επίσης, η ακρίβεια (accuracy) του τελικού μοντέλου είναι 94,2% για τα δεδομένα εκπαίδευσης, 87,3% για τα δεδομένα ελέγχου και 91,9% για όλα τα δεδομένα μαζί.

Παραθέτουμε τη ROC καμπύλη και τη Μήτρα Σύγχυσης (Confusion Matrix) του τελικού μας μοντέλου για τα δεδομένα εκπαίδευσης, τα δεδομένα ελέγχου και για ολόκληρο το σύνολο δεδομένων:



Σχήμα 5.9: ROC καμπύλες Τελικού Μοντέλου.



Σχήμα 5.10: Μήτρες Σύγχυσης Τελικού Μοντέλου.

5.5 Σύνοψη συμπερασμάτων αξιολόγησης

Στο σημείο αυτό συνοψίζουμε τα συμπεράσματα της αξιολόγησης.

1. Το καλύτερο υποσύνολο χαρακτηριστικών για την εκτίμηση κινδύνου εμφάνισης αμφιβληστροειδοπάθειας κρίθηκε το {Δείκτης μάζας σώματος, Γλυκοζυλιωμένη αιμοσφαιρίνη, Είδος θεραπείας, Λήψη διουρητικών, Λήψη ασπιρίνης, Οικογενειακό Ιστορικό Διαβήτη, Διάρκεια του Διαβήτη, Ηλικία}.
2. Το βέλτιστο πλήθος νευρώνων κρυφού επιπέδου για το συγκεκριμένο υποσύνολο είναι 17.
3. Οι τιμές των κριτηρίων απόδοσης που επιτυγχάνονται για το καλύτερο υποσύνολο στην διαδικασία της Διασταυρωμένης Επικύρωσης σε 10 μέρη στα δεδομένα ελέγχου είναι: TPR=73,69%, FPR=5,64%, TNR=94,36%, FNR=26,31%, PPV=81,29% και NPV=91,62%.

4. Η απόδοση του τελικού μοντέλου εκτίμησης κινδύνου εμφάνισης αμφιβληστροειδοπάθειας στα δεδομένα ελέγχου είναι: TPR=72,09%, FPR=7,96%, TNR=92,31%, FNR=27,91%, PPV=75,61% και NPV=90,91%.

6

Επίλογος

Στο κεφάλαιο αυτό συνοψίζεται η συνολική παρουσίαση της διπλωματικής τα συμπεράσματά της και προτείνονται κάποιες μελλοντικές επεκτάσεις της.

6.1 Σύνοψη και συμπεράσματα

Στα πλαίσια της παρούσας διπλωματικής αναζητήθηκαν προγνωστικοί παράγοντες για την εμφάνιση διαβητικής αμφιβληστροειδοπάθειας και μελετήθηκε η δυνατότητα ανάπτυξης ενός μοντέλου για την εκτίμηση του κινδύνου εμφάνισής της μακροπρόθεσμα. Για το σκοπό αυτό μελετήθηκαν εκτενώς εργασίες που έχουν γίνει σε αυτόν τον τομέα και μέθοδοι Τεχνητής Νοημοσύνης για επιλογή χαρακτηριστικών και ταξινόμηση. Υλοποιήθηκε ένας Γενετικός Αλγόριθμος με συνάρτηση καταλληλότητας που βασίζεται στην έννοια της Αμοιβαίας Πληροφορίας και χρησιμοποιήθηκε για τον προσδιορισμό υποσυνόλων χαρακτηριστικών που σχετίζονται με την εμφάνιση της νόσου. Υλοποιήθηκαν Τεχνητά Νευρωνικά Δίκτυα ως μοντέλα εκτίμησης του κινδύνου εμφάνισης της νόσου και χρησιμοποιήθηκαν για την αξιολόγηση της σχετικότητας των υποψήφιων υποσυνόλων με την εμφάνισή της. Αναπτύχτηκε, στη συνέχεια, το τελικό μοντέλο εκτίμησης του κινδύνου εμφάνισης της νόσου που βασίζεται στο περισσότερο σχετικό υποσύνολο χαρακτηριστικών. Τέλος, αξιολογήθηκε η απόδοση της μεθοδολογίας που ακολουθήθηκε.

