Πρόβλεψη του Ισομερισμού των Προλινών με χρήση Συμβολικών και Δομικών Χαρακτηριστικών

Παρασκευή Πασχάλη

Μεταπτυχιακή Εργασία Εξειδίκευσης

— ♦ —

Ιωάννινα, Φεβρουάριος 2016

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF IOANNINA

ΠΡΟΒΛΕΨΗ ΤΟΥ ΙΣΟΜΕΡΙΣΜΟΥ ΤΩΝ ΠΡΟΛΙΝΩΝ ΜΕ ΧΡΗΣΗ ΣΥΜΒΟΛΙΚΩΝ ΚΑΙ ΔΟΜΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Η
ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης
του Τμήματος Μηχανικών Η/Υ και Πληροφορικής
Εξεταστική Επιτροπή

από την

Παρασκευή Πασχάλη

ως μέρος των Υποχρεώσεων

για τη λήψη

του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Φεβρουάριος 2016

ΑΦΙΕΡΩΣΗ

Στην αδικοχαμένη μου αδερφή Βασιλική Ζ. Πασχάλη, απόφοιτη της Νομικής Σχολής Αθηνών.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον Επιβλέποντα Καθηγητή μου, κ. Αριστείδη Λύκα, για την καθοδήγηση και συνεχή ενθάρρυνση καθ' όλη τη διάρκεια εκπόνησης της μεταπτυχιακής μου εργασίας. Επίσης, θα ήθελα να ευχαριστήσω τον Επίκουρο Καθηγητή του Τμήματος Χημείας του Πανεπιστημίου Ιωαννίνων κ. Ανδρέα Τζάκο για την πολύτιμη βοήθεια που μου παρείχε. Τέλος, ευχαριστώ τους γονείς μου για την αμέριστη συμπαράσταση που μου έδειξαν όλο αυτό το διάστημα.

ΠΕΡΙΕΧΟΜΕΝΑ

Σελ

ΑΦΙΕΡΩΣΗ....iii
ΕΥΧΑΡΙΣΤΙΕΣ....iv
ΠΕΡΙΕΧΟΜΕΝΑ....v
ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ....vii
ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ....viii
ΠΕΡΙΛΗΨΗ....1
EXTENDED ABSTRACT IN ENGLISH....3
ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ....4
1.1. Οι πρωτεΐνες και η δομή τους 4
1.2. Τα δομικά συστατικά των πρωτεϊνών 5
1.3. Τα αμινοξέα συνδέονται μεταξύ τους με πεπτιδικούς δεσμούς (πρωτοταγής δομή) 8
1.4. Αναδίπλωση των πολυπεπτιδικών αλυσίδων (δευτεροταγής δομή) 15
1.5. Η μελέτη του ισομερισμού της προλίνης 20
ΚΕΦΑΛΑΙΟ 2. ΘΕΩΡΙΑ ΤΑΞΙΝΟΜΗΤΩΝ....21
2.1. Μέθοδοι Ταξινόμησης 21
2.2. Ο κανόνας του κοντινότερου γείτονα (k nearest neighbors) 26
2.3. Μηχανές διανυσμάτων υποστήριξης (support vector machines – svm) 28
2.4. Δέντρα απόφασης (decision trees) 38
2.4.1. Ο αλγόριθμος ID3 43
2.4.2. Ο αλγόριθμος C4.5 και C5.0 45
2.5. Ταξινομητής naive bayes 46
2.6. Συλλογική Ταξινόμηση (ensemble) 49
2.6.1. Μοντέλο Συλλογικών Ταξινομητών 50
2.7. Επικύρωση του σφάλματος των ταξινομητών 54
2.7.1. Η μέθοδος holdout (train set-test set) 55
2.7.2. Η μέθοδος Cross Validation (διασταυρωμένης επικύρωσης) 56
2.8. Η μέθοδος Random Subsampling 57

2.9. Απόδοση κατηγοριοποίησης (Αξιολόγηση ταξινομητών) 58
2.9.1. Ανισοκατανομή μεταξύ των κατηγοριών (class imbalance) 59
ΚΕΦΑΛΑΙΟ 3. Προβλεψη ισομερισμου προλινων 62
3.1. Πώς Ορίζεται το Πρόβλημα 62
3.2. Τρέχουσα έρευνα 63
3.3. Η δική μας Συνεισφορά 64
ΚΕΦΑΛΑΙΟ 4. Συνολο Δεδομενων και εξαγωγη χαρακτηριστικων 65
4.1. Ανάγνωση αρχείων PDB (Protein Data Bank) . 65
4.2. Παραδοχές της μεθόδου 69
4.3. Εξαγωγή χαρακτηριστικών 70
ΚΕΦΑΛΑΙΟ 5. Πειραματικα Αποτελεσματα 74
5.1. Μέθοδος Υποδειγματοληψίας (Under sampling) της Κλάσης Trans 74
5.1.1. Μέθοδος των k Κοντινότερων Γειτόνων (k nearest neighbors-kNN) 75
5.1.2. Δέντρα Απόφασης (Decision Trees-DT) 77
5.1.3. Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines-SVM) 78
5.1.4. Ταξινομητής Bayes (Naïve Bayes-NB) 79
5.2. Μέθοδος Ensemble Ταξινομητών 80
5.2.1. KNN Ensemble 81
5.2.2. Decision Tree Ensemble 81
5.2.3. Support Vector Machine Ensemble 82
5.2.4. Naïve Bayes Ensemble 83
ΚΕΦΑΛΑΙΟ 6. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ 86
ΑΝΑΦΟΡΕΣ 89
ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ 93

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

Πίνακας	Σελ
Πίνακας 1.1 Συντομογραφίες και σύμβολα για αμινοξέα	7
Πίνακας 2.1 Οι συνηθέστεροι πυρήνες εσωτερικού γινομένου σε SVM	38
Πίνακας 2.2 Μήτρα σύγχυσης	61
Πίνακας 4.1 Πίνακας διανυσμάτων 20 θέσεων στα οποία θέτουμε 1 μόνο τη θέση που αντιστοιχεί στο εκάστοτε αμινοξύ	71
Πίνακας 4.2 Διανύσματα κωδικοποίησης	72
Πίνακας 5.1 Αποτελέσματα αξιολόγησης των επιλεγμένων ταξινομητών.	80
Πίνακας 5.2 Αποτελέσματα αξιολόγησης των ensemble ταξινομητών.	85

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

Σχήμα Σελ

Σχήμα 1.1 Ευκαμψία και λειτουργία. Η πρωτεΐνη λακτοφερρίνη όταν δεσμεύει σίδηρο αλλάζει δομή, επιτρέποντας έτσι σε άλλα μόρια να ξεχωρίσουν το μόριο που έχει σίδηρο από εκείνο που δεν έχει. 5

Σχήμα 1.2 Τα ισομερή των αμινοξέων είναι D και L. Το σύμβολο R σημαίνει οποιαδήποτε πλευρική αλυσίδα. Τα D και L είναι εναντιομερή, δηλαδή αποτελούν αντίστροφες εικόνες το ένα του άλλου. 5

Σχήμα 1.3 Στις πρωτεΐνες απαντούν μόνο L-αμινοξέα. Όλα σχεδόν τα L-αμινοξέα έχουν απόλυτη διαμόρφωση S (από το λατινικό sinister που σημαίνει αριστερός). Το βέλος δείχνει την πρόοδο από τη μεγαλύτερη προς τη μικρότερη προτεραιότητα, η οποία είναι αντίθετη προς τη φορά των δεικτών του ρολογιού και επομένως το χειρόμορφο κέντρο έχει διαμόρφωση S. 6

Σχήμα 1.4 Κυκλικές δομές προλίνης. Η πλευρική αλυσίδα συνδέεται και με το άτομο του α-άνθρακα και με την αμινική ομάδα. 7

Σχήμα 1.5 Η αλληλουχία αμινοξέων διαβάζεται προς μία μόνο κατεύθυνση. Η εικόνα του πενταπεπτιδίου Tyr-Gly-Gly-Phe-Leu (YGGFL)δείχνει την αλληλουχία από το αμινο-τελικό προστο καρβοξυ-τελικό άκρο. Αυτό το πενταπεπτίδιο,η λευκινο-εγκεφαλίνη, είναι ένα ενδογενές οπιο-ειδές που τροποποιεί την αντίληψη του πόνου από τον εγκέφαλο. Το αντίθετο πενταπεπτίδιο, το Leu-Gly-Gly-Tyr (LFGGY), είναι ένα διαφορετικό μόριο χωρίς καμία λειτουργία στον εγκέφαλο. 9

Σχήμα 1.6 Τα τμήματα της πολυπεπτιδικής αλυσίδας. Η πολυπεπτιδική αλυσίδα αποτελείταιαπό έναν σταθερό κορμό (μαύρο χρώμα) και ποικιλία πλευρικών αλυσίδων (πράσινο χρώμα). 9

Σχήμα 1.7 Ο πεπτιδικός δεσμός είναι επίπεδος. Στο ζεύγος συνδεδεμένων αμινοξέων και τα έξι άτομα (Ca,C,O,N,H και C6) βρίσκονται στο ίδιο επίπεδο. Οι πλευρικές αλυσίδες έχουν πράσινο χρώμα στο σχήμα. 11

Σχήμα 1.8 Τυπικές αποστάσεις μεταξύ των πεπτιδικών ομάδων. Μια πεπτιδική ομάδα εμφανίζεται εδώ σε διαμόρφωση trans (ετερόπλευρη). 11

Σχήμα 1.9 Οι πεπτιδικοί δεσμοί trans και cis. Η μορφή trans είναι προτιμητέα διότι στη μορφή cis υπάρχουν προβλήματα χωροδιάταξης. Είναι το μέγεθος μέτρησης της ικανότητας στροφής γύρω από έναν δεσμό, που συνήθως βρίσκεται μεταξύ -180° και +180°. Οι δίεδρες γωνίες μερικές φορές λέγονται και γωνίες στρέψης. 12

Σχήμα 1.10 Trans και Cis X-Pro. Στην περίπτωση αυτή, η ενέργεια είναι περίπου ίδια διότι υπάρχουν αντίστοιχα προβλήματα χωροδιάταξης και για τις δύο μορφές. 13

Σχήμα 1.11 Η περιστροφή γύρω από τους δεσμούς ενός πολυπεπτιδίου. Η δομή κάθε αμινοξέος σε ένα πολυπεπτίδιο μπορεί να ρυθμιστεί από την περιστροφή γύρω από δύο απλούς δεσμούς. (Α) Η γωνία περιστροφής γύρω από τον δεσμό μεταξύ των ατόμων αζώτου και α-άνθρακα ονομάζεται φ, ενώ η γωνία περιστροφής γύρω από τον δεσμό μεταξύ του ατό-μου α-άνθρακα και των ανθράκων της καρβονυλικής ομάδας ονομάζεται ψ. (Β) Μια κάτοψη του δεσμού μεταξύ αζώτου και ατόμου α-άνθρακα δείχνει πώς μετράμε τη φ. (Γ) Μια κάτοψη του δεσμού μεταξύ του ατόμου α-άνθρακα και του άνθρακα της καρβονυλικής ομάδας, δείχνει πώς μετράμε την ψ. 13

Σχήμα 1.12 Ένα διάγραμμα Ramachandran που δείχνει τις τιμές των φ και ψ. Οι τιμές φ και ψ είναι περιορισμένες λόγω των συγκρούσεων μεταξύ ατόμων. Οι επιτρεπτές τιμές φ και ψ φαίνονται με σκούρο πράσινο, ενώ οι οριακές τιμές φαίνονται με ανοιχτό πράσινο. Η δομή που φαίνεται δεξιά είναι εκείνη που δημιουργεί τις λιγότερες συγκρούσεις μεταξύ των ατόμων. 14

Σχήμα 1.13 Η δομή μιας α-έλικας. (Α) Η απεικόνιση κορδέλας στην οποία ξεχωρίζουν τα άτομα άνθρακα και οι πλευρικές αλυσίδες. (Β) Μια πλάγια όψη του μοντέλου με σφαίρες και ράβδους όπου διακρίνονται οι δεσμοί υδρογόνου (διακεκομμένες γραμμές) μεταξύ των ομάδων NH

και CO. (Γ) Παρατηρώντας από το άκρο της έλικας και παράλληλα προς τον άξονα βλέπουμε τον περιελιγμένο κορμό να σχηματίζει το εσωτερικό της έλικας και τις πλευρικές αλυσίδες να προεξέχουν προς τα έξω. (Δ) Ένα χωροπληρωτικό μοντέλο του (Γ) δείχνει πόσο λίγος κενός χώρος μένει στο κέντρο της έλικας. 15

Σχήμα 1.14 Σχηματική απεικόνιση μιας *α*-έλικας. (Α) Μοντέλο με σφαίρες και ράβδους. (Β) Απεικόνιση κορδέλας. (Γ) Απεικόνιση κυλίνδρου. 17

Σχήμα 1.15 Δομή μιας μεικτής β-επιφάνειας 18

Σχήμα 1.16 Μια β-επιφάνεια όπου η κάθε πτύχωση είναι ελαφρώς στριμμένη σε σχέση με την προηγούμενη. (Α) Μοντέλο με σφαίρες και ράβδους. (Β) Σχηματικό μοντέλο. (Γ) Σχηματική διαμόρφωση που έχει στραφεί κατά 90° για να εμφανίσει καλύτερα το στρίψιμο. 19

Σχήμα 1.17 Οι θηλιές στην επιφάνεια μιας πρωτεΐνης. Ένα τμήμα του μορίου του αντισώματος έχει θηλιές στην επιφάνειά του (κόκκινο χρώμα) που αλληλεπιδρούν με άλλα μόρια. 20

Σχήμα 2.1 Το πρόβλημα της κατηγοριοποίησης. (1^{ο} διάγραμμα: ορισμός κατηγοριών, 2^{ο} διάγραμμα: σύνολο παραδειγμάτων προς κατηγοριοποίηση, 3^{ο} διάγραμμα κατηγοριοποιημένο σύνολο παραδειγμάτων) 24

Σχήμα 2.2 Κατηγοριοποίηση με χρήση KNN 27

Σχήμα 2.3 Παράδειγμα ταξινόμησης 2D 29

Σχήμα 2.4 Απεικόνιση περιθωρίου (margin) και διανυσμάτων στήριξης (support vector) 31

Σχήμα 2.5 Μη γραμμικά διαχωρίσιμα δεδομένα 34

Σχήμα 2.6 Μετασχηματισμός του χώρου εισόδου σε χώρο χαρακτηριστικών 36

Σχήμα 2.7 Ένα απλό Decision Tree 41

Σχήμα 2.8 Τυπική απεικόνιση δημιουργίας ενός δέντρου απόφασης. 43

Σχήμα 2.9 Σχηματική αναπαράσταση της μεθόδου bagging 52

Σχήμα 2.10 Σχηματική αναπαράσταση της μεθόδου AdaBoost 53

Σχήμα 2.11 Training- Test Set 55

Σχήμα 2.12 K-Fold Cross-validation 57

Σχήμα 2.13 Μέθοδος Random Subsampling 52

Σχήμα 4.1 Δομή ενός αρχείου PDB. 66

Σχήμα 4.2 Παραμετροποίηση του Pisces server για την εξαγωγή του συνόλου δεδομένων (πρωτεϊνών) για τη μελέτη. 68
Σχήμα 4.3 Πίνακας με την απόλυτη συχνότητα κάθε αμινοξέος στη σφαίρα κάθε προλίνης. 70
Σχήμα 5.1 Μετρική kloss του 10-fold cross validation για τις διαφορετικές τιμές του k. Τα διαφορετικά χρώματα απεικονίζουν τις διαφορετικές επαναλήψεις της διαδικασίας. 75
Σχήμα 5.2 Μέσο kloss του 10 fold cross validation για τις διαφορετικές τιμές του k. 76
Σχήμα 5.3 Μέσο kloss του 10-fold cross validation για τις διαφορετικές τιμές του MaxNumSplits. Τα διαφορετικά χρώματα απεικονίζουν τις διαφορετικές τιμές του MinNodeSize. 77
Σχήμα 5.4 Μετρική kloss του 10-fold cross validation για τις διαφορετικές συναρτήσεις πυρήνα για τις 10 διαφορετικές επαναλήψεις της διαδικασίας. 78
Σχήμα 5.5 Οι μετρικές sensitivity και specificity για τα ensemble του kNN. 81
Σχήμα 5.6 Οι μετρικές sensitivity και specificity για τα ensemble των δέντρων απόφασης. 82
Σχήμα 5.7 Οι μετρικές sensitivity και specificity για τα ensemble των μηχανών διανυσμάτων υποστήριξης. 83
Σχήμα 5.8 Οι μετρικές sensitivity και specificity για τα ensemble των ταξινομητών Bayes. 84

ΠΕΡΙΛΗΨΗ

Παρασκευή Ζ. Πασχάλη. ΜΔΕ, Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Φεβρουάριος, 2016. Πρόβλεψη του ισομερισμού των προλινών με χρήση συμβολικών και δομικών χαρακτηριστικών. Επιβλέπων: Αριστείδης Λύκας.

Η πλειονότητα των πεπτιδικών δεσμών σε πρωτεΐνες βρίσκονται σε διαμόρφωση trans. Ωστόσο, για τα κατάλοιπα προλίνης, ένα σημαντικό ποσοστό των προλυλο-πεπτιδικών δεσμών υιοθετεί την cis διαμόρφωση. Ο ισομερισμός της προλίνης σε cis ή trans είναι γνωστό ότι παίζει σημαντικό ρόλο στην αναδίπλωση των πρωτεϊνών, στην κυτταρική επικοινωνία και στη διαμεμβρανική ενεργή μεταφορά. Η ακριβής πρόβλεψη του ισομερισμού της προλίνης σε cis ή trans έχει πολλές σημαντικές εφαρμογές στην κατανόηση της δομής και της λειτουργίας των πρωτεϊνών.

Σε αυτή την εργασία, προτείνουμε μια νέα προσέγγιση για την πρόβλεψη cis-trans ισομερισμού των προλινών, χρησιμοποιώντας διάφορες μεθοδολογίες κατασκευής ταξινομητών από δεδομένα (παραδείγματα). Χρησιμοποιήσαμε πληροφορίες αλληλουχίας, πληροφορίες για το είδος των γειτονικών μορίων ως χαρακτηριστικά μιας προλίνης και τη δευτερεύουσα πληροφορία δομής που προβλέπεται από το DSSP. Η εκπαίδευση και ο έλεγχος των ταξινομητών πραγματοποιήθηκε σε ένα σύνολο δεδομένων 1900 μη ομόλογων πρωτεϊνών χρησιμοποιώντας 10-fold cross validation. Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) και τα Δέντρα Απόφασης παρουσιάζουν τις καλύτερες επιδόσεις για τον προσδιορισμό του ισομερισμού της προλίνη σε cis-trans ισομερισμό με βάση τα επιλεγμένα χαρακτηριστικά. Διαπιστώθηκε ότι η χρήση πολλαπλών ταξινομητών με τη μέθοδο bagging (ensemble bagging) βελτιώνει σημαντικά την απόδοση της πρόβλεψης, και συγκεκριμένα η μετρική Fmeasure (f1) αυξήθηκε από 76% σε 82% και τα ποσοστά της ειδικότητας και της ευαισθησίας εξισορροπήθηκαν εφόσον με αυτόν τον τρόπο αντιμετωπίστηκε το πρόβλημα του μη ισορροπημένου συνόλου δεδομένων (5,3 % cis και 94,7 % trans ισομερισμού).

Η επιτυχής εφαρμογή της προσέγγισης πολλαπλών ταξινομητών στη μελέτη αυτή ενίσχυσε ότι η μάθηση με πολλαπλούς ταξινομητές είναι ένα ισχυρό εργαλείο για την πρόβλεψη του ισομερισμού της προλίνης σε πρωτεΐνες, αλλά και γενικότερα για την κατασκευή ταξινομητών από ανισομερή ως προς τις κατηγορίες σύνολα εκπαίδευσης.

EXTENDED ABSTRACT IN ENGLISH

Paraskevi Z. Paschali, MSc, Department of Computer Science & Engineering, University of Ioannina, Greece. February, 2016. Prediction of proline isomerization by using symbolic and structural characteristics. Supervisor: Aristidis Likas.

The majority of peptide bonds in proteins are found to occur in trans conformation. However, for proline residues, a considerable fraction of Prolyl peptide bonds adopt the cis conformation. Proline cis-trans isomerization is known to play a significant role in protein folding and splicing, cell signaling and trans-membrane active transport. Accurate prediction of proline cis-trans isomerization in proteins would have many important applications towards the understanding of protein structure and function.

In this thesis, we propose a new approach to predict the proline cis-trans isomerization in proteins using several types of classifiers. The experimental results indicated that using ensemble of classifiers could lead to better prediction performance than using single classifiers. We used single sequence information, amino acid compositions of different local sequences and the secondary structure information predicted by DSSP. We explored these different sequence encoding schemes in order to investigate their effects on the prediction performance. The training and testing of this approach was performed on a dataset of 1900 non-homologous proteins (5,3 % cis and 94,7 % trans isomerization). SVMs and Decision Trees provided the best performance for determining the proline cis-trans isomerization based on the selected features. It was found that using multiple classifiers in the form of ensemble classification (bagging) significantly improve the prediction performance, the prediction f1 measure increased from 76% to 82% and balance specificity and sensitivity tackling with the problem of imbalanced dataset.

The successful application of Ensemble approach in this study reinforced that ensemble learning is a powerful tool in predicting proline *cis-trans* isomerization in proteins and in the case of imbalanced data sets in general.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

1.1 Οι πρωτεΐνες και η δομή τους

1.2 Τα δομικά συστατικά των πρωτεϊνών

1.3 Τα αμινοξέα συνδέονται μεταξύ τους με πεπτιδικούς δεσμούς (πρωτοταγής δομή)

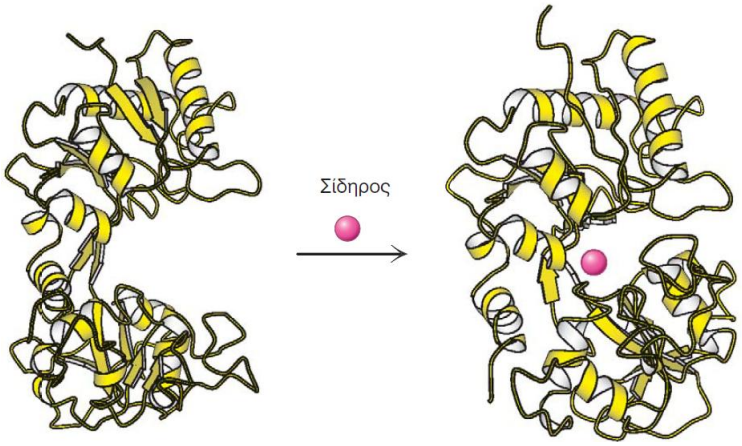
1.4 Αναδίπλωση των πολυπεπτιδικών αλυσίδων (δευτεροταγής δομή)

1.5 Η μελέτη του ισομερισμού της προλίνης

1.1. Οι πρωτεΐνες και η δομή τους

Οι πρωτεΐνες είναι μακρομόρια των ζωντανών οργανισμών και εξυπηρετούν βασικές λειτουργίες σε όλες σχεδόν τις βιολογικές διεργασίες. Λειτουργούν ως καταλύτες, μεταφορείς και αποθηκευτές άλλων μορίων, όπως το οξυγόνο, παρέχουν μηχανική στήριξη και ανοσοπροστασία, δημιουργούν κίνηση, διαβιβάζουν νευρικές ώσεις και ρυθμίζουν την ανάπτυξη και τη διαφοροποίηση [2]. Παρακάτω παρατίθενται βασικές ιδιότητες που επιτρέπουν στις πρωτεΐνες να συμμετέχουν σε ένα μεγάλο φάσμα λειτουργιών, σύμφωνα με το [2].

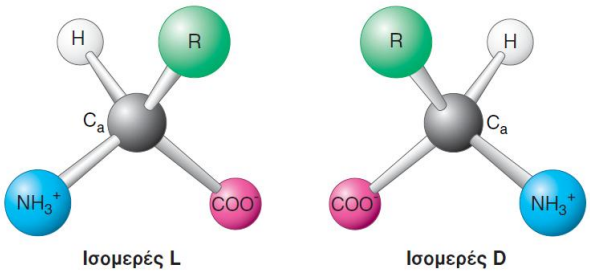
- Οι πρωτεΐνες είναι γραμμικά πολυμερή δομούμενα από μονομερή αμινοξέων.
- Οι πρωτεΐνες περιέχουν μια μεγάλη σειρά λειτουργικών ομάδων.
- Οι πρωτεΐνες μπορούν να αλληλεπιδράσουν μεταξύ τους και με άλλα βιολογικά μακρομόρια, για να δημιουργήσουν πολύπλοκα συσσωματώματα.
- Μερικές πρωτεΐνες είναι σχεδόν άκαμπτες, ενώ υπάρχουν άλλες που εμφανίζουν μια σχετική ευκαμψία (βλέπε Σχ. 1.1).



Σχήμα 1.1 Ευκαμψία και λειτουργία [2]. Η πρωτεΐνη λακτοφερρίνη όταν δεσμεύει σίδηρο αλλάζει δομή, επιτρέποντας έτσι σε άλλα μόρια να ξεχωρίσουν το μόριο που έχει σίδηρο από εκείνο που δεν έχει.

1.2. Τα δομικά συστατικά των πρωτεϊνών

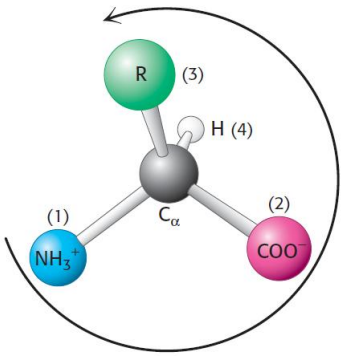
Τα αμινοξέα είναι οι δομικές μονάδες των πρωτεϊνών [5][32][38]. Ένα α-αμινοξύ αποτελείται από ένα κεντρικό άτομο άνθρακα, που λέγεται α-άνθρακας, συνδεδεμένο με μια αμινική ομάδα, μια καρβοξυλική ομάδα, ένα άτομο υδρογόνου και μια χαρακτηριστική ομάδα R. Η ομάδα R πολλές φορές ονομάζεται και πλευρική αλυσίδα. Έχοντας τέσσερις διαφορετικές ομάδες συνδεδεμένες στο τετράεδρο του ατόμου του α-άνθρακα, τα α-αμινοξέα είναι χειρόμορφα (chiral). Τα δύο κατοπτρικά είδωλα ονομάζονται L- και D-ισομερές (βλέπε Σχ. 1.2).



Σχήμα 1.2 Τα ισομερή των αμινοξέων είναι D και L [2]. Το σύμβολο R σημαίνει οποιαδήποτε πλευρική αλυσίδα. Τα D και L είναι εναντιομερή, δηλαδή αποτελούν αντίστροφες εικόνες το ένα του άλλου.

Μόνο L-αμινοξέα απαντούν στις πρωτεΐνες. Σε όλα σχεδόν τα αμινοξέα, το L-ισομερές έχει διαμόρφωση S και όχι R (βλέπε Σχ. 1.3). Αν και έχουν γίνει πολλές προσπάθειες να εξηγηθεί γιατί τα αμινοξέα των πρωτεϊνών έχουν αυτήν την απόλυτη διαμόρφωση, δεν υπάρχει ακόμη ικανοποιητική εξήγηση. Ίσως η επιλογή του L-ισομερούς σε σχέση με το D να έγινε τυχαία αλλά νωρίς στην εξέλιξη και στη συνέχεια, αφού έγινε, παρέμεινε σταθερή.

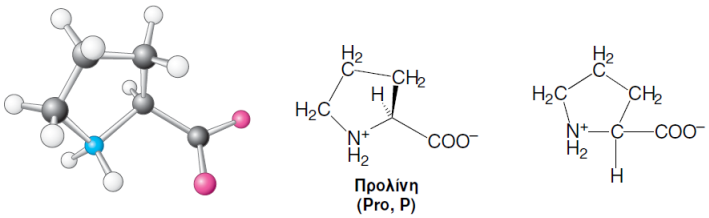
Υπάρχουν είκοσι είδη πλευρικών αλυσίδων στις πρωτεΐνες, που διαφέρουν μεταξύ τους ως προς το μέγεθος, το σχήμα, το φορτίο, τη δεσμευτική συγγένεια υδρογόνου, την υδροφοβικότητα και τη χημική αντιδραστικότητα [5][32][38]. Όλες οι πρωτεΐνες, σε όλα τα είδη βακτηριακές, αρχαϊκές και ευκαρυωτικές είναι δομημένες από τα ίδια 20 αμινοξέα. Αυτό το βασικό αλφάβητο των πρωτεϊνών δημιουργήθηκε πριν από αρκετά δισεκατομμύρια χρόνια. Η ποικιλία πρωτεϊνικών λειτουργιών είναι αποτέλεσμα της ποικιλότητας και ποικιλομορφίας αυτών των 20 δομικών στοιχείων. Η κατανόηση της χρήσης του αλφαβήτου στη δημιουργία των πολύπλοκων τριδιάστατων δομών που επιτρέπουν στις πρωτεΐνες να συμμετέχουν σε τόσες βιολογικές διεργασίες είναι ένα συναρπαστικό αντικείμενο της βιοχημείας.