Καταλήξαμε στο συμπέρασμα ότι τα χαρακτηριστικά που αποτελούν προδιαθεσικούς παράγοντες για την εμφάνιση διαβητικής αμφιβληστροειδοπάθειας είναι ο Δείκτης μάζας σώματος, τα επίπεδα της Γλυκοζυλιωμένης αιμοσφαιρίνης, το Είδος της θεραπείας που ακολουθείται, η Λήψη

διουρητικών, η Λήψη ασπιρίνης, το Οικογενειακό Ιστορικό όσον αφορά στο Διαβήτη, η Διάρκεια του Διαβήτη και η Ηλικία. Το αποτέλεσμα αυτό βρίσκεται σε συμφωνία με προηγούμενες επιδημιολογικές μελέτες σχετικά με την εμφάνιση της νόσου. Από το υποσύνολο στο οποίο καταλήξαμε λείπει η διαστολική αρτηριακή πίεση, που από κάποιες μελέτες έχει προκύψει ότι αποτελεί σημαντικό προδιαθεσικό παράγοντα. Το χαρακτηριστικό αυτό όμως περιλαμβάνεται στο αμέσως επόμενο σε απόδοση υποσύνολο, που δεν είχε σημαντικές διαφορές από το πρώτο στα κριτήρια αξιολόγησης που υπολογίστηκαν και μπορεί να χρησιμοποιηθεί εναλλακτικά.

Φαίνεται επομένως ότι ο Γενετικός Αλγόριθμος σε συνδυασμό με την Αμοιβαία Πληροφορία είναι σε θέση να καταλήξουν σε πολύ ικανοποιητικά αποτελέσματα για το πρόβλημα το οποίο πραγματευόμαστε, χάρις στην ευέλικτη αναζήτηση και την ικανότητα ανίχνευσης αλληλεξαρτήσεων που προσφέρουν.

Το μοντέλο που αναπτύχθηκε δεν πέτυχε πολύ υψηλές τιμές απόδοσης, αλλά σε μεγάλο βαθμό ικανοποιητικές. Μία πιθανή πηγή μείωσης της απόδοσής του είναι τα δεδομένα, στα οποία βασιστήκαμε. Η ποιότητα των δεδομένων είναι πολύ σημαντική κατά την ανάπτυξη ενός μοντέλου Μηχανικής Μάθησης και επηρεάζει πολύ την απόδοσή του. Τα δεδομένα μας περιείχαν πολλές ελλείψεις και χρειάστηκε να κάνουμε κάποιους συμβιβασμούς για τη διαμόρφωση ενός αξιοπρεπούς συνόλου δεδομένων. Παρά τις προσπάθειές μας, οι χειρισμοί αυτοί ενδέχεται να οδηγούν σε έλλειψη πληροφορίας και επομένως, σε χαμηλότερη απόδοση του μοντέλου. Επιπλέον, ενώ το απλό Τεχνητό Νευρωνικό Δίκτυο που χρησιμοποιήσαμε φαίνεται να παράγει αρκετά ικανοποιητικές προβλέψεις, ίσως ένα άλλο πιο πολύπλοκο είδος Νευρωνικού Δικτύου να ανταποκρίνεται καλύτερα στις απαιτήσεις ενός τέτοιου περίπλοκου προβλήματος. Πιθανώς, όπως αναφέρεται και στην επόμενη ενότητα, σε κάποια μελλοντική επέκταση της μεθοδολογίας μας να χρησιμοποιηθεί ένα άλλο είδος μοντέλου που να λύσει πιο αποτελεσματικά το πρόβλημα.

Το μοντέλο που υλοποιήθηκε σίγουρα δεν μπορεί να χρησιμοποιηθεί σαν κατ' εξοχήν εργαλείο επιλογής της κατάλληλης θεραπείας και τακτικής για το άτομο, αλλά μπορεί να χρησιμοποιηθεί ως "σύμβουλος" για την επισήμανση του κινδύνου. Επομένως, μπορούμε να πούμε ότι σε ένα βαθμό πέτυχε το στόχο του.