Σχήμα 1.3 Στις πρωτεΐνες απαντούν μόνο L-αμινοξέα [2]. Όλα σχεδόν τα L-αμινοξέα έχουν απόλυτη διαμόρφωση S (από το λατινικό sinister που σημαίνει αριστερός). Το βέλος δείχνει την πρόοδο από τη μεγαλύτερη προς τη μικρότερη προτεραιότητα, η οποία είναι αντίθετη προς τη φορά των δεικτών του ρολογιού και επομένως το χειρόμορφο κέντρο έχει διαμόρφωση S.

Σε αυτήν την εργασία θα ασχοληθούμε με ένα συγκεκριμένο αμινοξύ που ονομάζεται προλίνη. Η προλίνη αποτελείται από μια αλειφατική πλευρική αλυσίδα, αλλά διαφέρει όμως από τα άλλα 20 αμινοξέα διότι η πλευρική αλυσίδα της συνδέεται και με το άτομο

του αζώτου και με το άτομο του α-άνθρακα (βλέπε Σχ. 1.4). Η προλίνη μπορεί να επηρεάσει ιδιαίτερα την πρωτεϊνική αρχιτεκτονική διότι ο δακτύλιος της δομής της την κάνει πιο άκαμπτη από ότι τα άλλα αμινοξέα.



Σχήμα 1.4 Κυκλικές δομές προλίνης [2]. Η πλευρική αλυσίδα συνδέεται και με το άτομο του α-άνθρακα και με την αμινική ομάδα.

Τα αμινοξέα συχνά χαρακτηρίζονται με ένα ή με τρία γράμματα (Πίνακας 1.1) [2]. Οι συντομογραφίες των τριών γραμμάτων χρησιμοποιούν τα τρία πρώτα γράμματα του ονόματός τους στην αγγλική, εκτός από την ασπαραγίνη (Asn), τη γλουταμίνη (Gln), την ισολευκίνη (Ile) και τη θρυπτοφάνη (Trp). Τα σύμβολα ενός γράμματος για πολλά αμινοξέα αντιστοιχούν στο πρώτο γράμμα του ονόματός τους (π.χ. G για τη γλυκίνη, L για τη λευκίνη). Τα υπόλοιπα γράμματα συμφωνήθηκαν με σύμβαση. Τα σύμβολα αυτά και οι συντομογραφίες είναι βασικά στοιχεία του λεξιλογίου της βιοχημείας [2].

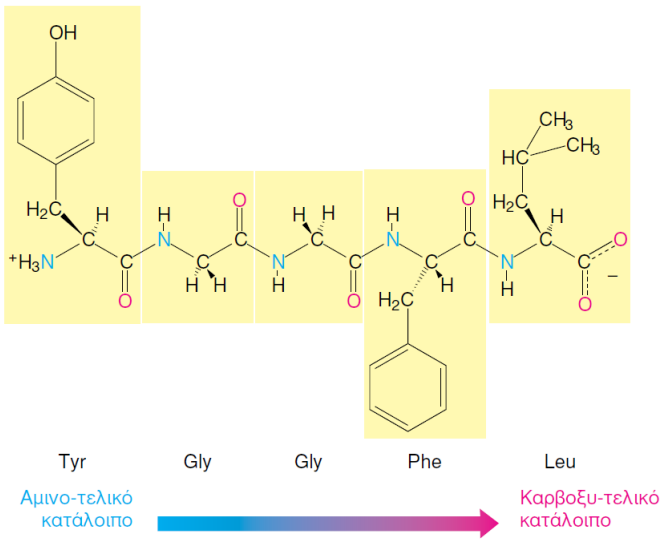
Πίνακας 1.1 Συντομογραφίες και σύμβολα για αμινοξέα [2]

Αμινοξύ	*Συντομογραφία τριών γραμμάτων*	*Συντομογραφία ενός γράμματος*	*Αμινοξύ*	*Συντομογραφία τριών γραμμάτων*	*Συντομογραφία ενός γράμματος*
Αλανίνη	Ala	A	Κυστεΐνη	Cys	C
Αργινίνη	Arg	R	Λευκίνη	Leu	L
Ασπαραγίνη	Asn	N	Λυσίνη	Lys	K
Ασπαραγινικό οξύ	Asp	D	Μεθειονίνη	Met	M
Βαλίνη	Val	V	Προλίνη	Pro	P
Γλουταμίνη	Gln	Q	Σερίνη	Ser	S
Γλουταμινικό οξύ	Glu	E	Τυροσίνη	Tyr	Y
Γλυκίνη	Gly	G	Φαινυλαλανίνη	Phe	F
θρεονίνη	Thr	T	Ασπαραγίνη ή ασπαραγινικό οξύ	Asx	B
θρυπτοφάνη	Trp	W			
Ισολευκίνη	Ile	I	Γλουταμίνη ή γλουταμινικό οξύ	Glx	Z
Ιστιδίνη	His	H			

1.3. Τα αμινοξέα συνδέονται μεταξύ τους με πεπτιδικούς δεσμούς (πρωτοταγής δομή)

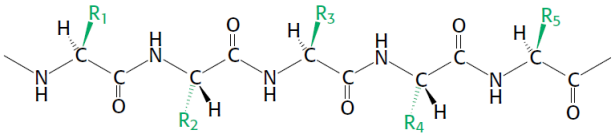
Οι πρωτεΐνες είναι γραμμικά πολυμερή που δημιουργούνται δεσμεύοντας την α-καρβοξυλική ομάδα ενός αμινοξέος στην α-αμινική ομάδα ενός άλλου αμινοξέος με έναν πεπτιδικό δεσμό (που λέγεται και αμιδικός δεσμός) [7][27]. Η δημιουργία ενός διπεπτιδίου από δύο αμινοξέα συνοδεύεται από την απώλεια ενός μορίου ύδατος. Η ισορροπία της αντίδρασης βρίσκεται μάλλον προς την πλευρά της υδρόλυσης παρά της σύνθεσης. Επομένως, η βιοσύνθεση του πεπτιδικού δεσμού χρειάζεται την προσθήκη ελεύθερης ενέργειας. Παρόλα αυτά οι πεπτιδικοί δεσμοί είναι αρκετά σταθεροί κινητικά, η διάρκεια ζωής ενός πεπτιδικού δεσμού σε υδάτινο διάλυμα, όταν δεν υπάρχει καταλύτης, πλησιάζει τα 1000 χρόνια. Στην εργασία μας, εκμεταλλευόμαστε την πρωτοταγή δομή των πρωτεϊνών με σκοπό να εξάγουμε συγκεκριμένα αμινοξέα πριν και μετά από κάθε προλίνη που περιέχεται σε μια πολυπεπτιδική αλυσίδα.

Μια σειρά αμινοξέων που ενώνονται με πεπτιδικούς δεσμούς δημιουργούν μια πολυπεπτιδική αλυσίδα, και κάθε αμινοξύ στο πολυπεπτίδιο ονομάζεται κατάλοιπο. Μια πολυπεπτιδική αλυσίδα έχει πολικότητα διότι τα δύο άκρα της είναι διαφορετικά: μια α-αμινική ομάδα στο ένα άκρο, μια α-καρβοξυλική ομάδα στο άλλο άκρο. Συμβατικά, έχουμε δεχθεί ότι το αμινο-τελικό άκρο θεωρείται η αρχή της πολυπεπτιδικής αλυσίδας και επομένως η αλληλουχία των αμινοξέων σε μια πολυπεπτιδική αλυσίδα γράφεται αρχίζοντας με το αμινο-τελικό κατάλοιπο. Για παράδειγμα, στο πενταπεπτίδιο Tyr-Gly-Gly-Phe-Leu (YGGFL),η τυροσίνη είναι το αμινο-τελικό κατάλοιπο και η λευκίνη το καρβοξυ-τελικό κατάλοιπο (βλέπε Σχ. 1.5). Το Leu-Phe-Gly-Gly-Tyr (CFGGY) είναι ένα άλλο πενταπεπτίδιο, με διαφορετικές χημικές ιδιότητες.



Σχήμα 1.5 Η αλληλουχία αμινοξέων διαβάζεται προς μία μόνο κατεύθυνση [2]. Η εικόνα του πενταπεπτιδίου Tyr-Gly-Gly-Phe-Leu (YGGFL)δείχνει την αλληλουχία από το αμινο-τελικό προς το καρβοξυ-τελικό άκρο. Αυτό το πενταπεπτίδιο,η λευκινο-εγκεφαλίνη, είναι ένα ενδογενές οπιο-ειδές που τροποποιεί την αντίληψη του πόνου από τον εγκέφαλο. Το αντίθετο πενταπεπτίδιο, το Leu-Gly-Gly-Tyr (LFGGY), είναι ένα διαφορετικό μόριο χωρίς καμία λειτουργία στον εγκέφαλο.

Μια πολυπεπτιδική αλυσίδα αποτελείται από ένα σταθερά επαναλαμβανόμενο τμήμα, που ονομάζεται κύρια αλυσίδα ή κορμός, και ένα μεταβλητό τμήμα, που αποτελείται από τις διαφορετικές πλευρικές αλυσίδες (βλέπε Σχ. 1.6). Ο κορμός του πολυπεπτιδίου έχει πολλές δυνατότητες δημιουργίας δεσμών υδρογόνου. Κάθε κατάλοιπο έχει μια καρβονυλική ομάδα, που είναι καλός δέκτης δεσμών υδρογόνου και μια αμιδική ομάδα (εκτός από την προλίνη), που είναι καλός δότης δεσμών υδρογόνου. Αυτές οι ομάδες αλληλεπιδρούν μεταξύ τους και με τις λειτουργικές ομάδες των πλευρικών αλυσίδων και έτσι σταθεροποιούνται συγκεκριμένες δομές για κάθε πολυπεπτιδική αλυσίδα, όπως θα αναλύσουμε στην ενότητα 1.5.



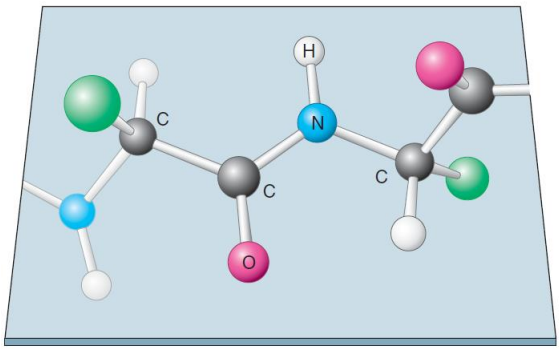
Σχήμα 1.6 Τα τμήματα της πολυπεπτιδικής αλυσίδας [2]. Η πολυπεπτιδική αλυσίδα αποτελείταιαπό έναν σταθερό κορμό (μαύρο χρώμα) και ποικιλία πλευρικών αλυσίδων (πράσινο χρώμα).

Οι περισσότερες φυσικές πολυπεπτιδικές αλυσίδες περιέχουν από 50 έως 2000 κατάλοιπα αμινοξέων και συνήθως ονομάζονται πρωτεΐνες. Τα πεπτίδια που έχουν μικρότερο αριθμό αμινοξέων ονομάζονται ολιγοπεπτίδια ή απλώς πεπτίδια. Η γνώση των αλληλουχιών αμινοξέων είναι σημαντική για αρκετούς λόγους.

Πρώτον, η γνώση της αλληλουχίας μιας πρωτεΐνης είναι συνήθως απαραίτητη για την κατανόηση του μηχανισμού δράσης της (π.χ. στην περίπτωση της καταλυτικής δράσης των ενζύμων). Ακόμη, τροποποιώντας την αλληλουχία γνωστών πρωτεϊνών μπορούμε να δημιουργήσουμε πρωτεΐνες με νέες ιδιότητες.
Δεύτερον, η γνώση της αλληλουχίας των αμινοξέων καθορίζει την τριδιάστατη δομή των πρωτεϊνών, πράγμα που είναι πολύ χρήσιμο στην εργασία μας. Η αλληλουχία των αμινοξέων είναι ο συνδετικός κρίκος μεταξύ της γενετικής πληροφορίας του DNA και της τριδιάστατης δομής που καθορίζει τη βιολογική λειτουργία μιας πρωτεΐνης.

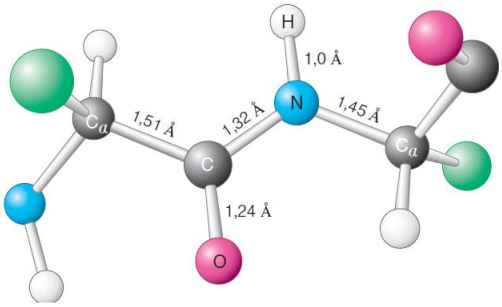
- Η στερεοδιάταξη των πεπτιδικών αλυσίδων

Η εξέταση της γεωμετρίας του πρωτεϊνικού κορμού αποκαλύπτει αρκετά σημαντικά στοιχεία. Πρώτον, ο πεπτιδικός δεσμός είναι βασικά επίπεδος (βλέπε Σχ. 1.7) [2]. Έτσι λοιπόν, για κάθε ζεύγος αμινοξέων τα οποία συνδέονται με πεπτιδικό δεσμό υπάρχουν έξι άτομα που βρίσκονται στο ίδιο επίπεδο: το άτομο α-άνθρακα και η ομάδα CO του πρώτου αμινοξέος καθώς και η ομάδα NH και το άτομο α-άνθρακα του δεύτερου αμινοξέος. Η εξήγηση αυτής της γεωμετρικής προτίμησης βρίσκεται στη φύση του χημικού δεσμού του πεπτιδίου. Ο πεπτιδικός δεσμός έχει, εν μέρει, χαρακτήρα διπλού δεσμού ο οποίος αποτρέπει την περιστροφή γύρω από τον εαυτό του.



Σχήμα 1.7 Ο πεπτιδικός δεσμός είναι επίπεδος [2]. Στο ζεύγος συνδεδεμένων αμινοξέων και τα έξι άτομα (Ca,C,O,N,H και C6) βρίσκονται στο ίδιο επίπεδο. Οι πλευρικές αλυσίδες έχουν πράσινο χρώμα στο σχήμα.

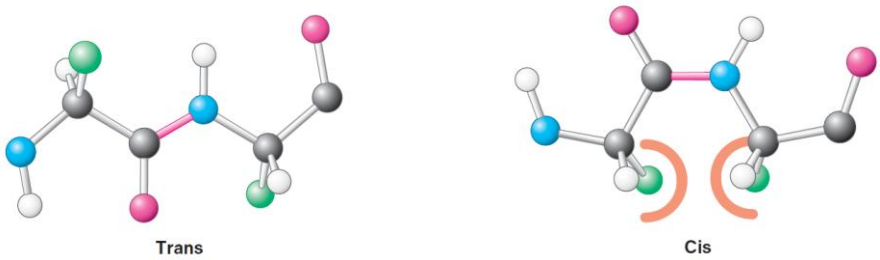
Το γεγονός ότι ο δεσμός δεν περιστρέφεται περιορίζει τις στερεοδιατάξεις του πεπτιδικού κορμού και εξηγεί την επίπεδη φύση του δεσμού [7]. Αυτός ο χαρακτήρας διπλού δεσμού εκφράζεται και στην απόσταση μεταξύ των ομάδων CO και NH. Η απόσταση C-N σε έναν πεπτιδικό δεσμό είναι 1,32A, δηλαδή μια τιμή που βρίσκεται μεταξύ των αναμενόμενων για έναν απλό C-N δεσμό (1,49 A) και έναν διπλό δεσμό C=N (1,27 A), όπως φαίνεται στο Σχ. 1.8. Τέλος, ο πεπτιδικός δεσμός δεν έχει φορτίο, επιτρέποντας έτσι στα πολυμερή των αμινοξέων που συνδέονται με πεπτιδικούς δεσμούς να δημιουργήσουν σφαιρικές κατασκευές χωρίς ενδιάμεσα κενά.



Σχήμα 1.8 Τυπικές αποστάσεις μεταξύ των πεπτιδικών ομάδων. Μια πεπτιδική ομάδα εμφανίζεται εδώ σε διαμόρφωση trans (ετερόπλευρη).

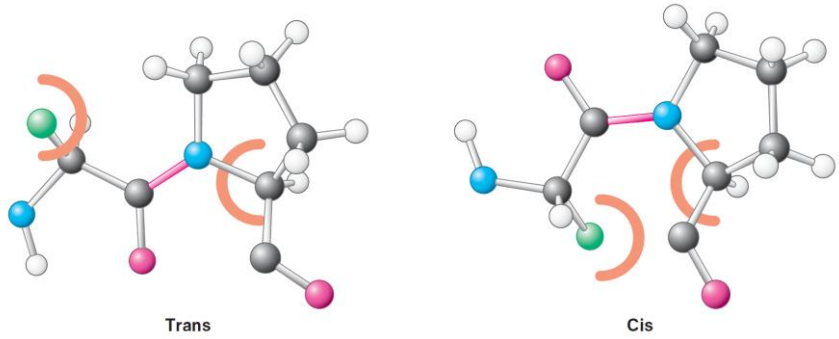
Υπάρχουν δύο δυνατές διαμορφώσεις για έναν επίπεδο πεπτιδικό δεσμό [7][27]. Στη διαμόρφωση trans τα δύο άτομα α-άνθρακα βρίσκονται απέναντι, ενώ στη διαμόρφωση cis

βρίσκονται στην ίδια πλευρά του πεπτιδικού δεσμού. Όλοι σχεδόν οι πεπτιδικοί δεσμοί των πρωτεϊνών είναι trans. Αυτή η προτίμηση της διαμόρφωσης trans σε σχέση με τη διαμόρφωση cis εξηγείται από το γεγονός ότι οι στερικές αλληλοεπικαλύψεις των ομάδων που συνδέονται στα άτομα α-άνθρακα παρεμποδίζουν τη διαμόρφωση cis αλλά αφήνουν ελεύθερη την trans (βλέπε Σχ.1. 9). Οι πιο κοινές περιπτώσεις δεσμών cis αφορούν το αμινοξύ X και την προλίνη (X-Pro). Οι δεσμοί αυτοί της προλίνης έχουν διαμόρφωση cis αντί trans διότι το άζωτο της προλίνης είναι δεσμευμένο σε δύο τετρασθενή άτομα άνθρακα, γεγονός που περιορίζει ουσιαστικά τις στερικές διαφοροποιήσεις μεταξύ μορφών trans και cis (βλέπε Σχ. 1.10).

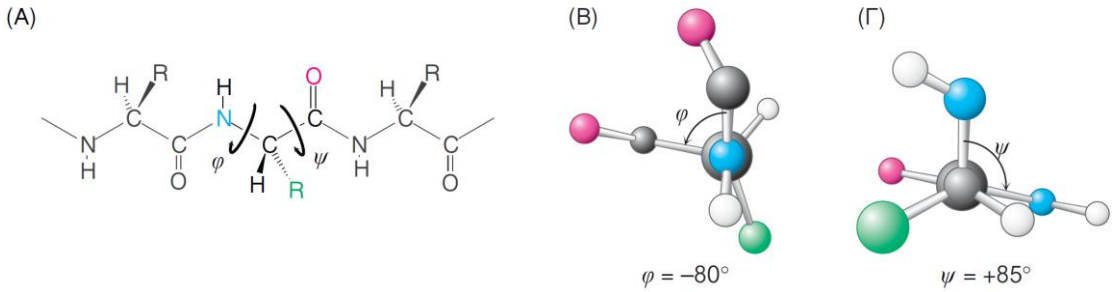


Σχήμα 1.9 Οι πεπτιδικοί δεσμοί trans και cis [2]. Η μορφή trans είναι προτιμητέα διότι στη μορφή cis υπάρχουν προβλήματα χωροδιάταξης. Είναι το μέγεθος μέτρησης της ικανότητας στροφής γύρω από έναν δεσμό, που συνήθως βρίσκεται μεταξύ -180° και +180°. Οι δίεδρες γωνίες μερικές φορές λέγονται και γωνίες στρέψης.

Σε αντίθεση με τον πεπτιδικό δεσμό, οι δεσμοί που ενώνουν τις αμινικές ομάδες με το άτομο α-άνθρακα και τις καρβονυλικές ομάδες με το άτομο α-άνθρακα είναι καθαροί απλοί δεσμοί. Τα δύο γειτονικά άκαμπτα πεπτιδικά επίπεδα μπορούν επομένως να περιστραφούν γύρω από τους δεσμούς αυτούς αποκτώντας διάφορους προσανατολισμούς. Η ελευθερία περιστροφής γύρω από τους δύο δεσμούς κάθε αμινοξέος επιτρέπει στις πρωτεΐνες να αναδιπλωθούν με πολλούς και διάφορους τρόπους. Οι περιστροφές γύρω από τους δεσμούς είναι δυνατόν να προσδιοριστούν από τις δίεδρες γωνίες (βλέπε Σχ. 1.11). Η γωνία περιστροφής γύρω από τον δεσμό μεταξύ των ατόμων αζώτου και α-άνθρακα ονομάζεται φ. Η γωνία περιστροφής γύρω από τον δεσμό ατόμου α-άνθρακα και της καρβονυλικής ομάδας ονομάζεται ψ. Για οποιονδήποτε από τους δεσμούς, η περιστροφή κατά τη φορά των δεικτών του ρολογιού (κοιτώντας από πίσω προς τα εμπρός) αντιστοιχεί σε θετική τιμή. Οι γωνίες φ και ψ καθορίζουν την κατεύθυνση της πολυπεπτιδικής αλυσίδας.

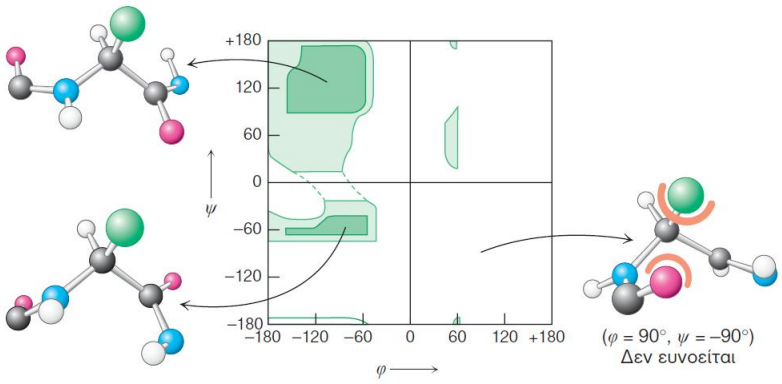


Σχήμα 1.10 Trans και Cis X-Pro [2]. Στην περίπτωση αυτή, η ενέργεια είναι περίπου ίδια διότι υπάρχουν αντίστοιχα προβλήματα χωροδιάταξης και για τις δύο μορφές.



Σχήμα 1.11 Η περιστροφή γύρω από τους δεσμούς ενός πολυπεπτιδίου [2]. Η δομή κάθε αμινοξέος σε ένα πολυπεπτίδιο μπορεί να ρυθμιστεί από την περιστροφή γύρω από δύο απλούς δεσμούς. (Α) Η γωνία περιστροφής γύρω από τον δεσμό μεταξύ των ατόμων αζώτου και α-άνθρακα ονομάζεται φ, ενώ η γωνία περιστροφής γύρω από τον δεσμό μεταξύ του ατό-μου α-άνθρακα και των ανθράκων της καρβονυλικής ομάδας ονομάζεται ψ. (Β) Μια κάτοψη του δεσμού μεταξύ αζώτου και ατόμου α-άνθρακα δείχνει πώς μετράμε τη φ. (Γ) Μια κάτοψη του δεσμού μεταξύ του ατόμου α-άνθρακα και του άνθρακα της καρβονυλικής ομάδας, δείχνει πώς μετράμε την ψ.

Ποιοι συνδυασμοί είναι δυνατοί για τις γωνίες φ και ψ; Ο G.N. Ramachandran διαπίστωσε ότι πολλοί συνδυασμοί δεν είναι δυνατοί λόγω των στερικών συγκρούσεων μεταξύ των ατόμων. Οι τιμές που επιτρέπονται μπορούν να τοποθετηθούν σε ένα διδιάστατο σχήμα που ονομάζεται διάγραμμα Ramachandran (βλέπε Σχ. 1.12). Τα τρία τέταρτα των πιθανών συνδυασμών των γωνιών φ και ψ είναι αδύνατον να πραγματοποιηθούν, διότι δημιουργούν τοπικές στερικές συγκρούσεις. Ο στερικός αποκλεισμός, δηλαδή το γεγονός ότι δύο άτομα δεν μπορούν να είναι στο ίδιο σημείο συγχρόνως, μπορεί να είναι σημαντικότατος κανόνας οργάνωσης της δομής των πρωτεϊνών [2].

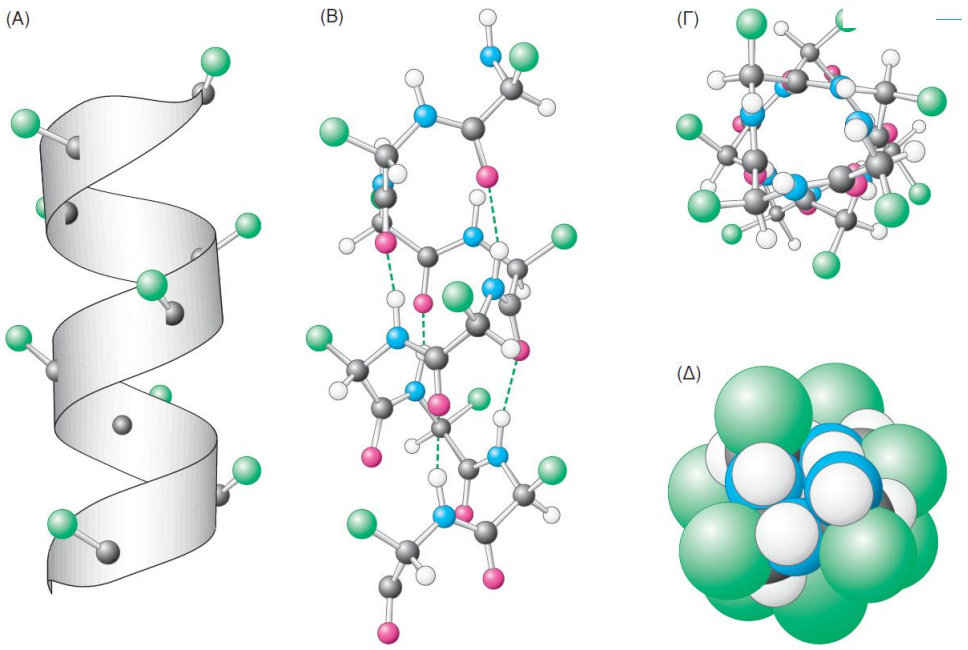


Σχήμα 1.12 Ένα διάγραμμα Ramachandran που δείχνει τις τιμές των φ και ψ [2]. Οι τιμές φ και ψ είναι περιορισμένες λόγω των συγκρούσεων μεταξύ ατόμων. Οι επιτρεπτές τιμές φ και ψ φαίνονται με σκούρο πράσινο, ενώ οι οριακές τιμές φαίνονται με ανοιχτό πράσινο. Η δομή που φαίνεται δεξιά είναι εκείνη που δημιουργεί τις λιγότερες συγκρούσεις μεταξύ των ατόμων.

Η ικανότητα των βιολογικών πολυμερών, όπως οι πρωτεΐνες, να αναδιπλώνονται σε καλά καθορισμένες δομές είναι ιδιαίτερα αξιοσημείωτη θερμοδυναμικά. Φθάνει να σκεφθούμε τις ισορροπίες μεταξύ ενός ξεδιπλωμένου πολυμερούς που έχει στερεοδιάταξη τυχαίου σπειράματος (δηλαδή μείγματος πολλών πιθανών στερεοδιατάξεων) και μιας αναδιπλωμένης μορφής που αποκτά μια μοναδική στερεοδιάταξη. Η απελευθέρωση της εντροπίας σε συνδυασμό με τον μεγάλο αριθμό στερεοδιατάξεων της ξεδιπλωμένης μορφής εμποδίζει την αναδίπλωση η οποία είναι δυνατόν να πραγματοποιηθεί μόνο με αλληλεπιδράσεις που προωθούν την αναδιπλωμένη μορφή. Επομένως, τα πολύ ευέλικτα πολυμερή που έχουν μεγάλο αριθμό πιθανών αναδιπλώσεων δεν οδηγούν σε απόλυτα καθορισμένες μοναδικές δομές. Η έλλειψη ευελιξίας στην πεπτιδική μονάδα και ο περιορισμένος αριθμός των επιτρεπόμενων γωνιών φ και ψ περιορίζουν τον αριθμό των δομών που μπορεί να επιτύχει η ξεδιπλωμένη μορφή της πρωτεΐνης κατά τη διεργασία αναδίπλωσής της.

1.4. Αναδίπλωση των πολυπεπτιδικών αλυσίδων (δευτεροταγής δομή)

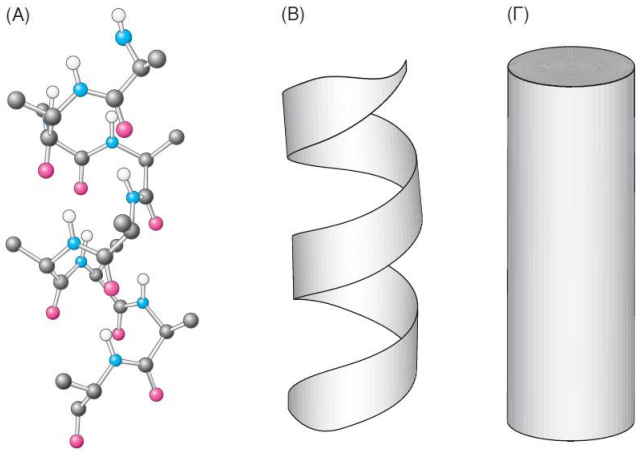
Μια πολυπεπτιδική αλυσίδα αναδιπλώνεται σε κανονικά επαναλαμβανόμενες δομές [23][33]. Το 1951 οι Linus Pauling και Robert Corey πρότειναν δύο περιοδικές δομές που τις ονόμασαν α-έλικα και β-πτυχωτή επιφάνεια. Στη συνέχεια καθορίστηκαν και άλλες δομές όπως η β-στροφή και η Ω-θηλιά. Παρόλο που οι δομές αυτές δεν παρουσιάζουν περιοδικότητα, αποτελούν καλά προσδιορισμένες στροφές ή θηλιές και συνοδεύουν τις α-έλικες και τις β-πτυχωτές επιφάνειες στην τελική τριδιάστατη δομή της πρωτεΐνης.



Σχήμα 1.13 Η δομή μιας α-έλικας [2]. (Α) Η απεικόνιση κορδέλας στην οποία ξεχωρίζουν τα άτομα άνθρακα και οι πλευρικές αλυσίδες. (Β) Μια πλάγια όψη του μοντέλου με σφαίρες και ράβδους όπου διακρίνονται οι δεσμοί υδρογόνου (διακεκομμένες γραμμές) μεταξύ των ομάδων NH και CO. (Γ) Παρατηρώντας από το άκρο της έλικας και παράλληλα προς τον άξονα βλέπουμε τον περιελιγμένο κορμό να σχηματίζει το εσωτερικό της έλικας και τις πλευρικές αλυσίδες να προεξέχουν προς τα έξω. (Δ) Ένα χωροπληρωτικό μοντέλο του (Γ) δείχνει πόσο λίγος κενός χώρος μένει στο κέντρο της έλικας.

- Η α-έλικα

Στην προσπάθειά τους να καθορίσουν τις πιθανές δομές των πεπτιδίων, οι Pauling και Corey έλαβαν υπ' όψιν τους στερικούς περιορισμούς καθώς και την ικανότητα του κορμού του πεπτιδίου να διαμορφώνει συνθήκες δεσμών υδρογόνου μεταξύ των ομάδων NH και CO. Σύμφωνα με το [17], η πρώτη δομή που πρότειναν ήταν εκείνη της α-έλικας, που είναι μια ραβδόμορφη δομή (βλέπε Σχ. 1.14). Ο κορμός, που έχει σχήμα σπειράματος, σχηματίζει το εσωτερικό της ράβδου και οι πλευρικές αλυσίδες εκτείνονται προς τα έξω, σε μια ελικοειδή διάταξη. Η α-έλικα σταθεροποιείται από δεσμούς υδρογόνου μεταξύ των ομάδων NH και CO της κύριας αλυσίδας. Συγκεκριμένα, η ομάδα CO κάθε αμινοξέος σχηματίζει έναν δεσμό υδρογόνου με την ομάδα NH του αμινοξέος που βρίσκεται τέσσερα κατάλοιπα μπροστά στην αλληλουχία. Έτσι, στη δομή αυτή όλες οι ομάδες CO και NH του πολυπεπτιδικού κορμού συνδέονται με δεσμούς υδρογόνου, εκτός από εκείνες που βρίσκονται στα άκρα της έλικας. Κάθε κατάλοιπο απέχει από το επόμενο 1,5 A κατά μήκος του άξονα της έλικας και είναι περιστραμμένο κατά 100°, δίνοντας 3,6 κατάλοιπα αμινοξέων ανά στροφή της έλικας. Συνεπώς, κατάλοιπα που απέχουν μεταξύ τους 3-4 αμινοξέα στην αλληλουχία βρίσκονται αρκετά κοντά το ένα στο άλλο λόγω της δομής της έλικας. Αντίθετα, αμινοξέα που είναι δίπλα στην αλληλουχία διότι βρίσκονται το ένα απέναντι στο άλλο στην έλικα δεν μπορούν να έρθουν σε επαφή. Το βήμα της α-έλικας, που ισούται με το προϊόν της μετατόπισης (1,5 A) επί τον αριθμό των καταλοίπων ανά στροφή (3,6), είναι 5,4 A. Η στροφή της έλικας μπορεί να είναι δεξιόστροφη (σύμφωνα με τους δείκτες του ρολογιού) ή αριστερόστροφη (αντίθετα από τους δείκτες του ρολογιού). Οι δεξιόστροφες έλικες είναι πιο ευνοούμενες ενεργειακά διότι παρουσιάζουν λιγότερες στερικές συγκρούσεις μεταξύ των πλευρικών αλυσίδων και του κορμού. Ουσιαστικά, όλες οι α-έλικες που απαντούν στις πρωτεΐνες είναι δεξιόστροφες. Στο σχηματικό διάγραμμα των πρωτεϊνών, οι α-έλικες εμφανίζονται σαν στριμμένες κορδέλες ή κύλινδροι (βλέπε Σχ.1.14).



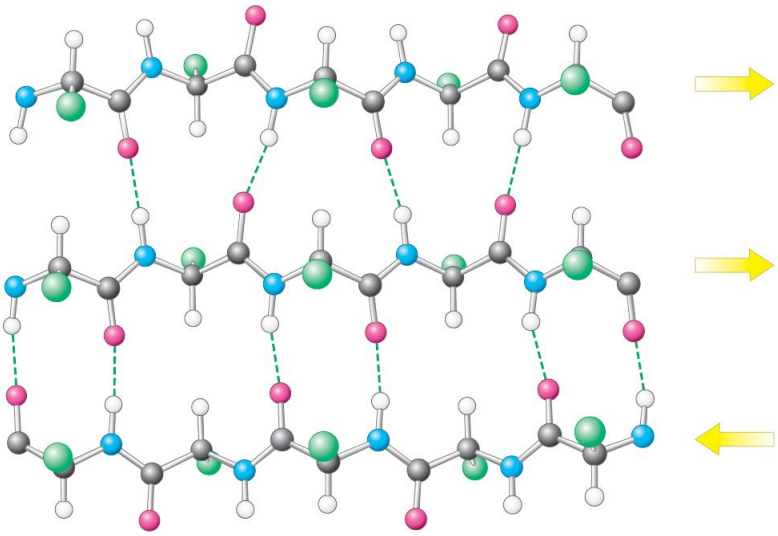
Σχήμα 1.14 Σχηματική απεικόνιση μιας *α*-έλικας [2]. (Α) Μοντέλο με σφαίρες και ράβδους. (Β) Απεικόνιση κορδέλας. (Γ) Απεικόνιση κυλίνδρου.

Το ποσοστό της α-έλικας των πρωτεϊνών ποικίλλει από 0% έως και 100% ανά περίπτωση. Παραδείγματος χάριν, 75% των καταλοίπων της φερριτίνης, της πρωτεΐνης που βοηθά στην αποθήκευση του σιδήρου, βρίσκονται σε α-έλικα. Οι απλές α-έλικες συνήθως έχουν μήκος μικρότερο από 45Α. Παρ' όλα αυτά, δύο ή περισσότερες α-έλικες μπορούν να περιελιχθούν και να δημιουργήσουν μια πολύ σταθερή δομή η οποία μπορεί να έχει μήκος 1000Α (100 nm ή 0,1 μm) ή περισσότερο. Τέτοια στερεοδιάταξη συσπειρωμένου σπειράματος α-ελίκων υπάρχει στη μυοσίνη και την τροπομυοσίνη των μυών, στο ινώδες των θρόμβων του αίματος και στην κερατίνη των μαλλιών. Οι ελικοειδείς ράβδοι των πρωτεϊνών αυτών έχουν μηχανικό ρόλο και δημιουργούν άκαμπτα δεμάτια ινιδίων, όπως τα αγκάθια του σκαντζόχοιρου. Ο κυτταρικός σκελετός είναι πλούσιος στα λεγόμενα ενδιάμεσα νημάτια τα οποία είναι επίσης συσπειρωμένα σπειράματα διπλών α-ελίκων. Πολλές πρωτεΐνες που διαπερνούν βιολογικές μεμβράνες περιέχουν επίσης α-έλικες.

- Οι β-επιφάνειες

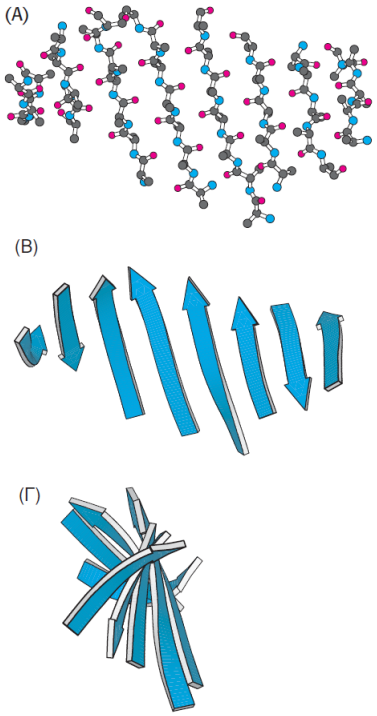
Η β-πτυχωτή επιφάνεια (ή απλώς β-επιφάνεια) διαφέρει σημαντικά από τη ραβδόμορφη α-έλικα [23][33]. Μια πολυπεπτιδική αλυσίδα, που ονομάζεται β-πτύχωση, σε μια β-επιφάνεια είναι σχεδόν απόλυτα απλωμένη, αντίθετα από το σφιχτό σπείραμα της α-έλικας. Η απόσταση μεταξύ γειτονικών αμινοξέων σε μια β-πτύχωση είναι περί-που 3.5 Α, ενώ στην α-έλικα υπενθυμίζεται ότι είναι 1.5 Α. Οι πλευρικές αλυσίδες των γειτονικών αμινοξέων έχουν αντίθετη κατεύθυνση. Μια β-επιφάνεια δημιουργείται όταν δύο ή

περισσότερες β-πτυχώσεις συνδεθούν με δεσμούς υδρογόνου. Οι διαδοχικές β-πτυχώσεις στη β-επιφάνεια μπορεί να έχουν την ίδια κατεύθυνση (παράλληλη β-επιφάνεια) ή να έχουν αντίθετη κατεύθυνση (αντιπαράλληλη β-επιφάνεια). Στην αντιπαράλληλη β-επιφάνεια οι ομάδες NH και CO ενός αμινοξέος συνδέονται αντίστοιχα με δεσμούς υδρογόνου με τις ομάδες CO και NH του αμινοξέος της γειτονικής β-πτύχωσης. Στην παράλληλη β-επιφάνεια η διάταξη των δεσμών υδρογόνου είναι λίγο πιο περίπλοκη. Για κάθε αμινοξύ η ομάδα NH συνδέεται στο CO του αμινοξέος της γειτονικής β-πτύχωσης, ενώ η ομάδα CO συνδέεται στο NH του αμινοξέος που βρίσκεται δύο κατάλοιπα πιο κάτω στην αλυσίδα. Πολλές πτυχώσεις, συνήθως 4-5 αλλά μπορεί και 10 ή και περισσότερες, συνδέονται προκειμένου να δημιουργήσουν β-επιφάνειες και μπορεί να είναι είτε καθαρά παράλληλες είτε αντιπαράλληλες είτε μεικτές (βλέπε Σχ. 1.15).



Σχήμα 1.15 Δομή μιας μεικτής β-επιφάνειας

Στα σχηματικά διαγράμματα, οι β-πτυχώσεις συνήθως εμφανίζονται ως φαρδιά βέλη με κατεύθυνση προς το καρβοξυ-τελικό άκρο, προσδιορίζοντας τον τύπο της β-επιφάνειας που σχηματίζεται, παράλληλη ή αντιπαράλληλη. Οι β-επιφάνειες παρουσιάζουν μεγαλύτερη ποικιλία από τις α-έλικες, μπορεί να είναι σχετικά ευθείες αλλά οι περισσότερες εμφανίζονται με την κάθε πτύχωση ελαφρά στριμμένη (βλέπε Σχ. 1.16). Η β-επιφάνεια είναι σημαντικό συστατικό πολλών πρωτεϊνών. Παραδείγματος χάριν, οι πρωτεΐνες που δεσμεύουν τα λιπαρά οξέα και είναι τόσο σημαντικές για τον μεταβολισμό των λιπών αποτελούνται σχεδόν αποκλειστικά από β-επιφάνειες.

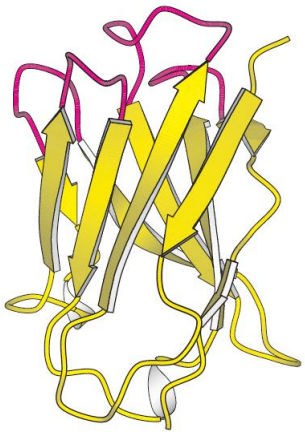


Σχήμα 1.16 Μια β-επιφάνεια όπου η κάθε πτύχωση είναι ελαφρώς στριμμένη σε σχέση με την προηγούμενη [2]. (Α) Μοντέλο με σφαίρες και ράβδους. (Β) Σχηματικό μοντέλο. (Γ) Σχηματική διαμόρφωση που έχει στραφεί κατά 90° για να εμφανίσει καλύτερα το στρίψιμο.

- Οι πολυπεπτιδικές αλυσίδες μπορούν να αλλάξουν κατεύθυνση δημιουργώντας αντίθετες στροφές και θηλιές

Οι πιο πολλές πρωτεΐνες έχουν συμπαγές σφαιρικό σχήμα και επομένως για τη δημιουργία τους απαιτούνται αναστροφές της πολυπεπτιδικής αλυσίδας τους. Πολλές από τις αναστροφές επιτυγχάνονται με ένα κοινό δομικό στοιχείο που ονομάζεται στροφή αναστροφής (γνωστό επίσης και ως β-στροφή ή κάμψη φουρκέτας) και απεικονίζεται στο Σχήμα 1.16. Σε πολλές β-στροφές η ομάδα CO του καταλοίπου i της αλυσίδας του πολυπεπτιδίου δημιουργεί δεσμό υδρογόνου με την ομάδα NH του καταλοίπου i+3. Η αλληλεπίδραση αυτή σταθεροποιεί την απότομη αλλαγή στην κατεύθυνση της πολυπεπτιδικής αλυσίδας. Υπάρχουν βέβαια και περιπτώσεις όπου η αναστροφή της αλυσίδας γίνεται μέσω πιο πολύπλοκων δομών που ονομάζονται θηλιές ή και Ω-θηλιές λόγω του σχήματός τους. Οι θηλιές, σε αντίθεση με τις α-έλικες και τις β-επιφάνειες, δεν έχουν κανονικές περιοδικές δομές. Παρ’ όλα αυτά, και οι δομές θηλιάς έχουν συχνά σταθερή και απόλυτα καθορισμένη δομή (βλέπε Σχ. 1.17). Οι στροφές και οι θηλιές

βρίσκονται κυρίως στην επιφάνεια των πρωτεϊνών και επομένως συμμετέχουν στην αλληλεπίδραση των πρωτεϊνών με άλλα μόρια. Η κατανομή της πρωτεϊνικής αλυσίδας σε α-έλικες, β-πτυχώσεις και στροφές αναφέρεται συνήθως ως δευτεροταγής δομή.



Σχήμα 1.17 Οι θηλιές στην επιφάνεια μιας πρωτεΐνης [2]. Ένα τμήμα του μορίου του αντισώματος έχει θηλιές στην επιφάνειά του (κόκκινο χρώμα) που αλληλεπιδρούν με άλλα μόρια.

1.5. Η μελέτη του ισομερισμού της προλίνης

Αξιοποιώντας τη γνώση για τη δομή των πρωτεϊνών όπως παρουσιάστηκε στο κεφ.1 και τις διάφορες μεθόδους κατηγοριοποίησης, που παρουσιάζονται στο κεφ.2, μοντελοποιήσαμε το πρόβλημα της πρόβλεψης του ισομερισμού της προλίνης σε cis ή trans. Αρχικά, για να το καταφέρουμε αυτό αναπτύξαμε λογισμικό το οποίο αντλεί πληροφορία από πρωτεομικά αρχεία και εξάγει συγκεκριμένα χαρακτηριστικά για κάθε προλίνη. Έπειτα αναλύοντας ένα συγκεκριμένο σύνολο πρωτεϊνών με αυτό το λογισμικό προχωρήσαμε στη μοντελοποίηση του προβλήματος με διάφορα είδη ταξινομητών χρησιμοποιώντας συγκεκριμένα χαρακτηριστικά.

ΚΕΦΑΛΑΙΟ 2. ΘΕΩΡΙΑ ΤΑΞΙΝΟΜΗΤΩΝ

2.1 Μέθοδοι Ταξινόμησης
2.2 Ο κανόνας του κοντινότερου γείτονα (k nearest neighbors)
2.3 Μηχανές διανυσμάτων υποστήριξης (support vector machines – svm)
2.4 Δέντρα απόφασης (decision trees)
2.5 Ταξινομητής naive bayes
2.6 Συλλογική Ταξινόμηση (ensemble)
    2.6.1 Μοντέλο Συλλογικών Ταξινομητών
2.7 Επικύρωση του σφάλματος των ταξινομητών
    2.7.1 Η μέθοδος holdout (train set-test set)
    2.7.2 Η μέθοδος Cross Validation (διασταυρωμένης επικύρωσης)
2.8 Η μέθοδος Random Subsampling
2.9 Απόδοση κατηγοριοποίησης (Αξιολόγηση ταξινομητών)
    2.9.1 Ανισοκατανομή μεταξύ των κατηγοριών (Class imbalance)

2.1. Μέθοδοι Ταξινόμησης

Οι τεχνικές εξόρυξης δεδομένων έχουν σκοπό την ανακάλυψη ενδιαφερόντων ή πρότυπων σχημάτων μέσα σε μεγάλα σύνολα δεδομένων ώστε αυτά να βοηθήσουν τους ειδικούς να λαμβάνουν αποφάσεις σχετικά με σημαντικές μελλοντικές δραστηριότητες. Η εξόρυξη γνώσης είναι στενά συνδεδεμένη με τις περιοχές των Βάσεων Δεδομένων, της Στατιστικής και της Τεχνητής Νοημοσύνης (Ανακάλυψη Γνώσης, Μηχανική Μάθηση).

Οι διαδικασίες εξόρυξης γνώσης σε Βάσεις Δεδομένων (ή Διαδικασίες KDD όπως αλλιώς ονομάζονται) απαιτούν τέσσερα στάδια. Το πρώτο στάδιο είναι η επιλογή των δεδομένων,

όπου επιλέγεται το σύνολο δεδομένων και τα χαρακτηριστικά (attributes) που μας ενδιαφέρουν σε σχέση με το στόχο μας. Το δεύτερο στάδιο καλείται καθαρισμός των δεδομένων. Στο στάδιο αυτό απομακρύνουμε το θόρυβο, χειριζόμαστε τις κενές τιμές, μετασχηματίζουμε τις τιμές των χαρακτηριστικών σε κοινές μονάδες μέτρησης και δημιουργούμε νέα χαρακτηριστικά συνδυάζοντας τα ήδη υπάρχοντα. Το επόμενο στάδιο είναι το στάδιο της εξόρυξης γνώσης όπου εξάγουμε τα πραγματικά πρότυπα σχήματα. Το τέταρτο και τελευταίο στάδιο είναι αυτό της αξιολόγησης όπου τα πρότυπα σχήματα παρουσιάζονται στο χρήστη, ο οποίος τα αξιολογεί.

Ένα από τα σημαντικότερα χαρακτηριστικά της εξόρυξης δεδομένων το οποίο λαμβάνεται υπόψη κυρίως από την περιοχή των Βάσεων Δεδομένων, είναι ο πολύ μεγάλος όγκος των δεδομένων. Η δυνατότητα κλιμάκωσης σε σχέση με το μέγεθος των δεδομένων αποτελεί ένα πολύ σημαντικό κριτήριο σε όλες τις τεχνικές της εξόρυξης δεδομένων. Μια τεχνική λέμε ότι κλιμακώνεται αν ο χρόνος εκτέλεσης αυξάνεται ανάλογα με το μέγεθος του συνόλου δεδομένων όπου ενεργεί (δηλ. γραμμικά), για δεδομένα μεγέθη των διαθέσιμων πόρων του συστήματος (π.χ μέγεθος της κύριας μνήμης, μέγεθος του δίσκου, κτλ). Όλες οι τεχνικές εξόρυξης γνώσης πρέπει να χαρακτηρίζονται από τη δυνατότητα της κλιμάκωσης.

Ένα άλλο επιθυμητό χαρακτηριστικό των τεχνικών εξόρυξης δεδομένων είναι το ότι πρέπει να εκτελούνται με τον ταχύτερο δυνατό τρόπο. Αυτό σημαίνει ότι η σειριακή επεξεργασία του συνόλου δεδομένων, το οποίο είναι πολύ μεγάλο, δεν αποτελεί σωστή προσέγγιση. Έτσι, τα παραδείγματα (ή αντικείμενα) του συνόλου δεδομένων, θα πρέπει να δεικτοδοτούνται ώστε αυτά να είναι άμεσα προσπελάσιμα.

Ίσως η πιο γνωστή εφαρμογή εξόρυξης δεδομένων είναι η κατηγοριοποίηση (ή ταξινόμηση) η οποία χρησιμοποιείται για να τοποθετήσει τα παραδείγματα ενός συνόλου παραδειγμάτων σε κατηγορίες. Διαφορετικά πεδία της πληροφορικής έχουν αναπτύξει διαφορετικές τεχνικές κατηγοριοποίησης. Κάποιες έχουν μεγάλη αποδοχή και άλλες λιγότερη.

Πολλές εταιρείες του ιδιωτικού και του δημόσιου τομέα χρησιμοποιούν σε καθημερινή βάση συστήματα κατηγοριοποίησης. Παραδείγματα τέτοιου είδους συστημάτων είναι τα συστήματα αναγνώρισης προτύπων, συστήματα ιατρικών διαγνώσεων, συστήματα

έγκρισης δανείων και πιστωτικών καρτών, συστήματα ανίχνευσης λαθών σε βιομηχανικές εφαρμογές, συστήματα κατηγοριοποίησης των τάσεων στην οικονομία κ.α

Όλες οι προσεγγίσεις στην εκτέλεση της κατηγοριοποίησης προϋποθέτουν γνώση των δεδομένων. Συνήθως χρησιμοποιούμε ένα σύνολο εκπαίδευσης για να καθορίσει τις συγκεκριμένες παραμέτρους που απαιτούνται από την τεχνική. Τα δεδομένα εκπαίδευσης (ή σύνολο εκπαίδευσης – training set) αποτελούνται από ένα δείγμα δεδομένων εισόδου καθώς επίσης και από την κατηγοριοποίηση που έχει δοθεί σε αυτά τα δεδομένα. Το πρόβλημα της κατηγοριοποίησης παρουσιάζεται στον πιο κάτω ορισμό.

Ορισμός: Η κατηγοριοποίηση (classification) είναι η διαδικασία η οποία απεικονίζει ένα σύνολο δεδομένων σε προκαθορισμένες ομάδες. Τις ομάδες αυτές συχνά τις καλούμε κατηγορίες ή κλάσεις.

Ο ορισμός θεωρεί την κατηγοριοποίηση σαν μια απεικόνιση από το σύνολο παραδειγμάτων στο σύνολο των κατηγοριών. Πρέπει να υπογραμμιστεί ότι οι κατηγορίες είναι προκαθορισμένες, δεν επικαλύπτονται και διαμερίζουν ολόκληρο το σύνολο παραδειγμάτων. Κάθε στοιχείο του συνόλου παραδειγμάτων τοποθετείται σε ακριβώς μια κατηγορία.

Η επίλυση των προβλημάτων κατηγοριοποίησης περιλαμβάνει δύο βασικά στάδια: Δημιουργούμε ένα μοντέλο από την αξιολόγηση και την ανάλυση των δεδομένων εκπαίδευσης. Αυτό το βήμα έχει σαν είσοδο τα δεδομένα εκπαίδευσης και σαν έξοδο το μοντέλο που αναπτύχθηκε. Το μοντέλο που δημιουργείται από αυτό το στάδιο είναι σε θέση να κατηγοριοποιεί τα δεδομένα εκπαίδευσης με όσο το δυνατό μεγαλύτερη ακρίβεια. Όταν είναι ήδη γνωστές οι κατηγορίες του συνόλου των δεδομένων εκπαίδευσης, δηλαδή το σύνολο των δεδομένων εκπαίδευσης περιλαμβάνει ένα χαρακτηριστικό το οποίο δείχνει την κλάση (κατηγορία) στην οποία κατηγοριοποιείται το κάθε παράδειγμα, τότε το βήμα αυτό καλείται εποπτευμένη μάθηση (supervised learning), σε αντίθετη περίπτωση, δηλαδή όταν δεν είναι γνωστές οι κατηγορίες του συνόλου των δεδομένων εκπαίδευσης, τότε το βήμα αυτό καλείται μη εποπτευμένη μάθηση (unsupervised learning – clustering).

Υπάρχουν δύο βασικές μέθοδοι που χρησιμοποιούνται για να λύσουν το πρόβλημα της κατηγοριοποίσης

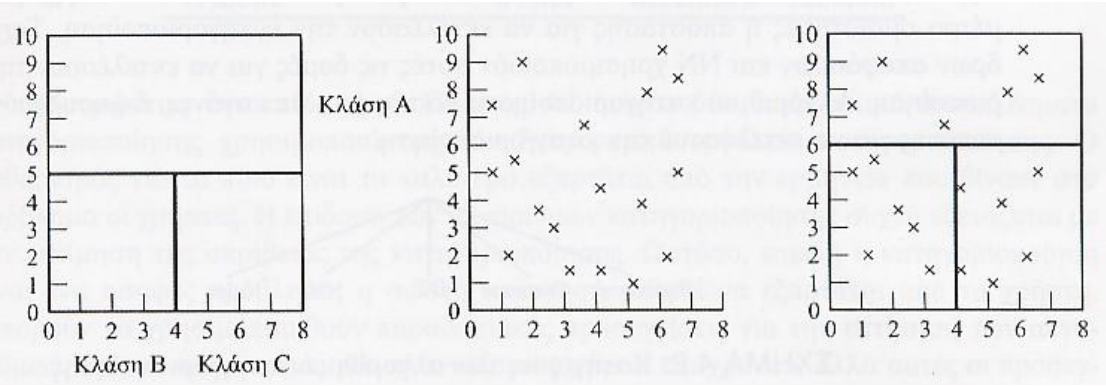
- Καθορισμός των ορίων:

  Η κατηγοριοποίηση εκτελείται με διαίρεση του χώρου της εισόδου των εν δυνάμει παραδειγμάτων του συνόλου παραδειγμάτων σε περιοχές όπου κάθε περιοχή συνδέεται με μια κατηγορία.