6.2 Μελλοντικές επεκτάσεις

Πιθανές μελλοντικές επεκτάσεις της εφαρμογής που αναπτύχθηκε στα πλαίσια της παρούσας διπλωματικής είναι οι εξής :

- Χρήση άλλων ειδών Τεχνητών Νευρωνικών Δικτύων για την υλοποίηση του μοντέλου εκτίμησης κινδύνου. Στη μεθοδολογία μας χρησιμοποιήθηκε ένα απλό πολυστρωματικό δίκτυο πρόσθιας τροφοδότησης. Είναι πολύ πιθανό με χρήση άλλων, πιο "δυνατών" δικτύων, όπως για παράδειγμα τα Νευρωνικά Δίκτυα Κυματίων (Wavelet Neural Networks),

να "αποτυπώνεται" πολύ καλύτερα το πρόβλημα και να επιτυγχάνεται υψηλότερη απόδοση. Άλλη μία πιθανή μελλοντική επέκταση είναι η χρήση Δυναμικών Νευρωνικών Δικτύων (Dynamic Neural Networks), στα οποία η έξοδος δεν εξαρτάται μόνο από τις τρέχουσες εισόδους, αλλά και από προηγούμενες εισόδους, εξόδους και καταστάσεις του δικτύου. Με τον τρόπο αυτό, μπορεί να ληφθεί καλύτερα υπόψη η χρονική εξέλιξη των χαρακτηριστικών και η εξάρτηση της κατάστασης του ασθενή από όλα τα προηγούμενα έτη.

- Εφαρμογή "δια βίου εκπαίδευσης". Η έννοια αυτή αφορά στην ικανότητα του δικτύου να συνεχίσει να εκπαιδεύεται συνεχώς με μελλοντικά δεδομένα, όταν η πραγματική έξοδος γίνει γνωστή. Με τον τρόπο αυτό, το μοντέλο γίνεται συνεχώς καλύτερο και πιο ικανό να παράγει την σωστή πρόβλεψη. Η επέκταση αυτή προϋποθέτει έλεγχο της εγκυρότητας των μελλοντικών δεδομένων, αποθήκευση της πρόβλεψης που είχε παραχθεί και ενσωμάτωση των νέων δεδομένων στα δεδομένα εκπαίδευσης.
- Χρήση πιο περίπλοκης συνάρτησης καταλληλότητας στο Γενετικό Αλγόριθμο, με σκοπό να λαμβάνονται υπόψη περισσότερα είδη αλληλεξαρτήσεων, όπως για παράδειγμα η αλληλεξάρτηση των χαρακτηριστικών ανά δύο με την έξοδο.
- Ενσωμάτωση και διερεύνηση περισσότερων χαρακτηριστικών, όπως για παράδειγμα η σωματική άσκηση.
- Χρήση της μεθοδολογίας για εκτίμηση κινδύνου εμφάνισης και άλλων μακροπρόθεσμων επιπλοκών, χάρις στη γενικότητά της.
- Χρήση πληρέστερου συνόλου δεδομένων. Η ανάπτυξη του μοντέλου που προτείνεται με καλύτερες "πρώτες ύλες" μπορεί να το βελτιώσει σημαντικά και να το αναδείξει σε καλύτερο εργαλείο εκτίμησης του κινδύνου εμφάνισης αμφιβληστροειδοπάθειας.
- Ενσωμάτωση του μοντέλου σε διαδικτυακή εφαρμογή (web application), ώστε να είναι ευρέως διαθέσιμο. Η κίνηση αυτή μπορεί να προσφέρει στην ενημέρωση και στην αφύπνιση των ασθενών σχετικά με τον κίνδυνο που διατρέχουν, με στόχο κάποια αλλαγή στη θεραπεία ή στη διατροφή τους.