- Χρήση κατανομών πιθανότητας:

  Για κάθε κατηγορία που δίνεται C_j $P(t_i \mid C_j)$ είναι η συνάρτηση κατανομής πιθανότητας (probability distribution function) για την κατηγορία υπολογισμένη σε ένα σημείο, t_i. Αν η πιθανότητα εμφάνισης κάθε κατηγορίας P (Cj), είναι γνωστή (ίσως να έχει οριστεί από κάποιον ειδικό του πεδίου εφαρμογής – domain expert), τότε $P(C_j) \cdot P(t_i \mid C_j)$ είναι η εκτίμηση της πιθανότητας όπου η t_i ανήκει στην κατηγορία $\mid C_j$.

Ας υποθέσουμε ότι μας δίνεται ένα σύνολο παραδειγμάτων που αποτελείται από παραδείγματα της μορφής t=<x,y> όπου $0 \leq x \leq 8$ και $0 \leq y \leq 10$. Το (Σχήμα 2.1) παρουσιάζει το πρόβλημα της κατηγοριοποίησης. Το (Σχήμα 2.1) (1^o διάγραμμα) παρουσιάζει τις προκαθορισμένες κατηγορίες – κλάσεις, το (Σχήμα 2.1) (2^o διάγραμμα) παρέχει δείγματα δεδομένων εισόδου και το (Σχήμα 2.1) (3^o διάγραμμα) παρουσιάζει την κατηγοριοποίηση των δεδομένων με βάση τις ορισμένες κατηγορίες.



Σχήμα 2.1 Το πρόβλημα της κατηγοριοποίησης. (1^o διάγραμμα: ορισμός κατηγοριών, 2^o διάγραμμα: σύνολο παραδειγμάτων προς κατηγοριοποίηση, 3^o διάγραμμα κατηγοριοποιημένο σύνολο παραδειγμάτων)

Ένα πολύ σημαντικό ζήτημα σχετικό με την κατηγοριοποίηση είναι η υπερπροσαρμογή. Συγκεκριμένα, λέγοντας υπερπροσαρμογή εννοούμε το φαινόμενο κατά το οποίο η τεχνική κατηγοριοποίησης ταιριάζει ακριβώς στα δεδομένα εκπαίδευσης και ίσως να μη μπορεί να εφαρμοστεί σε πιο ευρύ πληθυσμό δεδομένων. Για παράδειγμα, ας υποθέσουμε ότι τα δεδομένα εκπαίδευσης περιέχουν λανθασμένα δεδομένα ή δεδομένα με θόρυβο. Σε αυτή την περίπτωση, το ακριβές ταίριασμα των δεδομένων δεν είναι επιθυμητό.

Είναι πολύ συχνό το φαινόμενο κατά το οποίο τιμές χαρακτηριστικών (attributes) των παραδειγμάτων των συνόλων δεδομένων είναι λανθασμένες (θόρυβος), ελλιπείς (missing values) και ασυνεπείς. Το φαινόμενο του θορύβου προέρχεται από ανθρώπινα λάθη ή λάθη του υπολογιστή. Αντίστοιχα το φαινόμενο των ελλιπών δεδομένων προέρχεται από μη εισαγωγή στοιχείων για κάποια συγκεκριμένα παραδείγματα την ώρα της εισαγωγής αφού αυτά, την συγκεκριμένη στιγμή δεν είχαν αξία. Επίσης, το φαινόμενο των ασυνεπών δεδομένων προέρχεται από ενοποιήσεις δεδομένων όπου ένα χαρακτηριστικό έχει διαφορετικό όνομα στις Διαφορετικά σύνολα παραδειγμάτων.Έτσι, πριν ξεκινήσει η διαδικασία της εκπαίδευσης του αλγορίθμου κατηγοριοποίησης και της δοκιμής του, θα πρέπει να γίνει το λεγόμενο καθάρισμα των δεδομένων (data cleaning).

Το φαινόμενο των ελλιπών δεδομένων μπορεί να αντιμετωπιστεί με έναν από τους παρακάτω τρόπους:

- Αγνόησε το παράδειγμα: Αυτό γίνεται όταν λείπει η τιμή του χαρακτηριστικού κατηγορίας.
- Γέμισε τις τιμές που λείπουν με το χέρι: χρονοβόρα μέθοδος. Δεν είναι εφικτή αν είναι πάρα πολλά τα παραδείγματα.
- Χρησιμοποίησε μια σταθερά για το γέμισμα των τιμών που λείπουν (π.χ «unknown») . Το σύστημα θα χρησιμοποιήσει λανθασμένα αυτού του είδους τις τιμές. Αν και είναι απλή μέθοδος, δεν προτείνεται.
- Χρήση του μέσου όρου των τιμών του χαρακτηριστικού για την συμπλήρωση των τιμών που λείπουν για το συγκεκριμένο χαρακτηριστικό.
- Χρήση του μέσου όρου των τιμών του χαρακτηριστικού για όλα τα παραδείγματα που ανήκουν στην ίδια κατηγορία. π.χ συμπλήρωση των τιμών που λείπουν με το μέσο όρο του εισοδήματος για τους πελάτες με το ίδιο credit_risk.
- Χρήση της πιο πιθανής τιμής.

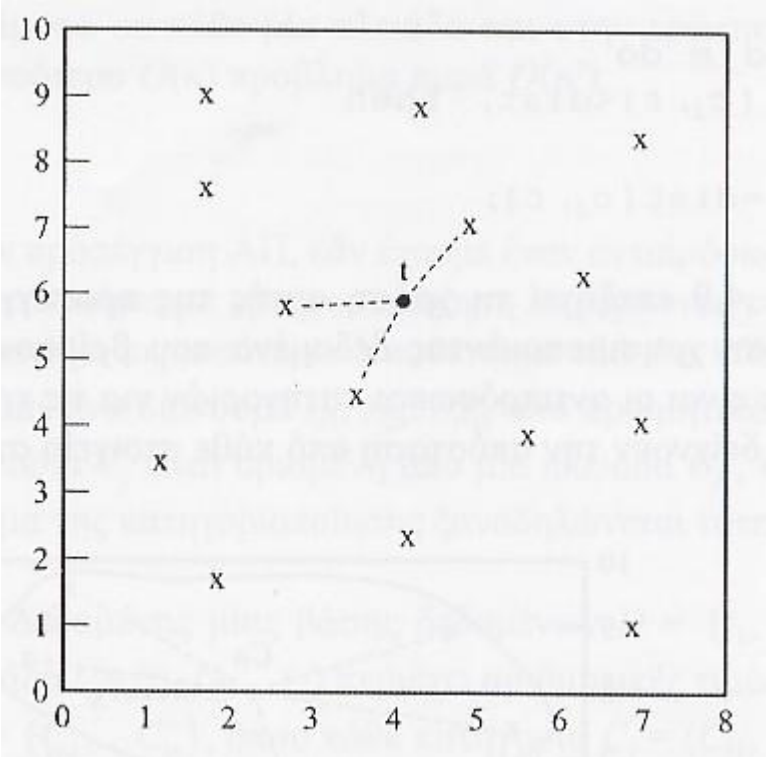
Το φαινόμενο του θορύβου αντιμετωπίζεται με έναν από τους εξής τρόπους:

- Binning: Εξομαλύνει τα δεδομένα λαμβάνοντας υπόψη τις τιμές των γειτόνων. Υπάρχουν πολλές τεχνικές binning.
- Clustering: Τα παραδείγματα που είναι outliers μπορούν να βρεθούν χρησιμοποιώντας τεχνικές clustering.
- Συνδυασμός ελέγχου υπολογιστή και ανθρώπου ώστε να εντοπιστούν οι outliers.
- Regression: μπορούν να χρησιμοποιηθούν τεχνικές Regression ώστε οι τιμές των μεταβλητών να μπορούν να βρεθούν από τις άλλες μεταβλητές.[21]

2.2. Ο κανόνας του κοντινότερου γείτονα (k nearest neighbors)

Μια από τις πιο δημοφιλείς τεχνικές κατηγοριοποίησης είναι αυτή των Κ- κοντινότερων γειτόνων (K nearest neighbors – KNN). Η κατηγοριοποίηση ενός παραδείγματος του συνόλου παραδειγμάτων βάση αυτής της τεχνικής πραγματοποιείται αναζητώντας τα παραδείγματα που μοιάζουν περισσότερο σε αυτό. Τα παραδείγματα αυτά ονομάζονται κοντινότεροι γείτονες. Η παράμετρος Κ της τεχνικής, ορίζει τον αριθμό των κοντινότερων γειτόνων που έχουν λόγο στην κατηγοριοποίηση. Το υπό εξέταση παράδειγμα τοποθετείται στην κατηγορία στην οποία ανήκουν οι περισσότεροι από τους Κ κοντινότερους γείτονες. Το πόσοι κοντινότεροι γείτονες θα ληφθούν υπόψη για να γίνει η κατηγοριοποίηση είναι ένα σημαντικό ζήτημα αφού η ακρίβεια της κατηγοριοποίησης εξαρτάται από αυτόν τον αριθμό. Έτσι, είναι απαραίτητο να βρούμε τον αριθμό Κ που δίνει την καλύτερη κατηγοριοποίηση. Από την άλλη, αν ο αριθμός αυτός είναι ιδιαίτερα μεγάλος, ο χρόνος που θα απαιτηθεί από την τεχνική θα είναι μακρύς και αυτό επειδή η αναζήτηση των Κ αντικειμένων που μοιάζουν περισσότερο με το υπό εξέταση αντικείμενο σε ένα μεγάλο σύνολο δεδομένων είναι μια χρονοβόρα διαδικασία ακόμη και στην περίπτωση όπου έχουμε δεικτοδότηση των δεδομένων. Άρα καταλήγουμε στο ότι μπορεί να μας συμφέρει το να αναζητούμε έναν μικρότερο αριθμό κοντινότερων γειτόνων, χάνοντας λίγο σε ακρίβεια της κατηγοριοποίησης αλλά κερδίζοντας πολύ σε χρόνο εκτέλεσης. Ένας τέτοιος συμβιβασμός είναι απαραίτητος σε συστήματα πραγματικού χρόνου.

Πιο αναλυτικά, η τεχνική KNN προϋποθέτει ότι το σύνολο εκπαίδευσης (training set) δεν περιλαμβάνει μόνο τα δεδομένα αλλά επίσης και την επιθυμητή κατηγοριοποίηση για κάθε στοιχείο. Όταν πρόκειται να γίνει μια κατηγοριοποίηση για ένα νέο στοιχείο, πρέπει να καθοριστεί η απόστασή του από κάθε στοιχείο του συνόλου εκπαίδευσης. Μόνο οι K κοντινότερες καταχωρήσεις στο σύνολο εκπαίδευσης λαμβάνονται υπόψη στη συνέχεια. Το νέο στοιχείο τοποθετείται στην κατηγορία που περιέχει τα περισσότερα στοιχεία από το σύνολο των K κοντινότερων στοιχείων. Το (Σχήμα 2.2) παρουσιάζει τη διαδικασία που χρησιμοποιείται από τον αλγόριθμο KNN. Στο σχήμα φαίνονται τα σημεία του συνόλου εκπαίδευσης. Παρουσιάζονται τα τρία κοντινότερα στοιχεία στο σύνολο εκπαίδευσης. Το t θα τοποθετηθεί στην κατηγορία στην οποία ανήκουν τα περισσότερα από αυτά τα K στοιχεία.



Σχήμα 2.2 Κατηγοριοποίηση με χρήση KNN

Τέλος, υπάρχει μια παραλλαγή του αλγορίθμου k κοντινότερων γειτόνων, που είναι γνωστή με το όνομα κοντινότερος γείτονας σταθμισμένης απόστασης. Σύμφωνα με την παραλλαγή αυτή, το πόσο συνεισφέρει κάθε γείτονας στην κατηγοριοποίηση υπολογίζεται βάσει ενός βάρους, ανάλογα με την απόστασή του από το ζητούμενο. Έτσι οι κοντινότεροι γείτονες έχουν μεγαλύτερη συνεισφορά αφού έχουν αυξημένο βάρος, ενώ οι μακρινότεροι

(από τους k γείτονες) έχουν μικρότερη συνεισφορά. Χρησιμοποιώντας αυτή τη μέθοδο, μπορούμε να λάβουμε υπόψη όλα τα παραδείγματα του συνόλου των δεδομένων εκπαίδευσης και όχι μόνο τις k κοντινότερες.[21]

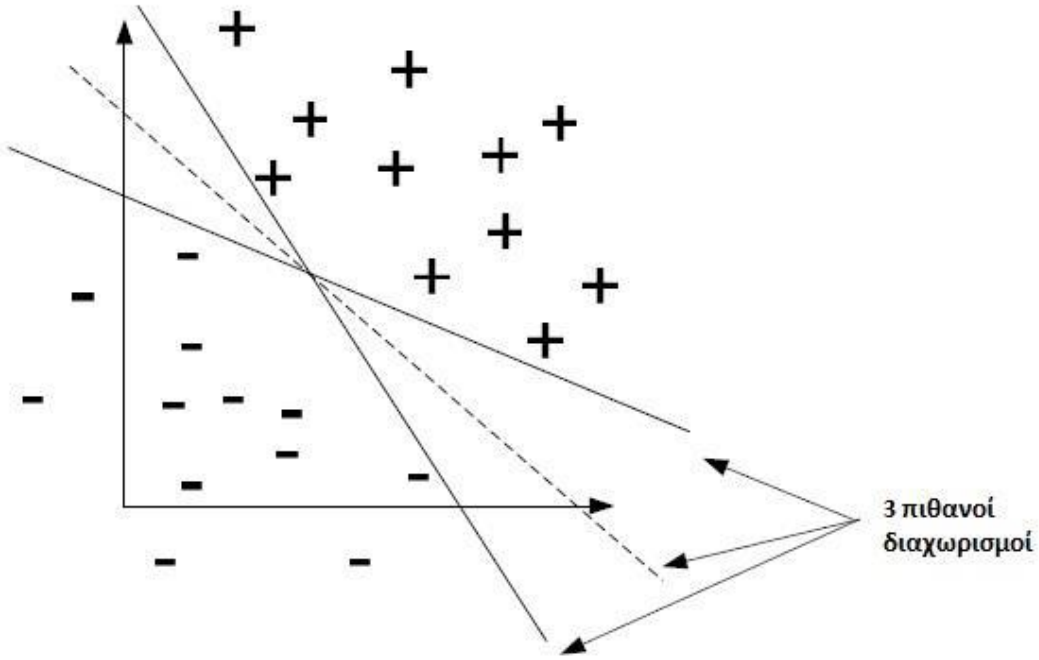
2.3. Μηχανές διανυσμάτων υποστήριξης (support vector machines – svm)

Οι μηχανές διανυσμάτων υποστήριξης (SVMs) ανήκουν στην ευρύτερη κατηγορία των γενικευμένων γραμμικών ταξινομητών, καθώς επιτυγχάνουν το διαχωρισμό των δεδομένων με τη χρήση υπερεπιπέδων. Η θεωρία των SVMs έχει τις αρχές της στα τέλη της δεκαετίας του ’70, αλλά η πρώτη παρουσίασή της έγινε το 1992 από τους Boser, Guyon και Vapnik στο COLT-92 (Computational Learning Theory – 92). Τα τελευταία χρόνια, ωστόσο, παρατηρείται πολύ συχνή χρήση της σε πολλούς τομείς αναγνώρισης προτύπων, όπως στην αυτόματη αναγνώριση των χαρακτήρων της αλφαβήτου, στην αναγνώριση αντικειμένων, στον εντοπισμό προσώπων σε εικόνες, κτλ. Τα SVMs αποτελούν μια κατηγορία μεθόδων μάθησης με επίβλεψη που χρησιμοποιούνται για ταξινόμηση και παλινδρόμηση. Με άλλα λόγια, η μηχανή διανυσμάτων υποστήριξης (SVM) είναι ένα εργαλείο ταξινόμησης και πρόβλεψης συναρτήσεων, που χρησιμοποιεί τη θεωρία εκμάθησης μηχανών (computer learning) για να μεγιστοποιεί την ακρίβεια πρόβλεψης, ενώ αποφεύγει την υπερπροσαρμογή (over-fitting) στα στοιχεία. Μπορούν να οριστούν ως τα συστήματα που μετατρέπουν το χώρο υπόθεσης των γραμμικών συναρτήσεων σε έναν χώρο χαρακτηριστικών μεγάλης διάστασης, που εκπαιδεύεται με έναν αλγόριθμο βελτιστοποίησης. Οι μηχανές SVM έγιναν διάσημες όταν έδωσαν ακρίβεια συγκρίσιμη με αυτή περίπλοκων νευρωνικών δικτύων στην προσπάθεια αναγνώρισης γραφής.

- Γραμμικά SVMs

Αποτελούν την απλούστερη περίπτωση SVM και εκπαιδεύονται σε γραμμικώς διαχωρίσιμα δεδομένα. Ας θεωρήσουμε ένα πρόβλημα δυαδικής ταξινόμησης. Ο στόχος είναι η εύρεση ενός υπερεπιπέδου που να χωρίζει το σύνολο κατά τέτοιον τρόπο ώστε τα διανύσματα ίδιας κατηγορίας να ανήκουν στην ίδια πλευρά του υπερεπιπέδου.

Έστω στοιχεία εκπαίδευσης $D=\{(\vec{x}_i, y_i), i=1,...,N\}$ με $y_i \in \{-1, +1\}$. Το στοιχείο $\vec{x}$ έχει διάσταση n και ανήκει σε μία από τις 2 κατηγορίες -1 ή + 1.



Σχήμα 2.3 Παράδειγμα ταξινόμησης 2D

Στο παραπάνω σχήμα φαίνεται πως υπάρχουν αρκετοί πιθανοί τρόποι διαχωρισμού των δεδομένων. Κατά συνέπεια, πρέπει να βρεθεί ο καλύτερος γραμμικός ταξινομητής (classifier) του τύπου:

$$f(\vec{x}) = \vec{w}^T \vec{x} + b = (w_1 x_1 + w_2 x_2 + w_3 x_3 + ... + w_n x_n + b) \qquad \text{Εξ. 2.1.}$$

από ένα άπειρο αριθμό υπερεπιπέδων που επιτυγχάνουν ακρίβεια 100% στην κατηγοριοποίηση των στοιχείων εκπαίδευσης. Είναι απαραίτητο το βέλτιστο υπερεπίπεδο, να μην προσεγγίζει περισσότερο το ένα από τα δύο σύνολα δεδομένων και κατά συνέπεια να παρέχει κάποια ανοχή και σε περίπτωση εισαγωγής νέων δεδομένων. Προφανής λύση

είναι ο ταξινομητής μέγιστου περιθωρίου (maximum margin) μεταξύ των δύο συνόλων δεδομένων (θετικών και αρνητικών). Βέλτιστος γραμμικός ταξινομητής είναι το υπερεπίπεδο στη μέση του περιθωρίου.

Δοθείσης της συνάρτησης f(x), η ταξινόμηση λαμβάνεται από τον τύπο:

$$\hat{y} = sign(f(x)) = \begin{cases} +1, f(x) > 0 \\ -1, f(x) < 0 \end{cases} \quad \text{Εξ. 2.2}$$

Διαφορετικά w και b μπορούν να οδηγήσουν στην ίδια ταξινόμηση. Δηλαδή:

$$\hat{y} = sign(a(w^T x + b)) = sign(w^T x + b) \quad \text{Εξ. 2.3}$$

Επομένως, υπάρχουν πολλές λύσεις.

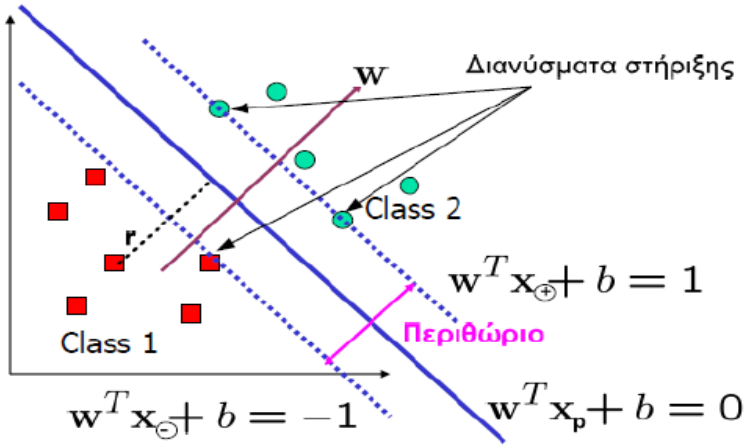
Από όλες τις πιθανές λύσεις αναζητείται η $f(x) = (w^T x + b)$ με το μέγιστο περιθώριο, έτσι ώστε για τα σημεία x_i πιο κοντά στο διαχωριστικό υπερεπίπεδο να ισχύει:

$$| w^T x + b | = 1 \quad \text{Εξ. 2.4}$$

Τα διανύσματα αυτά ονομάζονται **διανύσματα υποστήριξης.**

Για τα υπόλοιπα σημεία θα ισχύει :

$$| w^T x + b | > 1 \quad \text{Εξ. 2.5}$$



Σχήμα 2.4 Απεικόνιση περιθωρίου (margin) και διανυσμάτων στήριξης (support vector)

Αν υπάρχει ένα υπερεπίπεδο που να ικανοποιεί τις εξισώσεις (2.4) και (2.5), τότε το πρόβλημα είναι γραμμικά διαχωρίσιμο. Αν w και b είναι οι βέλτιστες τιμές για το διάνυσμα βαρών και το bias, τότε βέλτιστο υπερεπίπεδο περιγράφεται από τη σχέση:

$$w^T x + b = 0 \qquad \text{Εξ. 2.6}$$

Τα x_i για τα οποία η εξίσωση (2.4) ισχύει σαν ισότητα ονομάζονται «Διανύσματα Υποστήριξης» (Support Vectors). Τα x_i αυτά, βρίσκονται πιο κοντά στο επίπεδο απόφασης και κατά συνέπεια είναι δυσκολότερο να ταξινομηθούν. Πιθανή μετακίνησή τους οδηγεί σε διαφορετικό βέλτιστο υπερεπίπεδο και συνεπάγεται αλλαγή της λύσης του προβλήματος.

Όσον αφορά στο μήκος του περιθωρίου (margin) ως συνάρτηση του w, έστω ότι r είναι η απόσταση του σημείου x από το υπερεπίπεδο, όπως φαίνεται στην (Σχήμα 2.4). Λαμβάνοντας υπόψη ότι το w είναι διάνυσμα κάθετο στο υπερεπίπεδο, έχουμε:

$$f(x) = f(x_p + \frac{w}{\|w\|} r) = w^T x_p + w^T \frac{w}{\|w\|} r + b = \|w\| r \quad (\text{όπου } w^T x_p + b = 0) \qquad \text{Εξ. 2.7}$$

$$\text{Επομένως: } r = \frac{f(x)}{\|w\|} \qquad \text{Εξ. 2.8}$$

Το εύρος του περιθωρίου ρ είναι:

$$p = \frac{f(x_{\oplus})}{\|w\|} - \frac{f(x_{\ominus})}{\|w\|} = \frac{1}{\|w\|} - \frac{-1}{\|w\|} = \frac{2}{\|w\|} \qquad \text{Εξ. 2.9}$$

Επομένως η μεγιστοποίηση του περιθωρίου ισοδυναμεί με την ελαχιστοποίηση του $\|w\|$.[22]

Πρόβλημα εκμάθησης των SVM

Έχοντας ένα γραμμικά διαχωρίσιμο σύνολο δεδομένων, ο στόχος των συντελεστών μάθησης *w* και *b* της μηχανής διανυσμάτων υποστήριξης $f(x) = \left(w^T x + b\right)$ είναι η επίλυση του ακόλουθου προβλήματος βελτιστοποίησης με περιορισμούς:

- Να βρεθούν τα *w* και *b* που ελαχιστοποιούν την $\frac{1}{2}\|w\|^2$ Εξ. 2.10
- Με τον περιορισμό $y_i\left(w^T x_i + b\right) \geq 1, \forall i$ Εξ. 2.11

Αυτό το πρόβλημα βελτιστοποίησης μπορεί να λυθεί με τη χρήση της lagrangian συνάρτησης που ορίζεται ως

$$L(w,b,a) = \frac{1}{2} w^T w - \sum_{i=1}^{N} a_i \left[y_i \left(w^T x + b \right) - 1 \right], \text{ με } a_i \geq 0, \forall i \qquad \text{Εξ. 2.12}$$

όπου $(a_1, a_2, \ldots, a_N)$ είναι πολλαπλασιαστές Lagrange και $\alpha = [a_1, a_2, \ldots, a_N]^T$.

Η λύση του προβλήματος βελτιστοποίησης με περιορισμούς καθορίζεται από το σαγματικό σημείο της $L(w,b,\alpha)$ που πρέπει να ελαχιστοποιηθεί ως προς τα *w* και *b* και να μεγιστοποιηθεί ως προς το a.

- Εάν $y_i\left(w^T x + b\right) > 1$, η τιμή $\alpha_i = 0$ μεγιστοποιεί την $L(w,b,\alpha)$
- Εάν $y_i\left(w^T x + b\right) < 1$, η τιμή $\alpha_i = +\infty$ μεγιστοποιεί την $L(w,b,\alpha)$

Εφόσον τα w και b θέλουμε να ελαχιστοποιούν την $L(w,b,\alpha)$, πρέπει να αλλάξουν με τέτοιον τρόπο, ώστε $y_i\left(w^T x+b\right)=1$

- Οι Kuhn Tucker συνθήκες είναι: $\alpha_i\{y_i(w^T x+b)-1\}=0, \forall i$

Τα σημεία x_i με $\alpha_i>0$ είναι τα **διανύσματα υποστήριξης**.

Οι απαραίτητες συνθήκες για το σαγματικό σημείο της $L(w,b,\alpha)$ είναι

$$\frac{\partial L}{\partial w_j}=0,\forall j \qquad \frac{\partial L}{\partial a_i}=0,\forall i \qquad \text{Εξ. 2.13}$$

ή διαφορετικά $\nabla_w L=0$ $\nabla_a L=0$

βρίσκοντας τα αποτελέσματα με τις απαραίτητες συνθήκες στο

$$w=\sum_{i=1}^{N} a_i y_i x_i \qquad \text{Εξ. 2.14}$$

$$\sum_{i=1}^{N} a_i y_i=0 \qquad \text{Εξ. 2.15}$$

Με αντικατάσταση της εξίσωσης (2.14) στη Lagrangian και τη χρήση της εξίσωσης (2.15) ως νέο περιορισμό, το δυικό πρόβλημα βελτιστοποίησης μπορεί να εκφραστεί ως εξής:

Να βρεθεί το $\boldsymbol{\alpha}$ για το οποίο μεγιστοποιείται η παράσταση

$$\sum_i a_i-\frac{1}{2}\sum_i\sum_j a_i a_j y_i y_j x_i^T x_j \qquad \text{Εξ. 2.16}$$

με τη συνθήκη

$$\sum_{i=1}^{N} a_i y_i=0,\ a_i\geq 0,\forall i \qquad \text{Εξ. 2.17}$$

που αποτελεί ένα πρόβλημα κύρτωσης τετραγωνικού προγραμματισμού με υπαρκτό σφαιρικό ελάχιστο, που μπορεί να επιλυθεί από διάφορες ρουτίνες βελτιστοποίησης. Η βελτιστοποίηση μπορεί να επιλυθεί σε $O(N^3)$ χρόνο (κυβική πολυπλοκότητα ως προς το μέγεθος των στοιχείων κατάρτισης) και σε γραμμικό χρόνο ως προς τον αριθμό ιδιοτήτων.

Δεδομένων των τιμών (a_1,a_2,…,a_N) που λαμβάνονται από την επίλυση, ο τελικός SVM predictor μπορεί να εκφραστεί από την εξίσωση (Εξ.2.14) ως

$$f(x) = w^T x_i + b = \sum_{i=1}^{N} a_i y_i x_i^T x + b \qquad \text{Εξ. 2.18}$$

όπου

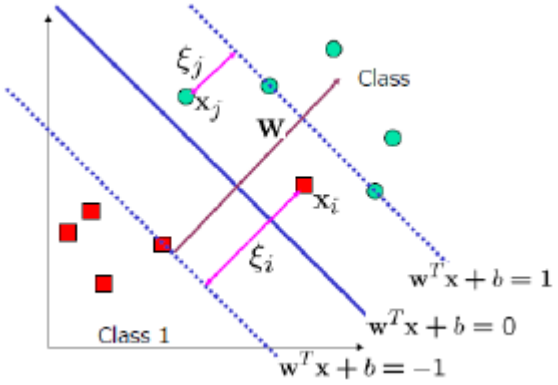
$$b = \frac{1}{\left| I_{support} \right|} \sum_{i \in I_{support}} \left(y_i - \sum_j a_j y_j x_j^T x_i \right) \qquad \text{Εξ. 2.19}$$

και $I_{support}$ είναι το σύνολο των διανυσμάτων υποστήριξης.