7

Βιβλιογραφία

[ΒΚΒ+06] Ι.Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, Η. Σακελλαρίου, Τεχνητή Νοημοσύνη, Γ' Έκδοση, Β. Γκιούρδας Εκδοτική, 2006.

[Δια07] Κ. Διαμαντάρας, Τεχνητά Νευρωνικά Δίκτυα, Εκδόσεις Κλειδάριθμος, 2007.

[ΠΣ04] Ν. Πετρόγλου, Λ. Σπάρος, Καμπύλη ROC στη διαγνωστική έρευνα, Εφαρμοσμένη Ιατρική Έρευνα, Αρχεία Ελληνικής Ιατρικής 21(2), 179-194, 2004.

[ΣΑΑ+07] Γ. Στεργίου, Η. Αβραμόπουλος, Ε. Ανδρεάδης, Α. Αχείμαστος, Ε. Βαρσαμής, Κ. Βέμμος, Δ. Βλαχάκος, Μ. Ελισάφ, Ν. Καρατζάς, Θ. Μουντοκαλάκης, Δ. Παπαδογιάννης, Κ. Σιαμόπουλος, 41 Πρακτικές Ερωτήσεις και Απαντήσεις για την Υπέρταση και τη Χοληστερίνη, Ελληνική Εταιρεία Μελέτης της Υπέρτασης, 2007, available at: http://www.hypertension.gr.

[Στε10] Χ. Στεφανάδης, Ασπιρίνη και σακχαρώδης διαβήτης, Ιούνιος 2010, available at: http://www.sugarfree.gr/content/view/1173/50/.

[ADC03] A. Al-Ani, M. Deriche, J. Chebil, A new mutual information based measure for feature selection, Intelligent Data Analysis 7, 43–57, 2003.

[BH00] I.A. Basheer, M. Hajmeer, Artificial neural networks: fundamentals,

computing, design, and application, Journal of Microbiological Methods 43, 3–31, 2000.

[BM09] K. Bailly, M. Milgram, Boosting feature selection for Neural Network based regression, Neural Networks 22, 748-756, 2009.

[BM92] K. P. Bennett, O. L. Mangasarian, Robust linear programming discrimination of two linearly inseparable sets, Optimization Methods and Software 1, 23-34 (Gordon & Breach Science Publishers), 1992.

[BML01] D. Boeri, M. Maiello, M. Lorenzi, Increased Prevalence of Microthromboses in Retinal Capillaries of Diabetic Individuals, Diabetes 50, 1432-1439, June 2001.

[BRC+00] J.B. Brown, A. Russell, W. Chan, et al. The global diabetes model user friendly version 3.0. Diabetes Research and Clinical Practice, vol.50, S15-S46, 2000.

[Bro09] G. Brown, A New Perspective for Information Theoretic Feature Selection, Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS), Vol. 5, 2009.

[Can04] E. Cantu-Paz, Feature Subset Selection, Class Separability and Genetic Algorithms, Genetic and Evolutionary Computation Conference, January 2004.

[CCJ92] J. K. Canner, Y.P. Chiang, and J. Javitt, A Monte Carlo based simulation network model for a chronic progressive disease: the case of diabetic retinopathy, WSC '92: Proceedings of the 24th conference on Winter simulation, pp. 1041—1049, 1992.

[CGB+04] P.M. Clarke, A.M. Gray, A. Briggs, A.J. Farmer, P. Fenn, R.J. Stevens, D.R. Matthews, I.M. Stratton, R.R. Holman, A model to estimate the lifetime health outcomes of patients with Type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no. 68), Diabetologia, vol.47, 1747–1759, 2004.

[CSG+96] T. Cibas, F. F. Soulie, P. Gallinari, S. Raudys, Variable selection with neural networks, Neurocomputing 12, 223-248, 1996.

[DBR+10] C. Deisy, S. Baskar, N. Ramraj, J. Saravanan Koori, P. Jeevanandam, A novel information theoretic-interact algorithm (IT-IN) for feature selection using three machine learning algorithms, Expert Systems with Applications 37, 7589–7597, 2010.