- Μη γραμμικά kernel based SVMs

Στις περισσότερες πραγματικές εφαρμογές το σύνολο των δεδομένων δεν είναι γραμμικά διαχωρίσιμο, όπως στην παραπάνω περίπτωση. Εισάγοντας χαλαρές μεταβλητές ξi, i=1,2,…,N, ο περιορισμός $y_i \left(w^T x + b \right) \geq 1$ αναδιατυπώνεται σε

$$y_i \left(w^T x + b \right) \geq 1 - \xi_i, \ \xi_i \geq 0 \qquad \text{Εξ. 2.20}$$



Σχήμα 2.5 Μη γραμμικά διαχωρίσιμα δεδομένα

Στην ιδανική περίπτωση που όλες οι χαλαρές μεταβλητές είναι ίσες με μηδέν, υπάρχει πλήρης ισοδυναμία με την γραμμικά διαχωρίσιμη περίπτωση της προηγούμενης παραγράφου. Στην περίπτωση των μη γραμμικά ταξινομημένων στοιχείων, το πρόβλημα μπορεί να γραφεί ως εξής:

Να βρεθούν τα w και b που ελαχιστοποιούν την $\frac{1}{2}\|w\|^2 + c\sum_i \xi_i^k$ Εξ. 2.21

με τη συνθήκη $y_i\left(w^T x_i + b\right) \geq 1 - \xi_i$, $\boldsymbol{\xi_i \geq 0}$, $\forall \boldsymbol{i}$ Εξ. 2.22

όπου $c>0$ είναι μία κατάλληλα επιλεγμένη παράμετρος.

Ο πρόσθετος όρος $c\sum_i \xi_i^k$ αναγκάζει όλες τις χαλαρές μεταβλητές, να έλθουν όσο το δυνατόν κοντύτερα στο μηδέν. Στην περίπτωση που k=0, ο δεύτερος όρος της εξίσωσης (Εξ. 2.21) μετράει τον αριθμό των λαθών κατάρτισης. Στην περίπτωση που k=2, έχουμε μεγαλύτερη ευαισθησία στα απομακρυσμένα στοιχεία (outliers).

Το παραπάνω πρόβλημα βελτιστοποίησης μπορεί να μετατραπεί σε δυϊκό πρόβλημα, γράφοντάς το στην ακόλουθη μορφή:

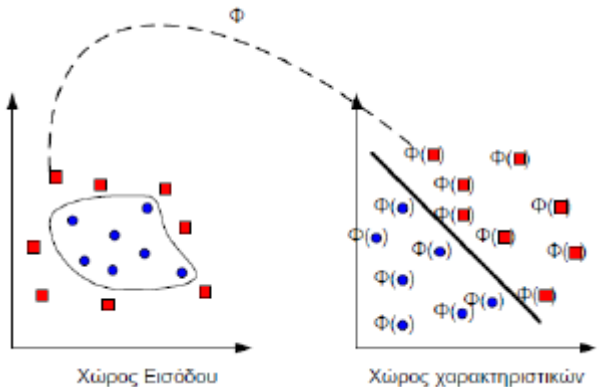
Να βρεθεί το α που ελαχιστοποιεί την $\sum_i a_i - \frac{1}{2}\sum_i \sum_j a_i a_j y_i y_j x_i^T x_j$ Εξ. 2.23

με τις συνθήκες $\sum_{i=1}^{N} a_i y_i = 0$, $0 \leq a_i \leq c, \forall i$ Εξ. 2.24

Η εισαγωγή της παραμέτρου c έχει ως σκοπό τον περιορισμό της σειράς των αποδεκτών τιμών των πολλαπλασιαστών Lagrange α_i. Το c αποτελεί μία παράμετρο κανονικοποίησης που ελέγχει την αλληλεπίδραση μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίησης του όρου λάθους κατάρτισης. Το καταλληλότερο c εξαρτάται από το εκάστοτε σύνολο διαθέσιμων στοιχείων. Εάν είναι πολύ μικρός αριθμός, τότε χρειάζεται προσαρμογή στον διαχωρισμό των δεδομένων κατάρτισης. Στην αντίθετη περίπτωση, που είναι πολύ μεγάλο, τότε ο αλγόριθμος παρουσιάζει επικαλύψεις στα στοιχεία εκπαίδευσης.

Το χρήσιμο εύρος του c βρίσκεται μεταξύ της τιμής για την οποία όλοι οι πολλαπλασιαστές Lagrange ισούνται με c και της τιμής για την οποία μόνο ένας πολλαπλασιαστής Lagrange έχει όριο το c. Μια απλή και πολύ διαδεδομένη μέθοδος προσδιορισμού του c είναι η μέθοδος της διασταύρωσης (cross-validation). Στην περίπτωση που $\alpha_i=c$, εάν $\xi_i>1$ τότε τα σημεία είναι λάθος κατηγοριοποιημένα, εάν $0<\xi_i\leq 1$ τότε τα σημεία είναι σωστά κατηγοριοποιημένα, αλλά σε απόσταση από το βέλτιστο υπερεπίπεδο μικρότερη από το μισό του περιθωρίου $\rho/2$, ή σε πολύ σπάνιες περιπτώσεις εάν $\xi_i=0$ τότε τα στοιχεία είναι σωστά κατηγοριοποιημένα και πάνω στο περιθώριο. Σχεδόν σε κάθε περίπτωση, τα διανύσματα υποστήριξης για τα οποία $\alpha_i=c$ αναφέρονται ως σφάλματα.

Το σύνολο των δεδομένων εισόδου D πολλές φορές μπορεί να μην είναι γραμμικά διαχωρίσιμο στον χώρο εισόδου, αλλά να είναι γραμμικά διαχωρίσιμο σε έναν άλλο χώρο διαφορετικής (κατά κανόνα μεγαλύτερης) διάστασης. Βάση του θεωρήματος του Cover για το διαχωρισμό των δεδομένων (Cover, 1965) γίνεται ο μετασχηματισμός του χώρου εισόδου σε έναν νέο χώρο μεγαλύτερης διάστασης που το D' είναι γραμμικά διαχωρίσιμο και ονομάζεται χώρος χαρακτηριστικών. Στο νέο αυτό χώρο διαχωρίζονται οι «εικόνες» των προτύπων (δεδομένων εισόδου).



Σχήμα 2.6 Μετασχηματισμός του χώρου εισόδου σε χώρο χαρακτηριστικών

Ο μετασχηματισμός γίνεται με τη βοήθεια μίας απεικόνισης $\Phi:\Re^M\rightarrow F^N$ από τον αρχικό χώρο M-διαστασεων σε έναν χώρο μεγαλύτερης διάστασης ιδιοτήτων F.

$$X=(x_1,...,x_p)\mapsto \varphi(x)=(\varphi_1,...,\varphi_p)$$
$$F=\{\varphi(x):x\in X\}$$

Εξ. 2.25

Με βάση τη συνάρτηση Φ (2.25) το δυϊκό πρόβλημα (2.23 – 2.24) μετασχηματίζεται σε:

Να βρεθεί α που να μεγιστοποιεί την $\sum_i a_i - \frac{1}{2}\sum_i \sum_j a_i a_j y_i y_j \Phi(x_i)^T \Phi(x_j)$ Εξ. 2.26

υπό συνθήκες $\begin{array}{l} \sum_{i-1}^{N} a_i y_i = 0 \\ 0 \le a_i \le c, \forall i \end{array}$ Εξ. 2.27

Ο τελικός predictor μπορεί να εκφραστεί και ως

$$f(x) = w^T \Phi(x_i) + b = \sum_{i=1}^{N} a_i y_i \Phi(x_i)^T \Phi(x) + b \qquad \text{Εξ. 2.28}$$

Αρκεί λοιπόν να βρεθεί μια έκφραση για το εσωτερικό γινόμενο στο χώρο ιδιοτήτων (χαρακτηριστικών) που να χρησιμοποιεί μόνο τα σημεία του χώρου εισόδου. Δηλαδή:

$$\Phi(x_i)\Phi(y_j) = \mathrm{K}(x_i, y_j) \qquad \text{Εξ. 2.29}$$

Η συμμετρική συνάρτηση *K* ονομάζεται πυρήνας (kernel).

Σύμφωνα με το θεώρημα Mercer, υπάρχει μια κατηγορία απεικόνισης *Φ* με την ακόλουθη ιδιότητα:

$$\Phi(x)\Phi(y) = \mathrm{K}(x, y) \qquad \text{Εξ. 2.30}$$

όπου *K* μία αντίστοιχη συνάρτηση kernel.

Το δυϊκό πρόβλημα (2.26 – 2.27) μπορεί πλέον να γραφεί ως

Να βρεθεί α που να μεγιστοποιεί την $\sum_i a_i - \frac{1}{2}\sum_i \sum_j a_i a_j y_i y_j \mathrm{K}(x_i, y_j)$ Εξ. 2.31

υπό συνθήκες $\begin{array}{l} \sum_{i-1}^{N} a_i y_i = 0 \\ 0 \le a_i \le c, \forall i \end{array}$ Εξ. 2.32

και αντίστοιχα ο SVM predictor είναι ο εξής:

$$f(x) = w^T \Phi(x_i) + b = \sum_{i=1}^{N} a_i y_i K(x_i, y_j) + b \qquad \text{Εξ. 2.33}$$

Μόνη απαίτηση για τον πυρήνα $K(x,x_i)$ είναι να ικανοποιεί το θεώρημα Mercer. Κατά συνέπεια υπάρχει αρκετή ελευθερία σχετικά με την επιλογή του. Στον Πίνακα 2.1 φαίνονται οι τρεις συνηθέστεροι πυρήνες που χρησιμοποιούνται σε SVM [21].

Πίνακας 2.1 Οι συνηθέστεροι πυρήνες εσωτερικού γινομένου σε SVM

Τύπος του SVM δικτύου	Πυρήνας Εσωτερικού Γινομένου $K(x,x_i)$	Παρατηρήσεις
Γραμμική	$x^T x_i$	Η δύναμη p καθορίζεται εκ των προτέρων
Πολυωνυμική μηχανή μάθησης	$(x^T x_i + 1)^p$	
Δίκτυο Ακτινικών Συναρτήσεων Βάσης	$\exp(-\frac{\gamma}{2\sigma^2}\lvert x - x_i\rvert^2)$	Το πλάτος σ^2 είναι κοινό για όλους τους πυρήνες και καθορίζεται εκ των προτέρων και $\gamma>0$
Διεπίπεδο Perceptron	$\tanh(\beta_0 x^T x_i + \beta_1)$	Το θεώρημα του Mercer δεν ικανοποιείται για όλες τις τιμές των β_0 και β_1

2.4. Δέντρα απόφασης (decision trees)

Τα Δέντρα Απόφασης ή Ταξινόμησης (Decision or Classification Trees) είναι ένα δημοφιλές εργαλείο για κατηγοριοποίηση δεδομένων. Είναι ουσιαστικά δεντροειδές δομές που ορίζουν μια ακολουθία αποφάσεων. Αυτές οι αποφάσεις παράγουν τους κανόνες για την κατηγοριοποίηση ενός συνόλου δεδομένων. Τα δέντρα αποφάσεων έχουν διάφορα πλεονεκτήματα, όπως το ότι είναι εύκολο να τα καταλάβουμε, μπορούν να μετασχηματιστούν σε κανόνες και πειραματικά έχει αποδειχθεί ότι λειτουργούν πολύ καλά.

Οι απαιτήσεις για τη σωστή ανάπτυξη ενός αλγορίθμου κατασκευής δέντρων αποφάσεων περιλαμβάνουν την ύπαρξη κατηγοριών δεδομένων και την ύπαρξη ενός συνόλου δεδομένων ήδη κατηγοριοποιημένων. Ο μηχανισμός αυτός κατασκευάζει δέντρα από ένα σύνολο παραδειγμάτων και η ποιότητά τους εξαρτάται από την ακρίβεια της

κατηγοριοποίησης, καθώς επίσης και από το μέγεθος του δέντρου. Είναι μια μέθοδος που υλοποιείται σε δύο φάσεις:

1. **Φάση κατασκευής του δέντρου:** Τα δεδομένα εκπαίδευσης χωρίζονται αναδρομικά με βάση κάποιο χαρακτηριστικό τους μέχρις ότου όλα τα στιγμιότυπα μιας ομάδας να ανήκουν στην ίδια κατηγορία.
2. **Φάση κλαδέματος του δέντρου:** Κάποιοι κόμβοι του δέντρου «κλαδεύονται» για να αποφευχθεί το overfitting του δέντρου, χωρίς όμως να επηρεάζεται σημαντικά η ακρίβεια της κατηγοριοποίησης.

Παρακάτω ακολουθούν κάποια ζητήματα σχετικά με την διαδικασία κατασκευής των δένδρων αποφάσεων που λαμβάνονται υπόψη από τους περισσότερους αλγορίθμους κατασκευής.

- Επιλογή των χαρακτηριστικών διάσπασης: Το ποια χαρακτηριστικά χρησιμοποιούνται ως χαρακτηριστικά διάσπασης, επηρεάζει την απόδοση του δένδρου απόφασης αφού μερικά χαρακτηριστικά είναι καλύτερα από τα άλλα. Η επιλογή του χαρακτηριστικού περιλαμβάνει όχι μόνο την εξέταση των δεδομένων του συνόλου εκπαίδευσης, αλλά και την εμπεριστατωμένη άποψη των ειδικών του συγκεκριμένου τομέα.
- Διάταξη των χαρακτηριστικών διάσπασης: Η σειρά με την οποία επιλέγονται τα χαρακτηριστικά διάσπασης είναι κάτι πολύ σημαντικό.
- Διασπάσεις: Ο αριθμός των διασπάσεων που έχουμε σχετίζεται με τη διάταξη των χαρακτηριστικών. Σε μερικά χαρακτηριστικά, το πεδίο είναι μικρό, πράγμα που σημαίνει ότι ο αριθμός των διασπάσεων είναι μικρός.Αντίθετα, αν το πεδίο είναι συνεχές ή έχει μεγάλο πλήθος διαφορετικών τιμών, ο αριθμός των διασπάσεων που θα γίνουν δεν είναι απλή διαδικασία.
- Δομή του δένδρου: Για να έχουμε καλύτερη απόδοση στην κατηγοριοποίηση, είναι επιθυμητό να έχουμε ένα ισοζυγισμένο δένδρο απόφασης με τα λιγότερα δυνατά επίπεδα. Ωστόσο, κάτι τέτοιο ίσως απαιτούσε πολύπλοκες συγκρίσεις με πολλές διακλαδώσεις.
- Κριτήρια τερματισμού: Έχει ήδη αναφερθεί ότι ο κάθε αλγόριθμος παραγωγής ενός δένδρου απόφασης έχει διαφορετικό κριτήριο τερματισμού. Η κατασκευή του δένδρου, όπως είναι φυσικό, τερματίζει όταν τα δεδομένα εκπαίδευσης δοκιμάζονται και κατηγοριοποιούνται τέλεια. Ωστόσο, ένα μεγάλο δένδρο απόφασης ίσως δεν είναι τόσο αποδοτικό. Έτσι, υπάρχουν περιπτώσεις που σταματάμε την κατασκευή του δένδρου. Κάτι τέτοιο αποτελεί συμβιβασμό μεταξύ ακρίβειας στην κατηγοριοποίηση και στην απόδοση. Επίσης, είναι επιθυμητό να σταματήσουμε την ανάπτυξη του δένδρου ώστε να αποφύγουμε φαινόμενα υπερπροσαρμογής.
- Δεδομένα εκπαίδευσης: Η δομή του δένδρου απόφασης εξαρτάται στο μεγαλύτερο ποσοστό, στα δεδομένα εκπαίδευσης που χρησιμοποιούνται.

Αν το σύνολο αυτό είναι πολύ μικρό, τότε ίσως το δένδρο να μην είναι αρκετά συγκεκριμένο ώστε να μπορεί να εφαρμοστεί σε γενικά δεδομένα. Από την άλλη πλευρά, αν το σύνολο εκπαίδευσης είναι μεγάλο, τότε υπάρχουν αυξημένες πιθανότητες να έχουμε φαινόμενα υπερπροσαρμογής.

Τα βασικά χαρακτηριστικά ενός δέντρου αποφάσεων είναι:

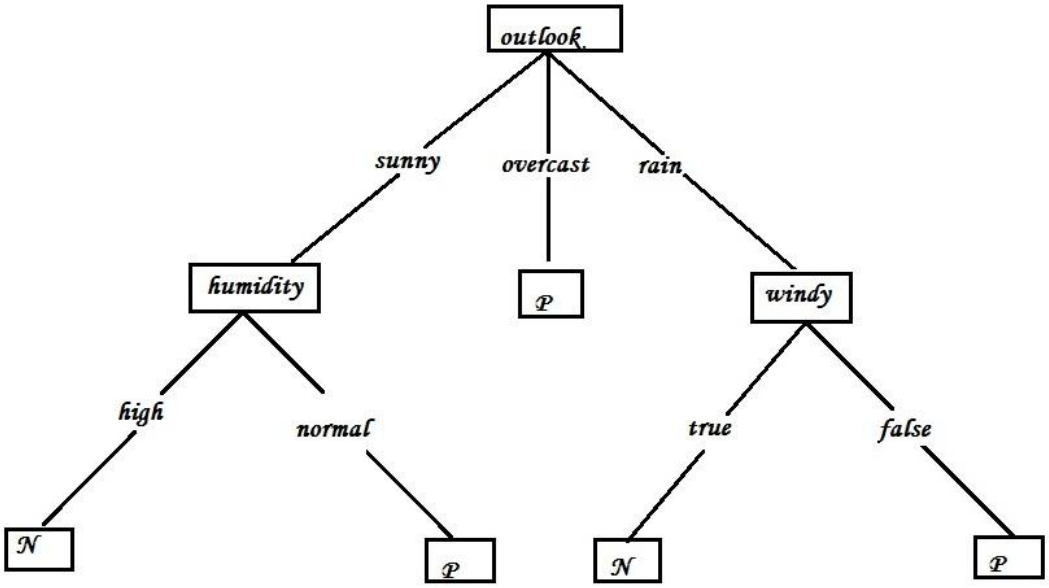
- Ρίζα: Το γνώρισμα που επιλέγεται ως η βάση, πάνω στην οποία χτίζεται το δέντρο.
- Εσωτερικός κόμβος: Ένα γνώρισμα το οποίο βρίσκεται στο εσωτερικό του δέντρου.
- Κλάδος: Μια από τις πιθανές τιμές του γνωρίσματος που βρίσκεται στον κόμβο από τον οποίο ξεκινά ο κλάδος.
- Φύλλο: Μια από τις καθορισμένες κλάσεις.

Η γενική ιδέα στα δέντρα απόφασης είναι η εξής:

Αρχικά, βασική προϋπόθεση αποτελεί το ότι κάθε δείγμα του συνόλου δεδομένων (data set) μπορεί να εκφράζεται ως μια συλλογή από τα χαρακτηριστικά (attributes) του data set. Αλγόριθμος λαμβάνει ως είσοδο κάθε case ως ένα input vector με τις τιμές των μεταβλητών και την αντίστοιχη κατηγορία που ανήκει. π.χ σε ένα data set, ένας ασθενής είναι μια case που εκφράζεται από τα attributes φύλλο και ηλικία.
Στη συνέχεια, παρατηρώντας και συγκρίνοντας αν τα vectors, των οποίων οι τιμές είναι κοντά η μία στην άλλη, ανήκουν ή όχι στην ίδια κατηγορία, κατασκευάζει ένα σετ από κανόνες αποφάσεων (decision rules) με σκοπό την κατηγοριοποίηση μελλοντικών δειγμάτων (test set) στις γνωστές κατηγορίες.

Το σημείο όπου γίνεται η υπόθεση για μια μεταβλητή είναι αυτό στο οποίο η μεταβλητή χωρίζεται (split) ανάμεσα σε δύο τιμές και ταυτόχρονα χωρίζει το training set σε δύο subsets.
Σχηματικά, στην απεικόνιση του δέντρου το σημείο αυτό ονομάζεται «κόμβος». Από κάθε κόμβο γεννιούνται δύο subsets-κλαδιά (branches) με την αντίστοιχη απόφαση ή ένα κλαδί το οποίο καταλήγει σε μία κατηγορία. Όταν καταλήξουμε σε μία κατηγορία, αυτή βρίσκεται σε ένα «φύλλο» (leaf). Σχηματικά έχει τη μορφή ενός «αντίστροφου δέντρου»

(τα φύλλα κάτω και η ρίζα στην κορυφή), γι' αυτό και ο αλγόριθμος έχει αυτήν την ονομασία - «Δέντρα Αποφάσεων».



Σχήμα 2.7 Ένα απλό Decision Tree

Στο παραπάνω σχήμα απεικονίζεται ένα απλό Decision Tree. Τα χαρακτηριστικά (attributes) του οποίου αποτέλεσαν ο καιρός, η θερμοκρασία, η υγρασία και ο άνεμος. Κάθε χαρακτηριστικό έχει το δικό του σύνολο τιμών:

Καιρός: {sunny, overcast, rain}
Θερμοκρασία: {cool, mild, hot}
Υγρασία: {high, normal}
Άνεμος: {true, false} or {strong, weak}

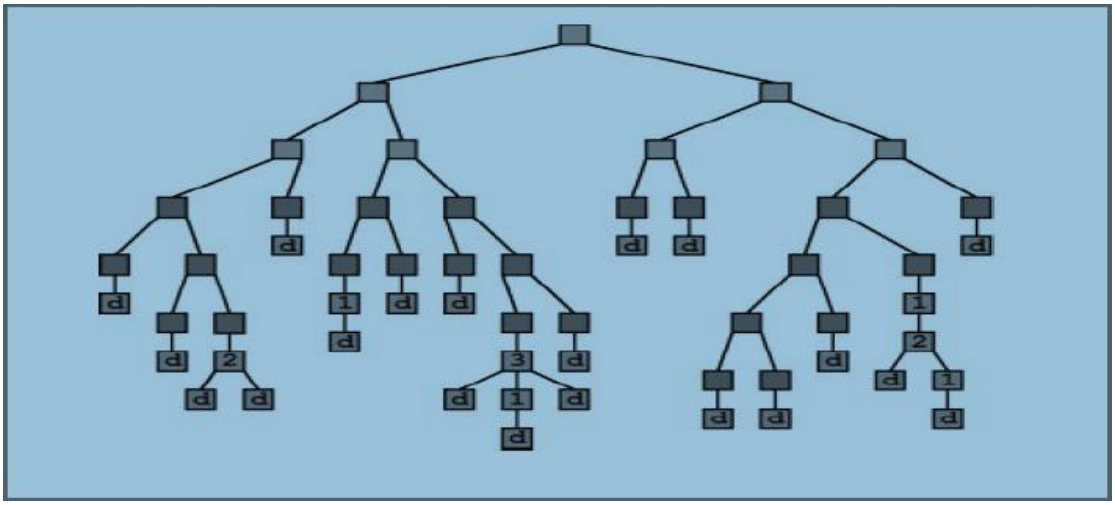
Οπότε μια απλή μέρα (που εκφράζει ένα αντικείμενο) μπορεί να περιγραφεί ως εξής και αποτελεί ένα input vector στον αλγόριθμο:
Νεφελώδης καιρός με χαμηλή θερμοκρασία και υψηλή υγρασία με την ύπαρξη ανέμου.
Όπως περιγράφηκε και παραπάνω κάθε αντικείμενο ενός training set ανήκει σε μια μοναδική κατηγορία, η οποία είναι γνωστή εκ των προτέρων. Στο παράδειγμα αυτό, για απλούστευση υπάρχουν μόνο δύο κατηγορίες P και N (θετικές και αρνητικές περιπτώσεις positive and negative). Σκοπός είναι να αναπτυχθούν κάποιοι κανόνες ταξινόμησης

(Classification rules) που να μπορούν να αποφασίζουν για ένα μελλοντικό αντικείμενο σε ποια από τις δύο κατηγορίες θα ανήκει.

Σε μαθηματικούς όρους ένα δέντρο αποφάσεων είναι μια ιεραρχημένη συλλογή σύνθετων διαζευκτικών προτάσεων οι οποίες αποτελούνται από ένα σύνολο λογικών συζεύξεων που αναφέρονται σε τιμές χαρακτηριστικών συγκεκριμένων παραδειγμάτων.

Τα δέντρα είναι εύκολα στη χρήση και αποτελεσματικά. Μπορούν να δημιουργηθούν κανόνες οι οποίοι είναι εύκολοι στο να κατανοηθούν και να ερμηνευθούν. Τα δέντρα αποφάσεων αποδίδουν καλά για μεγάλα σύνολα παραδειγμάτων. Κάθε παράδειγμα του συνόλου παραδειγμάτων πρέπει να φιλτραριστεί μέσα από το δέντρο. Αυτό παίρνει χρόνο ανάλογο με το ύψος του δέντρου, το οποίο είναι συγκεκριμένο. Μπορούμε να κατασκευάσουμε δέντρα για δεδομένα με άλλα γνωρίσματα.
Υπάρχουν όμως και μειονεκτήματα για τους αλγορίθμους δέντρων αποφάσεων. Πρώτον, δε χειρίζονται εύκολα συνεχή δεδομένα. Αυτά τα πεδία των γνωρισμάτων θα πρέπει να χωριστούν σε κατηγορίες για να μπορέσει να τα χειριστεί το δέντρο. Η προσέγγιση που χρησιμοποιείται είναι ότι ο χώρος του πεδίου διαιρείται σε ορθογώνιες περιοχές. Βέβαια δεν είναι όλα τα προβλήματα κατηγοριοποίησης αυτού του τύπου. Επίσης, ο χειρισμός που γίνεται στα ελλιπή δεδομένα είναι δύσκολος γιατί δεν μπορούν να βρεθούν σωστές διακλαδώσεις του δέντρου για να ακολουθηθούν. Επειδή το δέντρο απόφασης δημιουργείται από τα δεδομένα εκπαίδευσης, μπορεί να εμφανιστεί υπερπροσαρμογή. Αυτό θα μπορούσαμε να το προσπεράσουμε με τη φάση περικοπής του δέντρου. Τέλος οι συσχετίσεις μεταξύ των γνωρισμάτων του συνόλου παραδειγμάτων αγνοούνται από τη διαδικασία του δέντρου απόφασης [21].

Αξίζει να σημειωθεί ότι υπάρχουν διάφοροι τρόποι κατασκευής δέντρων αποφάσεων, οι οποίοι χρησιμοποιούν διαφορετικούς αλγόριθμους επιλογής του κριτηρίου διαχωρισμού του συνόλου των δεδομένων. Ωστόσο, η γενική μορφή ενός δέντρου απόφασης θα μπορούσε να ήταν όπως φαίνεται παρακάτω, στο Σχ. 2.6.



Σχήμα 2.8 Τυπική απεικόνιση δημιουργίας ενός δέντρου απόφασης.

2.4.1. Ο αλγόριθμος ID3

Μια από της διαδεδομένες και ταυτόχρονα απλές τεχνικές που χρησιμοποιείται για την κατασκευή δένδρων αποφάσεων είναι ο αλγόριθμος ID3. Αυτό που προσπαθεί να επιτύχει αυτός ο αλγόριθμος είναι να ελαχιστοποιήσει τον αριθμό των συγκρίσεων. Η βασική ιδέα ενός αλγορίθμου επαγωγής είναι να κάνει ερωτήσεις των οποίων οι απαντήσεις να περιέχουν την περισσότερη πληροφορία. Λέγοντας περισσότερη πληροφορία εννοούμε ερωτήσεις που απορρίπτουν μεγάλο μέρος του χώρου αναζήτησης. Η βασική ιδέα του αλγορίθμου ID3 είναι η επιλογή χαρακτηριστικών διάσπασης που περιέχουν μεγαλύτερο κέρδος πληροφορίας. Το ποσό της πληροφορίας, το οποίο σχετίζεται με την τιμή ενός χαρακτηριστικού, εξαρτάται από την πιθανότητα εμφάνισης του.