[Dia93] The Diabetes Control and Complications Trial Research Group , The Effect of Intensive Treatment of Diabetes on the Development and Progression of Long-Term Complications in Insulin-Dependent Diabetes Mellitus, Volume 329, Number 14, 977-986, September 1993.

[DL97] M. Dash, H. Liu, Feature Selection for Classification, Intelligent Data Analysis 1, 131–156, March 1997.

[EJG+00] R. O. Estacio, B. W. Jeffers, N. Gifford, R. W. Schrier, Effect of blood pressure control on diabetic microvascular complications in patients with hypertension and type 2 diabetes, Diabetes Care 23 (Suppl. 2), B54–B64, 2000.

[EJH+97] R. C. Eastman, J. C. Javitt, W. H. Herman et al., Model of complications of NIDDM. I. Model construction and assumptions. Diabetes Care 20, 725–734, 1997.

[ElA09] M.E. ElAlami, A filter model for feature subset selection based on genetic algorithm, Knowledge-Based Systems 22, 356–362, 2009.

[EOE+04] M. S. Eberhardt, C. Ogden, M. Engelgau, B. Cadwell, A. A. Hedley, S.H. Saydah, Prevalence of Overweight and Obesity among adults with Diagnosed Diabetes, United States 1988-1994 and 1999-2002, Centers for Disease Control and Prevention, Weekly, 53(45), 1066-1068, November 2004, available at: hptt://www.cdc.gov/mmwr/preview/mmwrhtml /mm5345a2.htm.

[FAG+04] D. S. Fong, L. Aiello, T. W. Gardner, G. L. King, G. Blankenship, F. L. Ferris, R. Klein, Retinopathy in Diabetes, Diabetes Care, Vol. 27, Supplement 1, January 2004.

[FM09] B. Firestone, J. W. Mold, Type 2 diabetes: Which interventions best reduce absolute risks of adverse events? , The Journal of family practice, vol. 58, 6:E1, 2009.

[GE03] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3, 1157-1182, 2003.

[GG10] D. Ghosh, R. Guha, Use of genetic algorithm and neural network approaches for risk factor selection: A case study of West Nile virus dynamics in an urban environment. Computers, Environment and Urban Systems 34, 189–203, February 2010.

[Gre86] J.J. Grefenstette, Optimization of Control Parameters for Genetic Algorithms, IEEE Trans. Systems, Man, and Cybernetics, SMC-16(1) 122-128, 1986.

[Hal99] M. A. Hall, Correlation-based Feature Selection for Machine Learning, Department of Computer Science, Hamilton, New Zealand, 1999.

[HCX07] J. Huang, Y. Cai, X. Xu, A hybrid genetic algorithm for feature selection wrapper based on mutual information, Pattern Recognition Letters 28, 1825–1844, May 2007.

[HMB+07] Y. Huang, P. McCullagh, N. Black, R. Harper, Feature selection and classification model construction on type 2 diabetic patients' data, Artificial Intelligence in Medicine 41, 251-262, 2007.

[HRK+99] H. Handels, Th. Roß, J. Kreusch, H.H. Wolff, S.J. Pöppl, Feature selection for optimized skin tumor recognition using genetic algorithms, Artificial Intelligence in Medicine 16, 283–297,1999

[IML+01] I. Inza, M. Merino, P. Larrañaga, J. Quiroga, B. Sierra, M. Girala, Feature Subset Selection by genetic algorithms and estimation of distribution algorithms. A case study in the survival of cirrhotic patients treated with TIPS, Artificial Intelligence in Medicine 23, 187-205, May 2001.

[IN00] H. Ishibuchi, T. Nakashima, Multi-objective pattern and feature selection by a genetic algorithm, Proceedings of Genetic and Evolutionary Computation Conference (Las Vegas, Nevada, U.S.A.), 1069-1076, July 2000.

[Jak05] A. Jakulin, Machine Learning Based on Attribute Interactions, PhD Dissertation, June 13, 2005.

[JB04] A. Jakulin, I. Bratko, Quantifying and Visualizing Attribute Interactions: An Approach Based on Entropy, March 2004.

[KC02] N. Kwak, C. Choi, Input Feature Selection for Classification Problems, IEEE Transactions on Neural Networks, vol. 13, No. 1, 143-159, January 2002.

[KIM10] M. Kabir, M. Islam, K. Murase, A new wrapper feature selection approach using neural network, Neurocomputing, 2010.

[Kit75] J. Kittler, Mathematical Methods of Feature Selection in Pattern Recognition, Int. J. Man-Machine Studies 7, 609-637, 1975.

[KJ97] R. Kohavi, G. H. John, Wrappers for feature subset selection, Artificial Intelligence 97, 273-324, 1997.

[KKL+09] R. Klein, M. D. Knudtson, K. E. Lee, R. Gangnon, B. E. K. Klein, The Wisconsin Epidemiologic Study of Diabetic Retinopathy XXIII. The Twenty-Five-Year Incidence of Macular Edema on Persons with Type 1 Diabetes, Ophthalmology 116(3), 497–503, March 2009.

[KKM+89] R. Klein, B. E. K. Klein, S. E. Moss, M. D. Davis, D. L. DeMets, The Wisconsin Epidemiologic Study of Diabetic Retinopathy IX. Four-Year Incidence and Progression of Diabetic Retinopathy When Age at Diagnosis Is Less Than 30 Years, Arch Ophthalmol, 107, 237-243, 1989.

[KKP06] S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, Data Preprocessing for Supervised Leaning, International Journal Of Computer Science, vol. 1, No. 2, 1306-4428, 2006.

[KO08] Y-W. Kim, Il-S. Oh, Classifier ensemble selection using hybrid genetic algorithms, Pattern Recognition Letters 29, 796–802, 2008.

[Koh95] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of International Joint Conference on Artificial Intelligence, 1137–1145, 1995.

[Kot07] S. B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31, 249-268, 2007.

[LDM+03] H.A van Leiden, J. M. Dekker, A.C. Moll, G. Nijpels, R. J. Heine, L. M. Bouter, C. D. A. Stehouser, B. C. P. Polak, Risk Factors for Incident Retinopathy in a Diabetic and Nondiabetic Population, The Hoorn Study, Arch Ophthalmol. 121, 245-251, 2003.

[LSL+09] H. Liua, J. Sun, L. Liua, H. Zhang, Feature selection with dynamic mutual information, Pattern Recognition 42, 1330-1339, 2009.

[MDJ09] Y. Marinakis, G. Dounias, J. Jantzen, Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection, Computers in Biology and Medicine 39 , 69-78, 2009.

[MW90] O. L. Mangasarian, W. H. Wolberg, Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, pp 1 & 18, September 1990.

[PBV+00] A. Palmer, A. Brandt, G. Valerio, C. Weiss, H. Stock, H. Wenzel, Outline of a diabetes disease management model. Principles and applications, Diabetes Research and Clinical Practice, vol. 50, S47–S56, 2000.

[PE01] J. Park, D. W Edington, A sequential neural network model for diabetes

prediction, Artificial Machine Learning in Medicine 23, 277-293, 2001.

[PLD05] H. Peng, F. Long, C. Ding, Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance and Min-Redundancy, IEEE Transactions of Pattern Analysis and Machine Intelligence, vol. 27, No. 8, August 2005.

[PRV+04] AJ. Palmer, S. Roze, WJ. Valentine, ME. Minshall, V. Foos, FM. Lurati, M. Lammert, GA. Spinas, The CORE Diabetes Model: projecting long-term clinical outcomes, costs and cost-effectiveness of interventions in diabetes mellitus (types 1 and 2) to support clinical and reimbursement decision making, Curr Med Res Opin, vol. 20, Suppl. 1, 5-26, 2004

[PSC+01] M. Porta, A-K. Sjoelie, N. Chaturvedi, L. Stevens, R. Rottiers, M. Veglio, J.H. Fuller and the EURODIAB Prospective Complications Study Group, Risk factors for progression to proliferative diabetic retinopathy in the EURODIAB Prospective Complications Study, Diabetologia 44, 2203-2209, 2001.