Η έννοια που χρησιμοποιείται για να μετρηθεί η πληροφορία καλείται εντροπία (Entropy). Χρησιμοποιούμε το μέτρο της εντροπίας ώστε να μετρήσουμε το πόσο ανομοιογενές είναι ένα σύνολο δεδομένων. Το μέτρο αυτό παίρνει τιμές στο διάστημα [0,1].

Ο τυπικός ορισμός της εντροπίας παρουσιάζεται στον παρακάτω ορισμό.

Ορισμός: Με δεδομένες τις πιθανότητες p_1, p_2, …, p_s με $\sum_{i=1}^{s} p_i = 1$ η εντροπία ορίζεται ως:

$$H\left(p_1, p_2, ..., p_s\right) = \sum_{i=1}^{s}\left(p_i \log\left(\frac{1}{p_i}\right)\right) \qquad \text{Εξ. 2.34}$$

Δεδομένης μια κατάστασης D, το H(D) , βρίσκει την ποσότητα της τάξης σε αυτή την κατάσταση. Όταν η κατάσταση D, χωρίζεται σε s νέες καταστάσεις S ={D1,D2,…,Ds}, το μέτρο της εντροπίας μπορεί να εφαρμοστεί σε κάθε μια από αυτές τις νέες καταστάσεις. Κάθε βήμα του ID3 επιλέγει την κατάσταση η οποία διατάσσει περισσότερο τη διάσπαση. Μια κατάσταση του συνόλου παραδειγμάτων είναι απολύτως διατεταγμένη αν όλα τα παραδείγματα σε αυτήν ανήκουν στην ίδια κατηγορία. Ο ID3 επιλέγει το χαρακτηριστικό διάσπασης με το μεγαλύτερο κέρδος πληροφορίας. Κέρδος (gain) πληροφορίας μετρά την μείωση της εντροπίας που θα προκληθεί αν χωριστεί το σύνολο δεδομένων με βάση κάποιο χαρακτηριστικό. Καταλήγοντας, Ο ID3 αλγόριθμος υπολογίζει το κέρδος μιας διάσπασης χρησιμοποιώντας τον εξής τύπο:

$$Gain\left(D, S\right) = H(D) - \sum_{i=1}^{s} P(D_i)H(D_i) \qquad \text{Εξ. 2.35}$$

Ο πρώτος όρος της διαφοράς είναι η εντροπία του συνόλου δεδομένων ενώ ο δεύτερος όρος είναι η εντροπία των δεδομένων μετά τη διάσπαση τους ανάλογα με την τιμή του χαρακτηριστικού S. Ο δεύτερος όρος αποτελείται από το άθροισμα της εντροπίας για το κάθε σύνολο που προκύπτει μετά τη διάσπαση. Μια γενική περιγραφή του ID3 παρουσιάζεται παρακάτω:

1. Αρχικά πρέπει να επιλεχτεί το πιο κατάλληλο χαρακτηριστικό για έλεγχο στη ρίζα.
2. Στη συνέχεια, για κάθε δυνατή τιμή του χαρακτηριστικού δημιουργούνται οι αντίστοιχοι απόγονοι της ρίζας. Τα δεδομένα μοιράζονται στους νέους κόμβους ανάλογα με την τιμή που έχουν για το χαρακτηριστικό που ελέγχεται στη ρίζα.
3. Η όλη διαδικασία επαναλαμβάνεται για κάθε νέο κόμβο. Η επιλογή του χαρακτηριστικού θα γίνει βάσει των δεδομένων που ανήκουν στον κάθε κόμβο.
4. Ένας κόμβος γίνεται φύλλο όταν όλα τα δεδομένα που ανήκουν σε αυτόν ανήκουν στην ίδια κατηγορία (αμιγής κόμβος). Η κατηγορία αυτή γίνεται και η τιμή του φύλλου.

5. Αν σε κάποιο βάθος τελειώσουν τα χαρακτηριστικά προς έλεγχο, τότε ο κόμβος γίνεται τερματικός και σαν τιμή παίρνει εκείνη που έχει την πλειοψηφία με βάση τα δεδομένα του κόμβου αυτού.

2.4.2. Ο αλγόριθμος C4.5 και C5.0

Ο Αλγόριθμος C4.5 βελτιώνει τον αλγόριθμο ID3. Συγκεκριμένα:

- Ελλιπή δεδομένα: Όταν το δένδρο απόφασης χτίζεται, τα ελλιπή δεδομένα αγνοούνται. Αυτό σημαίνει ότι το κέρδος υπολογίζεται λαμβάνοντας υπόψη μόνο στις εγγραφές που έχουν τιμή. Για να κατηγοριοποιήσουμε ένα παράδειγμα με ελλιπή τιμή σε ένα χαρακτηριστικό, η τιμή αυτή μπορεί να προβλεφτεί με βάση των υπόλοιπων τιμών αυτού του χαρακτηριστικού.
- Συνεχή δεδομένα: Τα χαρακτηριστικά που παίρνουν συνεχής τιμές, χωρίζονται σε διαστήματα.
- Κλάδεμα: Υπάρχουν δύο σημαντικές στρατηγικές κλαδέματος στον C4.5:
    - Αντικατάσταση του υποδένδρου (subtree replacement): ένα υποδένδρο αντικαθιστάται από ένα φύλλο αν αυτή η αντικατάσταση έχει ως αποτέλεσμα σφάλμα κοντά σε αυτό του αρχικού υποδένδρου. Η τεχνική αυτή εφαρμόζεται ξεκινώντας από τα φύλλα και ανεβαίνοντας προς τη ρίζα.
    - Ανύψωση υποδένδρου (subtree raising): αντικαθιστά ένα υποδένδρο με το περισσότερο χρησιμοποιούμενο υποδένδρο του. Έτσι ένα υποδένδρο ανυψώνεται αφού αντικαθιστά ένα υποδένδρο που βρίσκεται σε ψηλότερο επίπεδο. Και σε αυτή την περίπτωση πρέπει να λάβουμε υπόψη την αύξηση στη συχνότητα λαθών.

- Κανόνες: Ο C4.5 επιτρέπει την κατηγοριοποίηση είτε μέσω δένδρων αποφάσεων είτε μέσω κανόνων που δημιουργούνται από αυτό. Επίσης, προτείνονται κάποιες τεχνικές που απλουστεύουν τους πολύπλοκους κανόνες.
- Διάσπαση: Ο ID3 προτιμά τα χαρακτηριστικά με πολλές διαιρέσεις. Ωστόσο αυτό μπορεί να οδηγήσει σε υπερπροσαρμογή. Μια οριακή περίπτωση είναι να έχουμε ένα χαρακτηριστικό που έχει μια μοναδική τιμή για κάθε παράδειγμα. Το χαρακτηριστικό αυτό θα είναι το καλύτερο αφού θα υπήρχε μόνο ένα παράδειγμα (και έτσι μόνο μια κατηγορία) για κάθε διαίρεση. Μια βελτίωση θα μπορούσε να γίνει αν λάβουμε υπόψη την πληθικότητα της κάθε διαίρεσης. Αυτή η προσέγγιση χρησιμοποιεί το GainRatio και όχι το Gain.

$$GainRatio(D,S) = \frac{Gain(D,S)}{H(\frac{|D_1|}{|D|},...,\frac{|D_s|}{|D|})} \qquad \text{Εξ. 2.36}$$

Για την διάσπαση, ο C4.5 χρησιμοποιεί το μεγαλύτερο GainRatio το οποίο εξασφαλίζει ένα μεγαλύτερο από το μέσο όρο κέρδος στην πληροφορία. Αυτό αντισταθμίζει το γεγονός ότι η τιμή του GainRatio κλίνει προς διασπάσεις όπου το μέγεθος του ενός υποσυνόλου είναι κοντά προς αυτό του αρχικού.

Ο C5.0 είναι μια εμπορική έκδοση του C4.5 που χρησιμοποιείται πάρα πολύ στα πακέτα λογισμικού εξόρυξης γνώσης. Ο προσανατολισμός του είναι προς τη χρήση μεγάλων συνόλων δεδομένων. Η φάση της επαγωγής είναι όμοια με αυτή του C4.5 αλλά η δημιουργία κανόνων είναι διαφορετική. Σε αντίθεση με τον C4.5, οι αλγόριθμοι που χρησιμοποιούνται στον C5.0 δεν έχουν γίνει γνωστοί. Τα αποτελέσματα που έχουν δημοσιευτεί, δείχνουν ότι ο C5.0 βελτιώνει την χρήση της μνήμης κατά 90%, τρέχει πολύ πιο γρήγορα από τον C4.5 και παράγει πιο ακριβείς κανόνες. Μια πολύ σημαντική βελτίωση στην ακρίβεια του C5.0 βασίζεται στην ενίσχυση (boosting) η οποία είναι μια τεχνική που συνδυάζει διάφορους ταξινομητές.

2.5. Ταξινομητής naive bayes

Ο κατηγοριοποιητής Bayes χρησιμοποιείται για να εκτιμήσουμε την πιθανότητα ένα νέο παράδειγμα/ στιγμιότυπο να ανήκει σε μια από τις προκαθορισμένες κατηγορίες. Η απόδοση αυτού του αλγορίθμου συχνά είναι αρκετά υψηλή ενώ συγχρόνως επιτυγχάνει και μεγάλες ταχύτητες.

Η μέθοδος Bayes, έχει ως στόχο την εύρεση της πιο πιθανής υπόθεσης από ένα σύνολο υποθέσεων ή δεδομένου ενός συνόλου εκπαίδευσης D, αλλά και της γνώσης που πιθανόν εκ των προτέρων διαθέτουμε για τις πιθανότητες των διαφόρων υποθέσεων $h \in H$. Για τον προσδιορισμό των πιθανοτήτων αυτών θα χρησιμοποιήσουμε το θεώρημα του Bayes, σύμφωνα με το οποίο:

$$P(h \mid D) = \frac{P(h)P(D \mid h)}{P(D)} \qquad \text{Εξ. 2.37}$$

Όπου:

P(h|D): η ζητούμενη εκ των υστέρων πιθανότητα (a-posteriori probability) που εκφράζει την ισχύ της υπόθεσης h, δεδομένου του συνόλου εκπαίδευσης D.
P(h): η εκ των προτέρων γνωστή πιθανότητα που εκφράζει την ισχύ της υπόθεσης h, χωρίς να έχει προηγηθεί παρατήρηση των δεδομένων του συνόλου εκπαίδευσης D. Η πιθανότητα αυτή ονομάζεται εκ των προτέρων πιθανότητα της h (a-priori probability).
P(D|h): Η δεσμευμένη πιθανότητα η οποία εκφράζει το ενδεχόμενο παρατήρησης των δεδομένων εκπαίδευσης D, αποδεχόμενης της ισχύος της υπόθεσης h. Η πιθανότητα αυτή είναι δυνατόν να υπολογιστεί/ προσεγγιστεί από τη γνώση που διαθέτουμε για το συγκεκριμένο πρόβλημα (περίπτωση Naïve Bayes).
P(D): η εκ των προτέρων πιθανότητα παρατήρησης των δεδομένων εκπαίδευσης D. Ο όρος αυτός απλοποιείται και δε συμμετάσχει στους υπολογισμούς.

Σε αυτό που πρέπει να δώσουμε σημασία είναι ότι στην περίπτωση του αλγορίθμου κατηγοριοποίησης Bayes κάνουμε δύο παραδοχές για τα γνωρίσματα :

Ότι είναι εξίσου σημαντικά.
Ότι είναι στατιστικά ανεξάρτητα (δοθείσης της τιμής μιας κλάσης), δηλαδή αν γνωρίζουμε την τιμή ενός γνωρίσματος δεν μπορούμε να πούμε τίποτα για την τιμή ενός άλλου γνωρίσματος (με δεδομένο ότι γνωρίζουμε την κατηγορία).

Αφελής Ταξινομητής Naïve Bayes

Ας υποθέσουμε ότι η επίδραση της τιμής ενός χαρακτηριστικού σε μια δεδομένη κατηγορία είναι ανεξάρτητη από τις τιμές των υπολοίπων χαρακτηριστικών. Αυτή η υπόθεση είναι γνωστή ως υπό συνθήκη ανεξαρτησία (conditional independence). Τότε οι υπολογισμοί μας γίνονται πιο εύκολα, αλλά έχουμε κάνει μια απλοϊκή υπόθεση γι' αυτό και ο όρος απλοϊκός (naive), η υπόθεση αυτή όμως μπορεί να μην ισχύει πάντα.

'Εστω ότι κάθε στιγμιότυπο είναι ένα n-διάστατο διάνυσμα ανάλογα με τις τιμές των χαρακτηριστικών του.
Άρα:

$D=(D_1,D_2.....D_n)$

και ότι m είναι οι αντίστοιχες κατηγορίες

$h=(h_1,h_2,\ldots,h_m)$

Ένα στιγμιότυπο D ανήκει στην κατηγορία στην οποία έχει τη μεγαλύτερη εκ των υστέρων πιθανότητα (posterior probability). Ένα στιγμιότυπο D θα ανατεθεί στην κατηγορία h_j αν και μόνο αν :

$$P(h_i|D)>p(h_l|D),\ 1\leq l \leq m \text{ για κάθε } 1\neq i$$

Παίρνοντας τώρα υπόψη μας το θεώρημα του Bayes

$$P(h \mid D) = \frac{p(h)\,p(D \mid h)}{p(D)} \qquad \text{Εξ. 2.38}$$

επειδή η p(D) είναι σταθερή για όλες τις κατηγορίες μόνο το $p(D|h_i)p(h_i)$ χρειάζεται να μεγιστοποιηθεί. Επίσης αν δεν γνωρίζουμε τις πιθανότητες της κατηγορίας $p(h_i)$ τότε μπορούμε να υποθέσουμε ότι είναι ίσες, έτσι χρειάζεται να μεγιστοποιηθεί μόνο το $p(D|h_i)$. Εναλλακτικά μπορούμε να υπολογίσουμε το $p(h_i)$ από το σύνολο εκπαίδευσης.

Δοθέντος ενός συνόλου εκπαίδευσης, ο απλοϊκός αλγόριθμος Bayes αρχικά εκτιμά την εκ των προτέρων πιθανότητα $p(h_i)$ για κάθε κατηγορία μετρώντας πόσο συχνά κάθε κατηγορία εμφανίζεται στα δεδομένα εκπαίδευσης. Στα δεδομένα της κατηγορίας h_i, μπορεί να μετρηθεί ο αριθμός των εμφανίσεων κάθε τιμής, για να καθορίσει την $p(d_m|h_i)$. Αυτό πρέπει να γίνει για όλα τα δεδομένα της κατηγορίας h_i και για όλες τις τιμές τους.

Στη συνέχεια χρησιμοποιούμε αυτές τις πιθανότητες που έχουν υπολογιστεί με αυτόν τον τρόπο όταν πρέπει να κατηγοριοποιηθεί ένα παράδειγμα. Αυτός είναι ο λόγος γιατί η απλοϊκή κατηγοριοποίηση κατά Bayes μπορεί να θεωρηθεί σαν ένας τύπος αλγορίθμου που μπορεί να χρησιμοποιηθεί για την περιγραφή και για την πρόβλεψη δεδομένων. Οι πιθανότητες είναι περιγραφικές και στη συνέχεια χρησιμοποιούνται για να προβλέψουν τη συμμετοχή σε μια κατηγορία για το υπό μελέτη παράδειγμα. [21]

2.6. Συλλογική Ταξινόμηση (ensemble)

Πολλοί ερευνητές έχουν διερευνήσει την τεχνική συνδυασμού των προβλέψεων πολλών ταξινομητών για τη δημιουργία ενός ενιαίου ταξινομητή (Breiman-1996, Clemen-1989, Perrone-1993, Wolpert—1992). Ο ταξινομητής που προκύπτει ονομάζεται συλλογικός ταξινομητής (ensemble classifier) και είναι γενικά περισσότερο ακριβής από κάθε ταξινομητή που συμμετέχει στην ομάδα σχηματισμού του. Θεωρητικές και εμπειρικές μελέτες έχουν δείξει ότι ένας αποτελεσματικός συλλογικός ταξινομητής αποτελείται από ταξινομητές οι οποίοι είναι αποδοτικοί και τα σφάλματά τους βρίσκονται σε διαφορετικές περιοχές του διανυσματικού χώρου εισόδου. Οι δύο περισσότερο δημοφιλείς μέθοδοι δημιουργίας συλλογικών ταξινομητών είναι η μέθοδος Bagging και η μέθοδος Boosting. Οι τεχνικές αυτές βασίζονται στη δειγματοληψία του συνόλου εκπαίδευσης για να προκύψουν διαφορετικά σύνολα εκπαίδευσης για κάθε ταξινομητή του συνολικού συστήματος ταξινόμησης [16]

Η βασική ιδέα της συλλογικής ταξινόμησης είναι η στάθμιση διαφορετικών ταξινομητών και ο συνδυασμός τους σε ένα ενιαίο ταξινομητή, ο οποίος αποδίδει καλύτερα από κάθε ένα από τους επιμέρους ταξινομητές. Κατά τη λήψη μιας απόφασης οι άνθρωποι ακολουθούν την ίδια τεχνική, ακούγοντας διάφορες απόψεις και στη συνέχεια αξιολογώντας τις απόψεις αυτές για τη λήψη της τελικής απόφασης.

Για να παράγει ένας πληθυσμός σωστές αποφάσεις, θα πρέπει να πληροί τα ακόλουθα κριτήρια :

- Ποικιλομορφία απόψεων (Diversity of opinion). Κάθε μέλος του πληθυσμού θα πρέπει να έχει ιδιωτική πληροφορία, ακόμα και αν είναι μια υποκειμενική ερμηνεία των γεγονότων.
- Ανεξαρτησία (Independence). Οι απόψεις των μελών δεν θα πρέπει να επηρεάζονται από άλλα μέλη του πληθυσμού.
- Αποκέντρωση (Decentralization). Τα μέλη θα πρέπει να έχουν τη δυνατότητα να καταλήγουν σε συμπεράσματα χρησιμοποιώντας τοπική γνώση.
- Συνυπολογισμός (Aggregation). Υπάρχει κάποιος μηχανισμός για τη μετατροπή των επιμέρους αποφάσεων σε μια συλλογική απόφαση.

Ο συνδυασμός των εξόδων των ταξινομητών είναι χρήσιμος μόνο όταν υπάρχει διαφωνία μεταξύ τους. Όταν συνδυαστούν όμοιοι ταξινομητές που παράγουν όμοια αποτελέσματα, δεν υπάρχει κάποιο κέρδος ως προς την πληροφορία ταξινόμησης. Οι Hansen et al. (1990) απέδειξαν ότι εάν ο μέσος ρυθμός σφάλματος για ένα δείγμα είναι λιγότερο από 50% και οι επιμέρους ταξινομητές είναι ανεξάρτητοι στην παραγωγή των σφαλμάτων τους, το αναμενόμενο σφάλμα για ένα δείγμα μπορεί να ελαττωθεί στο μηδέν καθώς το πλήθος των επιμέρους ταξινομητών τείνει στο άπειρο.
Παρόλα αυτά, αυτές οι υποθέσεις σπάνια ισχύουν σε πραγματικές εφαρμογές. Οι Krogh et al. (1995) απέδειξαν ότι το σφάλμα του συλλογικού ταξινομητή μπορεί να διαιρεθεί σε δύο όρους, έναν που μετράει το μέσο σφάλμα γενίκευσης κάθε επιμέρους ταξινομητή και ένα δεύτερο όρο που μετράει το βαθμό ασυμφωνίας μεταξύ των επιμέρους ταξινομητών. Αυτό που έδειξαν με επίσημο τρόπο είναι ότι ένας ιδανικός συλλογικός ταξινομητής αποτελείται από πολύ αποτελεσματικούς επιμέρους ταξινομητές που διαφέρουν στις ταξινομήσεις τους όσο το δυνατό περισσότερο. Οι Opitz et al. [16] εμπειρικά επιβεβαίωσαν ότι αυτού του είδους οι συλλογικοί ταξινομητές γενικεύουν αποτελεσματικά.

Επομένως, οι μέθοδοι κατασκευής συλλογικών ταξινομητών έχουν ως κεντρική ιδέα τη χρήση ταξινομητών που διαφέρουν στις προβλέψεις τους. Αυτό επιτυγχάνεται με διάφορες τροποποιήσεις κατά τη διαδικασία εκπαίδευσης, ώστε να προκύψουν επιμέρους ταξινομητές με διαφορετικά αποτελέσματα.

2.6.1. Μοντέλο Συλλογικών Ταξινομητών

Μια τυπική μέθοδος κατασκευής συλλογικού ταξινομητή περιέχει τα εξής δομικά στοιχεία:
Σύνολο εκπαίδευσης (Training Set). Ένα σύνολο δεδομένων με κατηγορίες για την εκπαίδευση του συλλογικού ταξινομητή. Τα δεδομένα εκπαίδευσης μπορούν να εκφραστούν με διάφορους τρόπους αλλά συνήθως εκφράζονται ως διανύσματα με συγκεκριμένες τιμές στα διάφορα χαρακτηριστικά. Χρησιμοποιείται ο συμβολισμός Α για

την αναπαράσταση του συνόλου των μεταβλητών εισόδου με n χαρακτηριστικά: $A=\{a_1,\ldots, a_i,\ldots, a_n\}$ και y για την αναπαράσταση της κατηγορίας.

Επαγωγέας βάσης (Base Inducer). Ο επαγωγέας είναι ένας αλγόριθμος, ο οποίος λαμβάνει ως είσοδο το σύνολο εκπαίδευσης και κατασκευάζει ένα ταξινομητή. Έστω ότι ο συμβολισμός I αναπαριστά έναν επαγωγέα. Χρησιμοποιείται η έκφραση M= I(S) για την αναπαράσταση ενός ταξινομητή M ο οποίος προέκυψε από τον επαγωγέα I στο σύνολο εκπαίδευσης S.

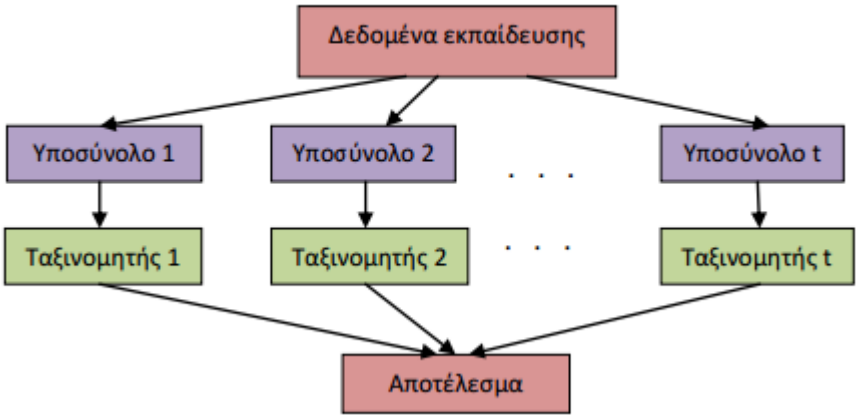
Γεννήτρια ποικιλομορφίας (Diversity Generator). Το στοιχείο αυτό είναι υπεύθυνο για τη διαφοροποίηση των επιμέρους ταξινομητών ώστε να είναι αποδοτικό το συλλογικό σύστημα.
Συνδυαστής (Combiner). Το στοιχείο αυτό συνδυάζει τις ταξινομήσεις των επιμέρους ταξινομητών για την παραγωγή της συλλογικής απόφασης.

Είναι χρήσιμο να γίνει διάκριση μεταξύ των εξαρτημένων μοντέλων συλλογικών ταξινομητών και των ανεξάρτητων μοντέλων. Στα εξαρτημένα μοντέλα η έξοδος του κάθε επιμέρους ταξινομητή χρησιμοποιείται ως είσοδος για τον επόμενο επιμέρους ταξινομητή, επομένως υπάρχει η δυνατότητα εκμετάλλευσης της γνώσης που παράγεται σε προηγούμενες επαναλήψεις για την καθοδήγηση της μάθησης στις επόμενες επαναλήψεις. Στα ανεξάρτητα μοντέλα οι επιμέρους ταξινομητές εκπαιδεύονται αυτόνομα και ανεξάρτητα και στη συνέχεια οι έξοδοι τους συνδυάζονται.

- Bagging

Μια απλή και αρκετά γνωστή μέθοδος είναι η μέθοδος bagging ή bootstrap aggregation. Ανήκει στην κατηγορία των τεχνικών όπου βασίζονται σε ένα αλγόριθμο εξόρυξης γνώσης και χρησιμοποιούν επαναληπτική δειγματοληψία (δειγματοληψία με επανατοποθέτηση) στα δεδομένα εκπαίδευσης για να παράγουν μια ομάδα ταξινομητών. Η κατηγορία αυτή εκμεταλλεύεται την αστάθεια που παρουσιάζουν ορισμένοι αλγόριθμοι στις μικρές αλλαγές στα δεδομένα εκπαίδευσης.

Ας δούμε πως δουλεύει η μέθοδος bagging. Χωρίζουμε το σύνολο δεδομένων σε t ίσα σύνολα με τυχαία επιλογή και επανατοποθέτηση. Τα σύνολα αυτά μπορεί να είναι ίδια, παρόμοια ή και τελείως διαφορετικά. Σε κάθε υποσύνολο που δημιουργήθηκε εφαρμόζουμε τον αλγόριθμο επαγωγής και έτσι έχουμε ένα σύνολο ταξινομητών. Συνδυάζοντας τα αποτελέσματα θα έχουμε την τελική εκτίμηση. Ο τρόπος συνδυασμού εξαρτάται από τις εξόδους των ταξινομητών. Αν οι έξοδοι είναι συνεχείς (ανήκουν σε κάποιο διάστημα) τότε παίρνουμε ως αποτέλεσμα το μέσο όρο των εξόδων. Σε περίπτωση που έχουμε επιγραφές κατηγοριών το τελικό αποτέλεσμα συμπίπτει με την απόφαση της πλειοψηφίας (η επιγραφή που προτείνουν οι περισσότεροι ταξινομητές).



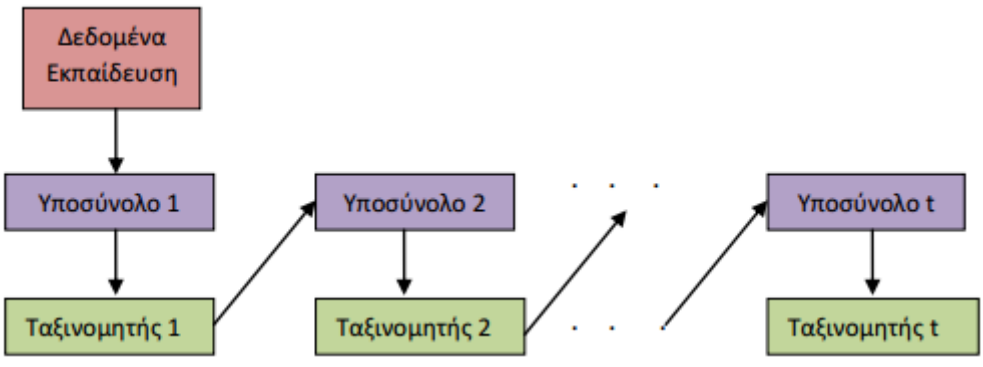
Σχήμα 2.9 Σχηματική αναπαράσταση της μεθόδου bagging

- Boosting

Μία άλλη μέθοδος που εμφανίστηκε τα τελευταία χρόνια και έχει επικεντρώσει το ενδιαφέρον πολλών ερευνητών είναι η μέθοδος boosting. Ανήκει στην ίδια κατηγορία με την μέθοδο bagging αλλά είναι λίγο πιο πολύπλοκη. Μπορεί να μειώσει κατά πολύ το σφάλμα που προκύπτει από αδύναμες μηχανές ταξινόμησης (δηλαδή αυτές που έχουν επίδοση λίγο καλύτερη από αυτήν της τυχαίας επιλογής) καθώς χρησιμοποιεί σειριακή εκπαίδευση σε μια ομάδα ταξινομητών. Η επίδοση δηλαδή μιας μηχανής επηρεάζεται από την επίδοση της προηγούμενης μηχανής σε αντίθεση με την μέθοδο bagging όπου η

εκπαίδευση του κάθε υποσυνόλου γινόταν ανεξάρτητα από των άλλων. Ακόμη τα νέα μοντέλα ταξινόμησης που δημιουργούνται ειδικεύονται στα παραδείγματα που στην προηγούμενη εκπαίδευση απέτυχαν να ταξινομηθούν σωστά. Και σε αυτήν την μέθοδο ο τρόπος συνδυασμού εξαρτάται από την έξοδο όπως στην μέθοδο bagging. Ο σημαντικότερος αλγόριθμος της κατηγορίας είναι ο AdaBoost (Adaptive Boosting).