[RTL08] P. Refaeilzadeh, L. Tang, H. Liu, Cross-Validation, Arizona State University, 2008.

[PWS+99] A.J. Palmer, C. Weiss, P.P. Sendi, K. Neeser, A. Brandt, G. Singh, H. Wenzel, G.A. Spinas, 'The cost-effectiveness of different management strategies for type I diabetes: a Swiss perspective', Diabetologia, vol. 43, 13–26, 1999.

[Ris97] K. M. Risvik, Discretization of Numerical Attributes: Preprocessing for Machine Learning, Computer Science Projects, Knowledge Systems Group Department of Computer and Information Science, Norwegian University of Science and Technology, April 1997.

[RN05] S.J. Russell, P.Norvig, Τεχνητή Νοημοσύνη: Μια σύγχρονη προσέγγιση, Δεύτερη Αμερικανική Έκδοση, εκδόσεις Κλειδάριθμος, 2005.

[SAG+00] R. J. Stevens, A. I. Adler, A. Gray, A. Briggs, R. Holman, Life-expectancy projection by modeling and computer simulation (UKPDS 46), Diabetes Research and Clinical Practice 50 Suppl. 3, S5–S13, 2000.

[SKA01] R.J. Stevens, V. Kothari, A.I. Adler, et al., The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56), Clinical Science, vol. 101, pp. 671-679, 2001.

[SP10] J. M. Sotoca, F. Pla, Supervised feature selection by clustering using conditional mutual information-based distances, Pattern Recognition 43,

2068–2081, 2010.

[SR06] M. Soryani, N. Rafat, Application of Genetic Algorithms to Feature Subset Selection in a Farsi OCR, World Academy of Science, Engineering and Technology 18, 2006.

[SR07] R. K. Sivagaminathan, S. Ramakrishnan, A hybrid approach for feature subset selection using neural networks and ant colony optimization, Expert Systems with Applications 33, 49–60, 2007.

[SS89] W. Siedlecki, J. Sklansky, A note on genetic algorithms for large-scale feature selection, Pattern Recognition Letters 10, 335-347, April 1989.

[SZK+10] M. Skevofilakas, K. Zarkogianni, B. G. Karamanos, K. S. Nikita, A hybrid Decision Support System for the Risk Assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus, 2010.

[TFM+01] G. D. Tourassi, E. D. Frederick, M. K. Markey, C. E. Floyd, Jr., Application of the mutual information criterion for feature selection in computer-aided diagnosis, Medical Physics, Vol. 28, No. 12, December 2001.

[TYT09] H. Temurtas, N. Yumusak, F. Temurtas, A comparative study on diabetes disease diagnosis using neural networks, Expert Systems with Applications 36, 8610–8615, 2009.

[UKP98a] UK Prospective Diabetes Study (UKPDS) Group: Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34), Lancet 12, 352, 854–865, 1998.

[UKP98b] UK Prospective Diabetes Study (UKPDS) Group, Tight blood pressure control and risk of macrovascular and microvascular complications in type 2diabetes (UKPDS 38), British Medical Journal, vol. 317, 703–713, 1998.

[VB02] A. Verikas, M. Bacauskiene, Feature selection with neural networks, Pattern Recognition Letters 23, 1323–1335, 2002.

[VDY04] V. Venkatraman, A. R. Dalby, Z. Yang, Evaluation of Mutual Information and Genetic Programming for Feature Selection in QSAR, J. Chem. Inf. Comput. Sci., Vol. 44, No. 5, 2004.

[VZ07] B. Verma, P. Zhang, A novel neural-genetic algorithm to find the most significant combination of features in digital mammograms, Applied Soft Computing 7, 612–625, 2007.

[YH97] J. Yang , V. Honavar, Feature Subset Selection using a Genetic Algorithm, Artificial Intelligence Research Group, May 1997.

[YM99] H.H. Yang, J. Moody, Feature Selection Based on Joint Mutual Information, Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis, Rochester New York, 1999.