Στην αρχή όλα τα στιγμιότυπα εκπαίδευσης έχουν το ίδιο βάρος $w = \frac{1}{N}$ όπου N το πλήθος των στοιχείων του συνόλου εκπαίδευσης. Έπειτα ξεκινάει η εκπαίδευση του κάθε μοντέλου και υπολογίζεται το σφάλμα e ως το άθροισμα των βαρών των λάθος ταξινομημένων στιγμιότυπων. Σε περίπτωση που το σφάλμα είναι 0 ή μεγαλύτερο ή ίσο του 0.5 (>= 0.5) τότε έχουμε τον τερματισμό των επαναλήψεων. Αλλιώς πολλαπλασιάζουμε τα βάρη των στιγμιότυπων που ταξινομήθηκαν λάθος με τον παράγοντα $\beta = \frac{e}{(1-e)}$ και έπειτα κάνουμε κανονικοποίηση των βαρών ώστε το άθροισμά τους να ισούται με ένα. Για κάθε ένα από τα μοντέλα προσθέτουμε στο βάρος της κάθε κατηγορίας την τιμή της ψήφου που είναι ίση με $\log\frac{(1-e)}{e}$. Το τελικό αποτέλεσμα είναι η κατηγορία με το μεγαλύτερο άθροισμα.



Σχήμα 2.10 Σχηματική αναπαράσταση της μεθόδου AdaBoost

2.7. Επικύρωση του σφάλματος των ταξινομητών

Οι περισσότεροι αλγόριθμοι ταξινόμησης προτύπων έχουν μια ή περισσότερες ελεύθερες παραμέτρους [6][42]. Για παράδειγμα ο αριθμός νευρώνων σε ένα νευρωνικό δίκτυο, ο αριθμός των κοντινών γειτόνων σε έναν kNN ταξινομητή, τα κέντρα και οι διακυμάνσεις στην παραμετρική εκτίμηση πυκνότητας και στις συναρτήσεις RBF, το εύρος h στους εκτιμητές πυκνότητας με Kernels και στα νευρωνικά δίκτυα PNN, ο αριθμός των βέλτιστων χαρακτηριστικών γνωρισμάτων κ.α.
Η ποιότητα των μοντέλων εξετάζεται με την εκτίμηση του σφάλματος γενίκευσης, δηλαδή την ικανότητα του μοντέλου να προβλέπει την κατηγορία μιας νέας περίπτωσης. Με βάση την απόδοση του κάθε ταξινομητή στα διαθέσιμα δεδομένα προκύπτει το εάν ένα μοντέλο καλύπτει καλά το σύνολο των δεδομένων λειτουργίας του. Υπάρχουν λοιπόν δύο σφάλματα.

Το σφάλμα εκπαίδευσης που δείχνει πόσο καλά προσεγγίζει το μοντέλο τα N_{train} το πλήθος παραδειγμάτων εκπαίδευσης με τα οποία εκπαιδεύτηκε και είναι αυτό που ελαχιστοποιείται κατά τη διάρκεια εκπαίδευσης, και το σφάλμα ελέγχου που δείχνει πόσο καλά γενικεύει το μοντέλο σε N_{test} το πλήθος νέα παραδείγματα ελέγχου.
Δύο ζητήματα που προκύπτουν [6][42] είναι η επικύρωση (validation) που δηλώνει «πώς υπολογίζεται το πραγματικό ποσοστό σφάλματος του» και αφετέρου η επιλογή μοντέλου (model selection) [6] που δηλώνει «πώς επιλέγεται η βέλτιστη παράμετρος (s) για ένα δεδομένο πρόβλημα ταξινόμησης». Το πραγματικό σφάλμα είναι το ποσοστό σφάλματος του ταξινομητή όταν δοκιμάζεται σε ολόκληρο τον πληθυσμό.
Κάποιος μπορεί να χρησιμοποιήσει το σύνολο εκπαίδευσης και για να επιλέξει το "βέλτιστο" ταξινομητή και για να εκτιμήσει το σφάλμα από αυτά ($N_{train}=N_{test}$). Αυτή η απλοϊκή προσέγγιση έχει δύο θεμελιώδη προβλήματα.
Πρώτον το τελικό μοντέλο θα έχει υπερπροσαρμογή (overfit) και δεν θα είναι σε θέση να γενικεύσει σε νέα δεδομένα. Το πρόβλημα υπερπροσαρμογής εμφανίζεται περισσότερο έντονα σε μοντέλα με μεγάλο αριθμό παραμέτρων.
Δεύτερο η εκτίμηση σφάλματος ελέγχου θα είναι υπερβολικά αισιόδοξη (χαμηλότερη από το αληθινό) και μεροληπτική. Στην πραγματικότητα, δεν είναι ασυνήθιστο να υπάρξει σωστή ταξινόμηση 100% στα δεδομένα εκπαίδευσης.

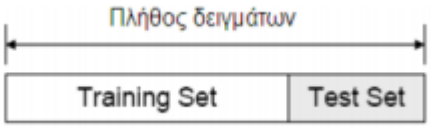
Άρα πρέπει να χωρίσουμε σε ανεξάρτητα σύνολα train και test. Ας εξετάσουμε τι πρέπει να κάνουμε όταν το ποσό των δεδομένων για train και test είναι περιορισμένο. Οι τεχνικές που παρουσιάζονται εδώ επιτρέπουν την καλύτερη χρήση των δεδομένων για

- εκπαίδευση (training)
- εκτίμηση απόδοσης (testing)
- επιλογή μοντέλου (model selection)

Αν και η χρήση των τεχνικών που συζητάμε αυξάνει το συνολικό χρόνο υπολογισμού, είναι ανεκτίμητες για την αποτίμηση της ποιότητας των μοντέλων κατηγοριοποίησης, την ακρίβεια αποτελεσμάτων, την επιλογή κατάλληλων ελεύθερων παραμέτρων και την επιλογή μεταξύ διαφορετικών ταξινομητών.

2.7.1. Η μέθοδος holdout (train set-test set)

Η μέθοδος holdout [42] εκτιμά το σφάλμα διαχωρίζοντας το σύνολο δεδομένων σε δύο ανεξάρτητα σύνολα, ένα σύνολο δεδομένων εκπαίδευσης (Training set), και ένα σύνολο δεδομένων ελέγχου (Test set). Κρατά ένα μέρος για δοκιμές ελέγχου, δηλαδή εκτίμηση του σφάλματος ταξινομητή και χρησιμοποιεί τα υπόλοιπα δεδομένα για εκπαίδευση για να καθορίσουν έτσι το μοντέλο, και τις ελεύθερες παραμέτρους του.



Σχήμα 2.11 Training- Test Set

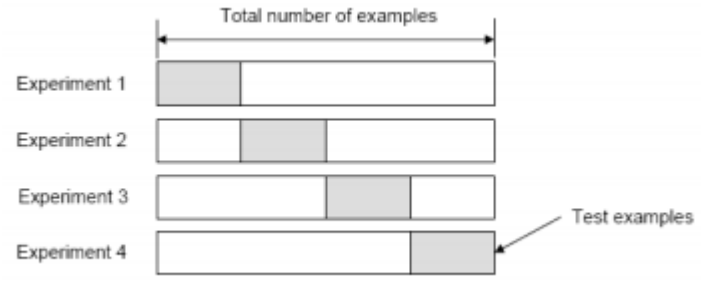
Στην πράξη συνηθίζεται να κρατά το 1/3 για έλεγχο και να χρησιμοποιεί τα 2/3 για εκπαίδευση. Φυσικά, μπορεί να είμαστε άτυχοι και τα δείγματα που έχουμε για εκπαίδευση (ή έλεγχο) να μην είναι αντιπροσωπευτικά. Γενικά, δεν μπορούμε να καθορίσουμε αν το δείγμα είναι αντιπροσωπευτικό ή όχι αλλά υπάρχει ένας έλεγχος: κάθε μια από τις κατηγορίες στο πλήρες σύνολο πρέπει να αναπαρίσταται με περίπου την σωστή αναλογία στα υποσύνολα. Πράγματι, πρέπει να εξασφαλίσουμε ότι η τυχαία

δειγματοληψία κατά το διαχωρισμό σε δύο υποσύνολα γίνεται με τέτοιο τρόπο ώστε να εγγυάται ότι κάθε κατηγορία αναπαρίσταται κατάλληλα και στο σύνολο εκπαίδευσης και στο ελέγχου. Η διαδικασία αυτή που καλείται stratification παρέχει σε κάποιο βαθμό μία ομαλή εκπροσώπηση κατηγοριών στα σύνολα εκπαίδευσης ή ελέγχου. Το stratified Sampling είναι χρήσιμο όταν υπάρχουν μεγάλες διαφορές στη συχνότητα εμφάνισης κάθε κατηγορίας. Η βασική ιδέα του stratification είναι απλή: ο αρχικός πληθυσμός διαμερίζεται σε υποσύνολα που λέγονται strata, χρησιμοποιώντας τις τιμές κάποιου χαρακτηριστικού ή ομάδας χαρακτηριστικών. Ένα τυχαίο δείγμα επιλέγεται έπειτα από κάθε stratum.

2.7.2. Η μέθοδος K-Fold Cross Validation (διασταυρωμένης επικύρωσης)

Η μέθοδος holdout μπορεί να έχει δύο μειονεκτήματα [42]. Πρώτον σε προβλήματα όπου υπάρχει ένα αραιό σύνολο δεδομένων (sparse data set) μπορεί να μην είμαστε σε θέση να αντέξουμε την "πολυτέλεια" του παραμερισμού μιας μερίδας του συνόλου δεδομένων. Το δεύτερο είναι ότι αφού αποτελεί ένα ενιαίο πείραμα εκπαίδευσης και δοκιμής, η εκτίμηση holdout του σφάλματος μπορεί να είναι παραπλανητική εάν συμβεί ένας "ανεπιτυχής μη αντιπροσωπευτικός" διαχωρισμός δεδομένων. Οι περιορισμοί του holdout μπορούν να υπερνικηθούν με μια οικογένεια μεθόδων αναδειγματοληψίας, που χρησιμοποιούν τελικά ολόκληρο το σύνολο δεδομένων [28][42] και για εκπαίδευση και για έλεγχο/επικύρωση [6][41], διεξάγοντας βέβαια πολλαπλάσια πειράματα εις βάρος του υψηλότερου υπολογιστικού κόστους. Αυτές είναι οι Cross Validation (K-Fold Cross Validation και Leave-one-out Cross Validation).

Στην K-Fold Cross-validation [6][41][42] αποφασίζουμε για ένα σταθερό αριθμό από folds (πτυχές), ή συνεχόμενες διαιρέσεις ή ομάδες δεδομένων. Υποθέτουμε ότι χρησιμοποιούμε K (π.χ. K=10). Τότε τα δεδομένα θα διαχωριστούν σε K προσεγγιστικά ίσα folds, και κάθε ένα στην συνέχεια θα χρησιμοποιηθεί επαναληπτικά για testing ενώ τα υπόλοιπα για training. Δηλαδή, χρησιμοποιούμε τα (K-1) folds για εκπαίδευση και το 1 fold για έλεγχο, και επαναλαμβάνουμε την διαδικασία K φορές. Έτσι κάθε παράδειγμα του συνόλου των διαθέσιμων δεδομένων χρησιμοποιείται ακριβώς μία φορά ως μέλος τους συνόλου επικύρωσης και k-1 φορές ως μέλος του συνόλου εκπαίδευσης.

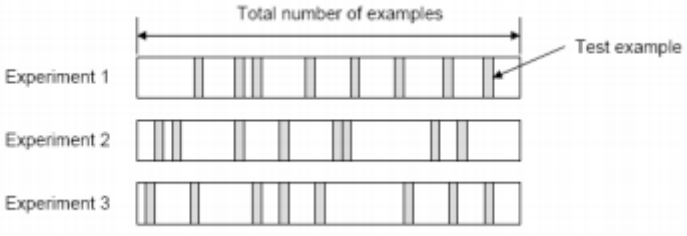


Σχήμα 2.12 K-Fold Cross-validation

Όπως πριν, το πραγματικό σφάλμα Ε υπολογίζεται ως μέσος όρος των χωριστών εκτιμήσεων E_i των πειραμάτων i (i=1,2….K) δηλαδή $E = (1/K)\sum_{i=1}^{K} E_i$. Η τεχνική αυτή καλείται K-fold cross-validation, και εάν συνδυάζεται με stratification (που είναι κοινή πρακτική), τότε αναφέρεται ως stratified k-fold cross-validation. Η K Fold Cross-validation είναι παρόμοια με την μέθοδο Random Subsampling. Το πλεονέκτημα της πρώτης είναι ότι όλα τα παραδείγματα στο σύνολο δεδομένων χρησιμοποιούνται τελικά και για την εκπαίδευση και για τη δοκιμή.

2.8. Η μέθοδος Random Subsampling

Η μέθοδος Random Subsampling (τυχαία υποδειγματοληψία) εκτελεί Κ διασπάσεις ολόκληρου του συνόλου δεδομένων. Για κάθε ομάδα δεδομένων που χωρίζεται επιλέγεται τυχαία ένας (σταθερός) αριθμός παραδειγμάτων χωρίς αντικατάσταση. Για κάθε ομάδα δεδομένων εκπαιδεύεται [41][42] ο ταξινομητής από την αρχή με το σύνολο εκπαίδευσης και έπειτα εκτιμάται το σφάλμα εκπαίδευσης E_i από τα παραδείγματα δοκιμής ελέγχου.



Σχήμα 2.13 Μέθοδος Random Subsampling

Η πραγματική εκτίμηση σφάλματος Ε λαμβάνεται ως μέσος όρος των χωριστών εκτιμήσεων E_i των πειραμάτων i (i=1,2….K) δηλαδή $E = (1/K)\sum_{i=1}^{K} E_i$. Η εκτίμηση με τη μέθοδο αυτή μπορεί να είναι σημαντικά καλύτερη από την εκτίμηση holdout. Για δεδομένα που περιέχουν θόρυβο ή συστάδες με διαφορετικές πυκνότητες, η τεχνική biased sampling προσφέρει τη δυνατότητα ένα σημείο να συμπεριλαμβάνεται στο δείγμα ανάλογα με την τοπική του πυκνότητα.

2.9. Απόδοση κατηγοριοποίησης (Αξιολόγηση ταξινομητών)

Η απόδοση των αλγορίθμων ταξινόμησης εξετάζεται με την εκτίμηση της ακρίβειας (accuracy) της κατηγοριοποίησης, δηλαδή την ικανότητα του μοντέλου να προβλέπει την κατηγορία μιας νέας περίπτωσης. Η εκτίμηση της ακρίβειας είναι ένα πολύ σημαντικό ζήτημα στο χώρο της κατηγοριοποίησης αφού κάτι τέτοιο μας δείχνει το πόσο καλά ανταποκρίνεται ο αλγόριθμος μας για δεδομένα με τα οποία δεν έχει εκπαιδευτεί. Η εκτίμηση της ακρίβειας επίσης επιτρέπει τη σύγκριση των διαφόρων αλγορίθμων κατηγοριοποίησης.

Αν και η ακρίβεια είναι το πιο σημαντικό μέτρο αποτίμησης της απόδοσης του αλγορίθμου κατηγοριοποίησης που χρησιμοποιούμε, υπάρχουν και άλλα μέτρα σύγκρισης:

- Ταχύτητα: Κόστος υπολογισμού (συμπεριλαμβανομένης της παραγωγής και της χρήσης του μοντέλου)
- Rebustness: Σωστή πρόβλεψη με ελλιπή δεδομένα ή δεδομένα με θόρυβο.
- Scalability: Αποδοτική κατασκευή του μοντέλου δοθέντος μεγάλης ποσότητας δεδομένων (μπορεί να εκτιμηθεί μετρώντας τις λειτουργίες I/O που απαιτεί ο αλγόριθμος)
- Interpretability: Επίπεδο κατανόησης και γνώσης που παρέχεται από το μοντέλο. (Μπορεί να εκτιμηθεί μετρώντας το πόσο πολύπλοκο είναι το μοντέλο π.χ αριθμός κόμβων στα δέντρα απόφασης, αριθμός επιπέδων στα νευρωνικά δίκτυα κ.α)

Όσο αφορά την ακρίβεια στην πρόβλεψη της κατηγορίας, το μέτρο αυτό είναι το πιο σημαντικό, ωστόσο δε θα πρέπει να υπολογίζεται ανεξάρτητα από τα υπόλοιπα μέτρα. Για παράδειγμα, δεν έχει νόημα το να έχουμε έναν αλγόριθμο κατηγοριοποίησης που μας δίνει αποτελέσματα με πολύ υψηλή ακρίβεια μετά από πολύ χρόνο. Ίσως να ήταν καλύτερη επιλογή το να έχουμε έναν αλγόριθμο κατηγοριοποίησης που να μας δίνει αποτελέσματα με λίγο χαμηλότερη ακρίβεια από τον καλύτερο (ως προς την ακρίβεια) αλγόριθμο κατηγοριοποίησης αλλά πιο σύντομα. Η ακρίβεια της κατηγοριοποίησης συνήθως υπολογίζεται με τον καθορισμό του ποσοστού των παραδειγμάτων που τοποθετούνται στη σωστή κατηγορία.

2.9.1. Ανισοκατανομή μεταξύ των κατηγοριών (class imbalance)

Έστω ότι θέλουμε να εκπαιδεύσουμε έναν αλγόριθμο κατηγοριοποίησης με ένα σύνολο ιατρικών δεδομένων ώστε αυτός να είναι σε θέση να κατατάσσει τα μελλοντικά παραδείγματα στις κατηγορίες «cancer» ή «not cancer». Μια εκτίμηση της ακρίβειας γύρω στο 90% μπορεί να παρουσιάζει τον αλγόριθμό μας αρκετά ακριβή, ωστόσο τι γίνεται αν μόνο το 3-4 % των δεδομένων εκπαίδευσης ανήκει στην κατηγορία «cancer». Σε μια τέτοια περίπτωση, ένας αλγόριθμος κατηγοριοποίησης με ακρίβεια 90% ίσως να μην είναι δεκτός αφού στην πραγματικότητα θα μπορεί να αναγνωρίζει και να κατηγοριοποιεί μόνο τα παραδείγματα που ανήκουν στην κατηγορία “cancer”. Αντίθετα εμείς θέλουμε να είμαστε σε θέση να καταλαβαίνουμε το πόσο καλά ο αλγόριθμός μας αναγνωρίζει τα “cancer” παραδείγματα (positive samples) και πόσο καλά αναγνωρίζει τα “non cancer” παραδείγματα (negative samples). Για να γίνει αυτό μπορούμε να χρησιμοποιήσουμε τα μέτρα sensitivity, specificity, precision, accuracy, recall και Fmeasure(f1).

Τα παραπάνω μέτρα μπορούν να οριστούν ως εξής:

$$sensitivity = \frac{tp}{tp + fn}$$

$$specificity = \frac{tn}{tn + fp}$$

$$precision = \frac{tp}{tp + fp}$$

$$accuracy = \frac{tp + tn}{tp + fp + fn + tn}$$

$$recall = \frac{tp}{tp + fn}$$

$$Fmeasure(f1) = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

tp είναι ο αριθμός των true positives (“cancer” παραδείγματα που σωστά κατηγοριοποιήθηκαν σαν “cancer”), tn είναι ο αριθμός των true negatives (“not_cancer” παραδείγματα που σωστά κατηγοριοποιήθηκαν σαν “not_cancer” , fp είναι ο αριθμός των false positives (“not_cancer” παραδείγματα που λανθασμένα κατηγοριοποιήθηκαν σαν “cancer”) και fn είναι ο αριθμός των false negative (“cancer” παραδείγματα που λανθασμένα κατηγοριοποιήθηκαν σαν “not cancer”).

Ένας άλλος τρόπος που επιδεικνύει την ακρίβεια της λύσης σε ένα πρόβλημα κατηγοριοποίησης είναι **ο πίνακας σύγχυσης ή μήτρα σύγχυσης (confusion matrix).** Με δεδομένες m κατηγορίες μια μήτρα σύγχυσης είναι μια m x m μήτρα όπου κάθε καταχώρηση $C_{i,j}$ δείχνει τον αριθμό των παραδειγμάτων τα οποία τοποθετήθηκαν στην κατηγορία C_j αλλά των οποίων η πραγματική κατηγορία είναι η $C_{i.}$ Όπως καταλαβαίνουμε, οι καλύτερες λύσεις θα έχουν μόνο μηδενικές τιμές έξω από την κύρια διαγώνιο. Μια μήτρα σύγχυσης για τρεις κατηγορίες παρουσιάζονται στον (Πίνακα 2.2).

Πίνακας 2.2 Μήτρα σύγχυσης

Πραγματική Κατηγορία	Εκχώρηση		
	Short	Medium	Tall
Short	0	5	0
Medium	0	4	3
Tall	0	1	2

ΚΕΦΑΛΑΙΟ 3. ΠΡΟΒΛΕΨΗ ΙΣΟΜΕΡΙΣΜΟΥ ΠΡΟΛΙΝΩΝ

3.1 Πως Ορίζεται το Πρόβλημα
3.2 Τρέχουσα έρευνα
3.3 Η δική μας Συνεισφορά

3.1. Πώς Ορίζεται το Πρόβλημα

Είναι γνωστό ότι οι επίπεδοι πεπτιδικοί δεσμοί συμβαίνουν κυρίως σε trans διαμόρφωση (conformation) [34], cis πεπτιδικοί δεσμοί εμφανίζονται σπάνια σε πρωτεΐνες στις οποίες υπάρχει ένα ενεργειακό φράγμα (energy barrier) περίπου 20 kcal/mol μεταξύ της trans και της cis διαμόρφωσης. Ωστόσο, στην περίπτωση του πεπτιδικού δεσμού Xaa-Pro (όπου Xaa είναι οποιοδήποτε αμινοξύ), η διαφορά ενέργειας είναι μόνο 0,5 kcal/mol μεταξύ trans και cis ισομερισμού, και το ενεργειακό φράγμα είναι περίπου 13 kcal/mol. Έτσι, ένα σημαντικό ποσοστό (περίπου 4-5%) των Xaa-Pro πεπτιδικών δεσμών υιοθετεί τη cis διαμόρφωση, ενώ μόνο το 0,03-0,05% των πεπτιδικών δεσμών Xaa-nonPro εμφανίζονται σε cis διαμόρφωση [12][20][40].

Κατά τα τελευταία χρόνια, υπάρχει ένας αυξανόμενος αριθμός γνωστών πρωτεϊνικών δομών οι οποίες έχουν προσδιοριστεί και εμφανίζουν ετερογένεια στη διαμόρφωση ενός ή περισσοτέρων προλυλο-πεπτιδικών δεσμών [1]. Οι πεπτιδικοί δεσμοί προλίνης cis φέρουν μεγάλη βιολογική σημασία στη δομή και λειτουργία των πρωτεϊνών. Η σημασία του ισομερισμού της προλίνης cis/trans ως περιοριστικό στο ποσοστό της πρωτεϊνικής αναδίπλωσης (protein folding) έχει διερευνηθεί αρκετά [9][25][39], για παράδειγμα, έχει προταθεί ότι κυριαρχεί στην αναδίπλωση της άλφα υπομονάδας (alpha subunit) της trp

συνθάσης (synthase) στην E. coli [43]. Η διαδικασία ισομερισμού των πεπτιδικών δεσμών Xaa-Pro μπορεί να καταλύεται και επιταχύνεται από το λεγόμενο πεπτιδυλο-προλυλο cis/trans ισομερισμό [31], οι οποίοι ανακαλύφθηκε να εμπλέκονται σε σηματοδότηση των κυττάρων και τον κυτταρικό πολλαπλασιασμό, και να εμπλέκονται στην επαγωγή των σοβαρών ασθενειών, όπως ο καρκίνος, το AIDS, η νόσος του Alzheimer και άλλες νευροεκφυλιστικές διαταραχές [8]. Επιπλέον, ο ισομερισμός της προλίνης λειτουργεί ως μοριακός διακόπτης λόγω της δυνητικής του ικανότητας να ελέγχει τη δραστηριότητα της πρωτεΐνης εντός των ορίων της εγγενούς παραλλαγής διαμόρφωσης [2].
Για τους παραπάνω λόγους, η ακριβής πρόβλεψη του ισομερισμού της προλίνης σε cis/trans σε πρωτεΐνες έχει πολλές σημαντικές εφαρμογές στη μελέτη της πρόβλεψης δομής της πρωτεΐνης και τον ορθολογικό μοριακό σχεδιασμό.

3.2. Τρέχουσα έρευνα

Πιο πρόσφατα, οι Pahlke et al. εφάρμοσαν διαφορετικές στατιστικές μεθόδους, όπως τη μέθοδο Chou-Fasman για υπολογισμό των παραμέτρων και μήτρες συχνότητας εμφάνισης (occurrence matrices) για να αναλύσουν την πιθανότητα διαμόρφωσης της προλίνης σε cis και trans και να εξάγουν τα μοτίβα για πιθανή πρόβλεψή της [18]. Πρόσφατη μελέτη σχετικά με τη διατήρηση του cis προλυλο-δεσμού έδειξε ότι τα αμινοξέα με cis πρόλυλο-δεσμό είναι πιο συχνά συντηρημένα σε σχέση με τα αμινοξέα με trans προλυλο-δεσμό σε πρωτεΐνες που σχετίζονται εξελικτικά, και η συνολική ομολογία αλληλουχίας πρωτεΐνης είναι ένας ισχυρότερος δείκτης για την εμφάνιση αμινοξέα με cis προλυλο-δεσμό σε αντίθεση με τα τοπικά μοτίβα ακολουθιών (local sequence motives) [15].

Ωστόσο, οι περισσότερες από αυτές τις μελέτες βασίζονταν στην στατιστική ανάλυση της συχνότητας εμφάνισης γειτονικών καταλοίπων με επίκεντρο την προλίνη, χωρίς περαιτέρω συστηματική πρόβλεψη του ισομερισμού της προλίνης σε cis/trans από την πρωτογενή αλληλουχία της πρωτεΐνης. Από όσο γνωρίζουμε, η πρώτη απόπειρα πρόβλεψης του ισομερισμού πεπτιδικού δεσμού της προλίνης σε cis/trans με βάση τις αλληλουχίες αμινοξέων έγινε από τους Frömmel και Preissner [10]. Χρησιμοποίησαν έξι διαφορετικά μοτίβα για να ορίσουν σωστά περίπου το 72,7% των γνωστών αμινοξέων προλίνης cis

(176 προλίνες cis σε ένα σχετικά μικρό σύνολο δεδομένων με 242 δεσμούς Xaa-Pro), λαμβάνοντας υπόψη τα γειτονικά ± 6 αμινοξέων με κέντρο την προλίνη, καθώς και τις φυσικοχημικές ιδιότητες. Αργότερα, οι μηχανές διανυσμάτων υποστήριξης (support vector machines-SVM) εισήχθησαν στη συνέχεια για να υλοποιήσουν αυτό το σκοπό και πέτυχε 76,7% ακρίβεια πρόβλεψης με τη χρήση της μεθόδου jack-knife για τα cis αμινοξέα προλίνης, χρησιμοποιώντας την πληροφορία της αλληλουχίας αμινοξέων που κωδικοποιείται από δυαδικά ψηφία (0 και 1), ως διάνυσμα εισόδου [37]. Ο αλγόριθμος COPS αναπτύχθηκε για να προβλέψει τον ισομερισμό του πεπτιδικού δεσμού cis/trans με βάση τις παραμέτρους της διαμόρφωσης [19], αλλά αυτή η μέθοδος εκμεταλλευόταν μόνο την πληροφορία της δευτερογενούς δομής τριάδων αμινοξέων, παραλείποντας σημαντικές πληροφορίες που μπορούν να εξαχθούν από την αλληλουχία των αμινοξέων.

3.3. Η δική μας Συνεισφορά

Στην παρούσα εργασία προτείνουμε διάφορες μεθόδους ταξινόμησης, όπως τη μέθοδο των k-κοντινότερων γειτόνων (k nearest neighbors), τα δέντρα απόφασης (decision trees), τις μηχανές διανυσμάτων υποστήριξης, του αφελούς Μπεϋζιανού (Naïve Bayes) ταξινομητή και την κατασκευή συνδυαστικών ταξινομητών (ensembles) με βάση αυτούς για την πρόβλεψη της διαμόρφωση του πεπτιδικού δεσμού της προλίνης σε cis ή trans. Επίσης προτείνουμε αρκετά καινοτόμα χαρακτηριστικά (συμβολικά και δομικά) σε σύγκριση με τα χαρακτηριστικά που χρησιμοποιήθηκαν σε προηγουμένως δημοσιευμένες εργασίες. Συγκεκριμένα, τα χαρακτηριστικά που επιλέξαμε βασίζονται στα δομικά συστατικά των πρωτεϊνών (που παρουσιάσαμε στην ενότητα 1.1) την πρωτοταγή δομή των πρωτεϊνών (που παρουσιάσαμε στην ενότητα 1.3) για την εξαγωγή μοτίβων αμινοξέων και τον υπολογισμό δίεδρων γωνιών μεταξύ των αμινοξέων και την αναδίπλωση των πολυπεπτιδικών αλυσίδων στο χώρο.

ΚΕΦΑΛΑΙΟ 4. ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΞΑΓΩΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

4.1 Ανάγνωση αρχείων PDB (Protein Data Bank) [3].
4.2 Παραδοχές της μεθόδου
4.3 Εξαγωξή χαρακτηριστικών

4.1. Ανάγνωση αρχείων PDB (Protein Data Bank) [3].

Δύο είναι οι βάσεις δεδομένων από τις οποίες μπορεί κανείς να συγκεντρώσει πρωτεϊνικά δεδομένα. Η πρώτη βάση δεδομένων είναι η PDB (Protein Data Bank) [3]. Η συγκεκριμένη βάση δεδομένων περιέχει δομικά δεδομένα βιολογικών μακρομορίων τα οποία έχουν προκύψει στην πλειονότητά τους είτε από πειράματα κρυσταλλογραφίας είτε πειράματα πυρηνικού μαγνητικού συντονισμού (NMR - Nuclear Magnetic Resonance). Η δεύτερη βάση δεδομένων είναι η Uniprot. Η συγκεκριμένη βάση δεδομένων περιέχει αναλυτικά χαρακτηρισμένες πρωτεϊνικές ακολουθίες, με πολλές συνδέσεις προς τρίτες βάσεις δεδομένων. Τα δεδομένα με τα οποία ασχοληθήκαμε προέρχονται από την πρώτη βάση δεδομένων για αυτό και παρακάτω παρουσιάζουμε αναλυτικά τη δομή τους. Τα αρχεία αυτά ονομάζονται pdb αρχεία και είναι αρχεία κειμένου με συγκεκριμένη δομή και καταληξη .pdb.

Κάθε αρχείο pdb αποτελείται από μία σειρά εγγραφών οι οποίες αποτελούνται από 80 στήλες. Κάθε εγγραφή διαθέτει ένα κωδικό έξι χαρακτήρων, στην αρχή της, επεξηγηματικό της πληροφορίας που η εγγραφή περιέχει. Κάθε εγγραφή δύναται να

καταλαμβάνει μία ή περισσότερες γραμμές, ενώ ορισμένες εγγραφές είναι προαιρετικές. Οι εγγραφές που συναντώνται σε ένα αρχείο pdb (Σχήμα 4.1) είναι οι εξής:

```
HEADER    OXYGEN TRANSPORT                        18-SEP-92   1CMY      1CMY   2
COMPND    HEMOGLOBIN YPSILANTI (CARBON MONOXY FORM)                       1CMY   3
SOURCE    HUMAN (HOMO SAPIENS) BLOOD                                      1CMY   4
AUTHOR    F.R.SMITH,E.E.LATTMAN, C.W.CARTER JUNIOR                        1CMY   5
REVDAT   1   31-OCT-93 1CMY    0                                          1CMY   6
JRNL        AUTH   F.R.SMITH,E.E.LATTMAN,C.W.CARTER JUNIOR               1CMY   7
JRNL        TITL   THE MUTATION BETA99 ASP-TYR STABILIZES Y-A NEW,        1CMY   8
...
REMARK   1                                                              1CMY  13
REMARK   2                                                              1CMY  14
REMARK   2 RESOLUTION. 3.0  ANGSTROMS.                                  1CMY  15
REMARK   3                                                              1CMY  16
REMARK   3 REFINEMENT.                                                  1CMY  17
REMARK   3   PROGRAM                    PROLSQ                          1CMY  18
REMARK   3   AUTHORS                    KONNERT,HENDRICKSON             1CMY  19
...
SEQRES   1 D  146  VAL HIS LEU THR PRO GLU GLU LYS SER ALA VAL THR ALA  1CMY  90
SEQRES   2 D  146  LEU TRP GLY LYS VAL ASN VAL ASP GLU VAL GLY GLY GLU  1CMY  91
SEQRES   3 D  146  ALA LEU GLY ARG LEU LEU VAL VAL TYR PRO TRP THR GLN  1CMY  92
...
HET    HEM  A 142      43     PROTOPORPHYRIN IX WITH FE(II)              1CMY 102
HET    HEM  B 147      43     PROTOPORPHYRIN IX WITH FE(II)              1CMY 103
HET    HEM  C 142      43     PROTOPORPHYRIN IX WITH FE(II)              1CMY 104
HET    HEM  D 147      43     PROTOPORPHYRIN IX WITH FE(II)              1CMY 105
FORMUL   5  HEM    4(C34 H32 N4 O4 FE1 ++)                              1CMY 106
HELIX   23 DA  THR D    4  VAL D   18  1                                 1CMY 129
HELIX   24 DB  ASN D   19  VAL D   34  1                                 1CMY 130
HELIX   25 DC  TYR D   35  PHE D   41  1                                 1CMY 131
...
CRYST1   93.100   93.100  144.600  90.00  90.00 120.00 P 32 2 1     12  1CMY 137
...
HETATM 1071 FE   HEM A 142      -1.337  64.961 152.106  1.00 15.00      1CMY1217
HETATM 1072  CHA HEM A 142      -0.732  65.412 155.438  1.00 15.00      1CMY1218
HETATM 1073  CHB HEM A 142      -0.190  68.087 151.470  1.00 15.00      1CMY1219
...
ATOM   3398  N   VAL D   1      -5.807  33.934 128.823  1.00 15.00      1CMY3544
ATOM   3399  CA  VAL D   1      -5.080  32.676 128.654  1.00 15.00      1CMY3545
ATOM   3400  C   VAL D   1      -5.517  31.975 127.349  1.00 15.00      1CMY3546
ATOM   3401  O   VAL D   1      -6.700  31.665 127.136  1.00 15.00      1CMY3547
ATOM   3403  CG1 VAL D   1      -5.732  32.404 131.123  1.00 15.00      1CMY3549
```

Σχήμα 4.1 Δομή ενός αρχείου PDB.

HEADER: Τετραψήφιος κωδικός για την αναγνώριση της εγγραφής στην (Protein Data Bank) PDB, περιέχει μια γενική ταξινόμηση του μακρομορίου καθώς και την ημερομηνία κατάθεσης της δομής στην PDB.
TITLE: Περιλαμβάνει συνήθως τα περιεχόμενα της εγγραφής, την πειραματική διαδικασία που χρησιμοποιήθηκε, την ύπαρξη μεταλλάξεων. Είναι το πεδίο όπου ο ερευνητής που καταθέτει τη δομή, αποδεικνύει τη σημαντικότητα της εργασίας του.
COMPOUND: Στο σημείο αυτό περιέχονται πληροφορίες για το μακρομόριο οι οποίες αναφέρονται στη δομή του καθώς και σε άλλα μικρότερα μόρια με τα οποία, πιθανώς, έχει αλληλεπιδράσει.
SOURCE: Βιολογική προέλευση του μακρομορίου.
KEYWDS: Χαρακτηριστικές λέξεις κλειδιά για το χαρακτηρισμό της δομής και την αναζήτηση της στην PDB.
EXPDATA: Πειραματική τεχνική προσδιορισμού της δομής.
AUTHOR: Λίστα με τα ονόματα των ερευνητών που συνετέλεσαν στην καταγραφή του μακρομορίου.
JRNL: Βιβλιογραφική αναφορά για τον προσδιορισμό της δομής που αναφέρει η παρούσα εγγραφή.
REMARK: Στη θέση αυτή, περιλαμβάνονται μια σειρά από πληροφορίες σχετικές με τη δομή που περιέχεται στο αρχείο. Περιέχονται βασικές βιβλιογραφικές αναφορές που σχετίζονται άμεσα με το προς εξέταση μακρομόριο. Στο πεδίο αυτό

επιπλέον περιλαμβάνονται και στοιχεία σχετικά με την πειραματική διαδικασία που ακολουθήθηκε για την ανάλυση της δομής, όπως τα διάφορα προγράμματα που οι ερευνητές χρησιμοποίησαν, οι τιμές διάφορων δεικτών και γενικότερα πληροφορίες που δείχνουν την ορθότητα της εγγραφής.

SEQRES: Περιέχει την αμινοξεϊκή αλληλουχία του προκείμενου μακρομορίου. Στην περίπτωση των πρωτεϊνών ακολουθείται ο κώδικας τριών γραμμάτων για τα αμινοξέα.

HET: Αναφορά στα μόρια (ετεροάτομα) που δεν είναι αμινοξέα ή νουκλεοτίδια. Αυτά μπορεί να είναι προσθετικές ομάδες και ιόντα για τα οποία έχουν προσδιοριστεί οι συντεταγμένες τους. Τα στοιχεία που παρουσιάζονται εδώ είναι ένας κώδικα αναγνώρισης για τη διάκριση με τα υπόλοιπα κατάλοιπα της εγγραφής, η αρίθμηση που καταλαμβάνουν μέσα στο αρχείο των συντεταγμένων και τέλος ο αριθμός των ατόμων από τα οποία αποτελούνται.

HETNAM: Ονοματολογία των καταλοίπων που περιέχονται στο πεδίο HET.

FORMUL: Μοριακός τύπος των καταλοίπων που αναφέρονται στο πεδίο HET.

HELIX: Τμήματα της αλληλουχίας που έχουν ελικοειδή δομή.

SHEET: Τμήματα της αλληλουχίας που έχουν δομή πτυχωτού φύλλου.

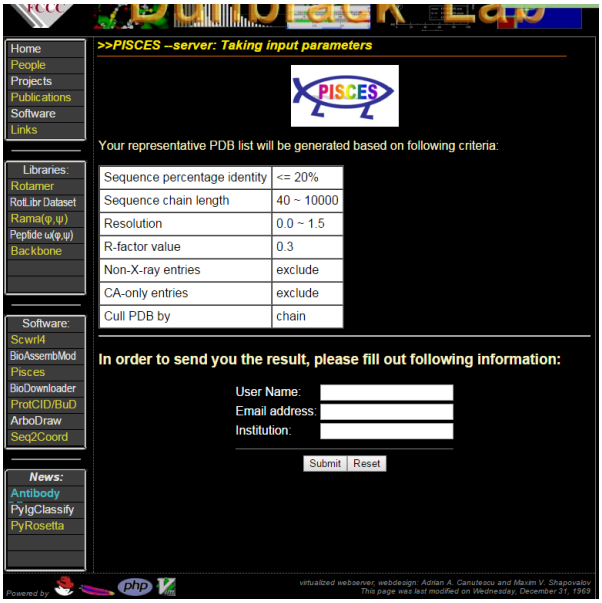
TURN: Τμήματα της αλληλουχίας που έχουν στροφές.

CRYST1: Στο πεδίο αυτό περιέχονται οι παράμετροι μοναδιαίας κυψελίδας και της ομάδας συμμετρίας.

ATOM: Εδώ περιέχονται οι συντεταγμένες των ατόμων στους άξονες X, Y, Z. Το πεδίο περιλαμβάνει επίσης και άλλα στοιχεία όπως τα άτομα για τα οποία αναφέρονται οι συντεταγμένες και σε ποια κατάλοιπα ανήκουν (20 βασικά αμινοξέα).

HETATM: Εδώ περιέχονται οι συντεταγμένες των ατόμων στους άξονες X, Y, Z. Το πεδίο περιλαμβάνει επίσης και άλλα στοιχεία όπως τα άτομα για τα οποία αναφέρονται οι συντεταγμένες και σε ποια κατάλοιπα ανήκουν (όλα τα υπόλοιπα αμινοξέα και τα μόρια νερού).

Για τη μελέτη αυτή, χρησιμοποιήθηκαν 1900 αρχεία PDB που προέκυψαν σύμφωνα με την επιλογή (cull) πρωτεϊνών που προέκυψαν από το server PISCES, συγκεκριμένα η παραμετροποίηση που επέστρεψε αυτά τα αρχεία δίνεται στο Σχήμα 4.2.



Σχήμα 4.2 Παραμετροποίηση του Pisces server [33] για την εξαγωγή του συνόλου δεδομένων (πρωτεϊνών) για τη μελέτη.

Αρχικά ο χρήστης, ξεκινάει να τρέχει τον κώδικα *matlab* από το script main.m. Έπειτα ο κώδικας επιτρέπει στο χρήστη να επιλέξει το είδος της ανάλυσης, ενός ή περισσοτέρων pdb αρχείων. Αφού ο χρήστης επιλέξει ένα από τα δύο είδη, διαλέγει είτε ένα pdb αρχείο κατά την πρώτη περίπτωση είτε ένα φάκελο, ο οποίος πρέπει να περιέχει τουλάχιστον ένα pdb αρχείο, κατά τη δεύτερη περίπτωση. (Τα pdb αρχεία μας δίνουν πληροφορία για κάθε πρωτεΐνη, πως απλώνεται στο χώρο. Εξάγουμε απ' αυτά τις x,y,z συντεταγμένες, όπως παρουσιάζονται παραπάνω. Έτσι δημιουργούμε ένα μοτίβο προλίνης, το οποίο αποτελείται από 11 συνολικά αμινοξέα, 5 πριν την προλίνη και 5 μετά. Στη συνέχεια βρίσκουμε τη δίεδρη γωνία που σχηματίζεται μεταξύ της προλίνης και του προηγούμενου αμινοξέος, συγκεκριμένα μεταξύ των συντεταγμένων της προλίνης και των συντεταγμένων του προηγούμενου αμινοξέος και ακόμα πιο συγκεκριμένα μεταξύ των ατόμων των συντεταγμένων τους). Έπειτα ο χρήστης επιλέγει την ακτίνα R, της σφαίρας με κέντρο το κεντροειδές (μέση τιμή x,y,z συντεταγμένων) κάθε αμινοξέος προλίνης (PRO) μέσα στην οποία βρίσκει ο κώδικας των αριθμό των αμινοξέων (res) και των νερών (hoh) της πρωτεϊνικής αλυσίδας. Τελικά, αν ο χρήστης έχει επιλέξει προηγουμένως ανάλυση πολλαπλών αρχείων pdb, του δίνεται η δυνατότητα να δημιουργήσει και ένα αρχείο, βάση δεδομένων (prolineDB.csv), που περιέχει τις προλίνες από όλα τα αρχεία pdb που βρεθήκαν στο φάκελο με τα pdb αρχεία. Κατά την ανάλυση πολλαπλών αρχείων pdb το

λογισμικό δημιουργεί ένα φάκελο μέσα στον οποίο αποθηκεύει ένα αρχείο csv (proteinID.csv) για κάθε pdb αρχείο στον φάκελο εισόδου και το αρχείο prolineDB.csv, αν έχει προηγουμένως επιλεγεί να δημιουργηθεί. Επίσης δημιουργείται ένα αρχείο motivesDB.csv στο οποίο αποθηκεύονται τα μοτίβα και ο χαρακτηρισμός της γωνίας της προλίνης (cis, trans).

Για να υπολογίσουμε πόσα αμινοξέα και πόσα νερά βρίσκονται στη σφαίρα με κέντρο κάθε προλίνη και ακτίνα R ο αλγόριθμος κάνει τα παρακάτω βήματα.

- Εντοπίζει τις προλίνες μιας αλυσίδας και υπολογίζει τα κεντροειδή τους (με βάσει τις συντεταγμένες των γνωστών ατόμων τους).
- Υπολογίζει τον πρώτο κοντινότερο γείτονα (1 nearest neighbor) με βάση την ευκλείδεια απόσταση. Δηλαδή υπολογίζει την ευκλείδεια απόσταση του κεντροειδούς κάθε προλίνης με κάθε άτομο ενός αμινοξέος και παίρνει την μικρότερη. Αν αυτή η απόσταση είναι μικρότερη ή ίση από το R, τότε ο αλγόριθμος αποθηκεύει την απόσταση και το αμινοξύ αυτό που ανήκει στη σφαίρα της προλίνης. Τελικά ταξινομούμε με αύξουσα σειρά τα αμινοξέα αυτά σύμφωνα με την απόσταση που απέχουν από το κεντροειδές της προλίνης.

$$\min_{a \in Atoms} \left\{ \left(X_a - X_{PRO}\right)^2 + \left(Y_a - Y_{PRO}\right)^2 + \left(Z_a - Z_{PRO}\right)^2 \right\} \le R^2 \qquad \text{Εξίσωση 4.1}$$

- Υπολογίζει την ευκλείδεια απόσταση του κεντροειδούς κάθε προλίνης με ένα άτομο νερού. Αν αυτή η απόσταση είναι μικρότερη ή ίση από το R, τότε ο αλγόριθμος αυξάνει τον αριθμό των νερών που ανήκουν στη σφαίρα της προλίνης.

$$\left(X_{hoh} - X_{PRO}\right)^2 + \left(Y_{hoh} - Y_{PRO}\right)^2 + \left(Z_{hoh} - Z_{PRO}\right)^2 \le R^2 \qquad \text{Εξίσωση 4.2}$$

4.2. Παραδοχές της μεθόδου

- Ο αλγόριθμος λαμβάνει υπόψη μόνο το πρώτο μοντέλο σε ένα pdb με περισσότερα από ένα μοντέλα.
- Ο αλγόριθμος αγνοεί πανομοιότυπες αλυσίδες που ανήκουν στο ίδιο pdb αρχείο.
- Ο αλγόριθμος ενσωματώνει τα αμινοξέα που δεν έχουν εντοπιστεί σε ένα πείραμα (missing residues-remark 465) στην αλυσίδα αμινοξέων που προκύπτει από τα αμινοξέα της ετικέτας ATOMS σύμφωνα με τον σειριακό αριθμό του αμινοξέος. Αυτό πραγματοποιείται με σκοπό αυτή η ακολουθία των αμινοξέων να ταυτίζεται με την ακολουθία αμινοξέων από την ετικέτα SEQRES, ούτως ώστε να υπολογίζουμε σωστά τη δίεδρη γωνία ωμέγα

μεταξύ της προλίνης και του προηγούμενου αμινοξέος. Έτσι λοιπόν μπορούμε να εντοπίζουμε τις προλίνες τις οποίες δεν μπορούμε να υπολογίσουμε την γωνία ωμέγα είτε επειδή αγνοούνται οι συντεταγμένες των ατόμων της ίδιας της προλίνης είτε επειδή αγνοούνται οι συντεταγμένες των ατόμων του προηγούμενου αμινοξέος.

- Ο αλγόριθμος δεν τυπώνει στα αρχεία εξόδου τις προλίνες για τις οποίες δεν μπορεί να υπολογίσει τη γωνία ωμέγα.

4.3. Εξαγωγή χαρακτηριστικών

Δομή του πίνακα με το τελικό αποτέλεσμα:

	Proline Motif							Absolute Freq. of each Residue in P-sphere																				DSSP		Omega Angle		
	R1	…	R5	P	R6	…	R10	G	A	V	S	L	T	I	C	F	Y	W	N	M	Q	P	K	D	R	E	H	Number of HoHs in P sphere	Acc	Structure	Switch (Yes/No)	Degrees	Type (trans/cis)
P_1	Code V1	…	Code V5	Code P	Code V6		Code V10																							Code SV	1or 0		1 or 2
P_2																																	
	…																																
P_N																																	

Σχήμα 4.3 Πίνακας με την απόλυτη συχνότητα κάθε αμινοξέος στη σφαίρα κάθε προλίνης.

Κάθε γραμμή αυτού του πίνακα είναι εγγραφή για μία προλίνη.

Στις πρώτες 10x20 στήλες αποθηκεύουμε τα αμινοξέα του μοτίβου της προλίνης (μοτίβο μήκους 10 αφού αφαιρούμε την προλίνη-κόκκινη στήλη) με την προτεινόμενη μοντελοποίηση. Μετατρέπουμε τα γράμματα/συντομογραφίες των αμινοξέων σε διανύσματα 20 θέσεων, στα οποία θέτουμε σε 1, μόνο τη θέση που αντιστοιχεί στο εκάστοτε αμινοξύ. Σύμφωνα με τη συγκεκριμένη μοντελοποίηση προκύπτουν τα παρακάτω 20 διαφορετικά διανύσματα (ανά γραμμή του πίνακα 4.1):

Πίνακας 4.1 Πίνακας διανυσμάτων 20 θέσεων στα οποία θέτουμε 1 μόνο τη θέση που αντιστοιχεί στο εκάστοτε αμινοξύ

Διάνυσμα	$1^η$	$2^η$	$3^η$	$4^η$	$5^η$	$6^η$	$7^η$	$8^η$	$9^η$	$10^η$	$11^η$	$12^η$	$13^η$	$14^η$	$15^η$	$16^η$	$17^η$	$18^η$	$19^η$	$20^η$
V_G	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V_A	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V_V	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V_S	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V_L	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V_T	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V_I	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
V_C	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
V_F	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
V_Y	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
V_W	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
V_N	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
V_M	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
V_Q	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
V_P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
V_K	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
V_D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
V_R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
V_E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
V_H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Για παράδειγμα αν το μοτίβο είναι το εξής:

HMIQA**P**DGTDI

Τότε στον πίνακα θα αποθηκεύονται οι κωδικοί/διανύσματα (μήκους 20) των αντίστοιχων γραμμάτων εκτός της προλίνης η οποία αφαιρείται γιατί ειναι σταθερή στο κέντρο κάθε μοτίβου:

H M I Q A **P** D G T D I

V_H V_M V_I V_Q V_A $\mathbf{V_P}$ V_D V_G V_T V_D V_I

Επομένως το παραπάνω μοτίβο μήκους 11, μετατρέπεται σε ένα διάνυσμα μήκους 10x20, δηλαδή μήκους 200 εφόσον αφαιρούμε το διάνυσμα που αντιστοιχεί στην προλίνη.

Στις υπόλοιπες 20 (201-220) στήλες του πίνακα αποθηκεύεται η απόλυτη συχνότητα εμφάνισης κάθε αμινοξέος στη σφαίρα κάθε προλίνης. Για παράδειγμα στη 12η θέση του πίνακα αναγράφεται πόσες φορές εμφανίστηκε το αμινοξύ με συντομογραφία G στη σφαίρα τις i-οστής προλίνης.

Στη στήλη 221 αποθηκεύεται ο αριθμός των HOH (μόρια νερού) που ανήκουν στη σφαίρα της προλίνης.

Στη στήλη 222 αποθηκεύεται ο αριθμός Acc (ο οποίος δείχνει πόση από η επιφάνεια ενός αμινοξέος εκτίθεται σε νερά με μονάδες $Angstrom^2$) που προκύπτει από το λογισμικό DSSP [14] για κάθε προλίνη.

Στις υπόλοιπες 8 (223-230) στήλες του πίνακα αποθηκεύεται το λογικό διάνυσμα κωδικοποίησης της δευτεροταγούς δομής στην οποία ανήκει η προλίνη, structure vector (SV), όπως αυτή προέκυψε από το λογισμικό DSSP [14]. Επειδή τα διαφορετικά είδη δευτεροταγούς δομής πρωτεΐνης είναι 8, προκύπτουν τα παρακάτω 8 διαφορετικά διανύσματα κωδικοποίησης (ανά γραμμή του παρακάτω πίνακα) τα οποία έχουμε θέσει σε 1, μόνο τη θέση που αντιστοιχεί στην κάθε δομή:

Πίνακας 4.2 Διανύσματα κωδικοποίησης

Διάνυσμα	$1^η$	$2^η$	$3^η$	$4^η$	$5^η$	$6^η$	$7^η$	$8^η$
SV_G	1	0	0	0	0	0	0	0
SV_H	0	1	0	0	0	0	0	0
SV_I	0	0	1	0	0	0	0	0
SV_T	0	0	0	1	0	0	0	0
SV_E	0	0	0	0	1	0	0	0
SV_B	0	0	0	0	0	1	0	0
SV_S	0	0	0	0	0	0	1	0
SV_C	0	0	0	0	0	0	0	1

Οι υπόλοιπες στήλες από 231-233 περιέχουν τις πληροφορίες που αφορούν τη γωνία ωμέγα κάθε προλίνης οι οποίες αποτελούν features της κατηγοριοποίησης αλλά ορίζουν τις κατηγορίες που θα προβλέπουν οι ταξινομητές μας (cis ή trans).

ΚΕΦΑΛΑΙΟ 5. ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

5.1 Μέθοδος Υποδειγματοληψίας (Under sampling) της Κλάσης Trans
5.1.1 Μέθοδος των k Κοντινότερων Γειτόνων (k nearest neighbors-kNN)
5.1.2 Δέντρα Απόφασης (Decision Trees-DT)
5.1.3 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines-SVM)
5.1.4 Ταξινομητής Bayes (Naïve Bayes-NB)
5.2 Μέθοδος Ensemble Ταξινομητών
5.2.1 KNN Ensemble
5.2.2 Decision Tree Ensemble
5.2.3 Support Vector Machine Ensemble
5.2.4 Naïve Bayes Ensemble

5.1. Μέθοδος Υποδειγματοληψίας (Under sampling) της Κλάσης Trans

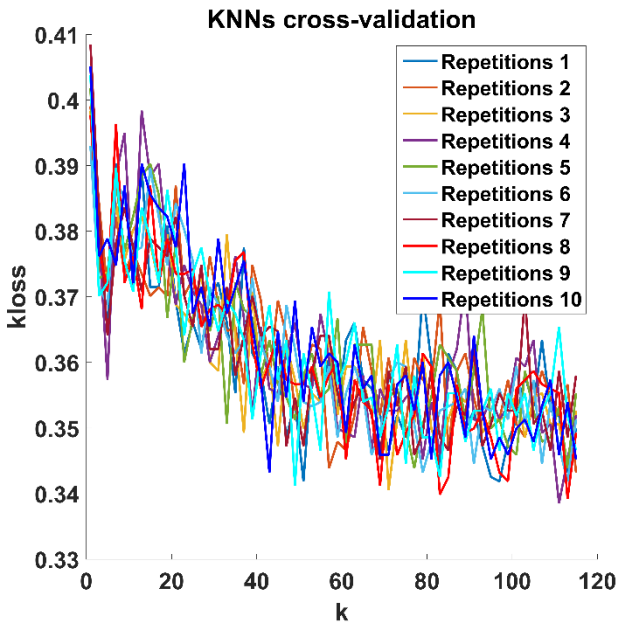
Τα αποτελέσματα που θα παρουσιάσουμε σε αυτή την ενότητα προέκυψαν από κατηγοριοποίηση των ταξινομητών που εκπαιδεύτηκαν με δείγματα μέσω under sampling στην κατηγορία που βρίσκεται σε πλεόνασμα. [3]. Η κατηγορία trans αποτελεί το 94.7% (14875) ενώ η κατηγορία cis αποτελεί το 5.3% (826) του δείγματος μας. Αρχικά, διαχωρίσουμε το δείγμα μας σε 90% για training (743 δείγματα cis και 13388 δείγματα trans) και σε 10% για testing για κάθε κατηγορία. Έπειτα για να αντιμετωπίσουμε την ανισορροπία του δείγματος, παίρνουμε τυχαία δείγματα από την κατηγορία trans, ίσα με τον αριθμό των δειγμάτων του training set της κατηγορίας cis, δηλαδή 743. Έτσι προκύπτει ένα ισορροπημένο training set. Έπειτα για να αξιολογήσουμε τους ταξινομητές επαναλαμβάνουμε τη διαδικασία 10 φορές δημιουργώντας τυχαία 10 διαφορετικά training

και test sets, κρατώντας σταθερό το training set της κατηγορίας cis και κάνοντας under sampling από τα δείγματα της κατηγορίας trans.

Αρχικά για να βρούμε το βέλτιστο ταξινομητή από κάθε είδος εφαρμόζουμε τη μέθοδο του 10-fold cross validation. Για να αξιολογήσουμε τους ταξινομητές κατά το cross validation χρησιμοποιήσαμε τη μετρική kloss του Matlab.Το kloss είναι η μέση απώλεια (loss) κάθε μοντέλου cross-validation (που είναι 10, εφόσον K=10) όταν προβλέπει τα δεδομένα που δεν έχουν χρησιμοποιηθεί για το training.

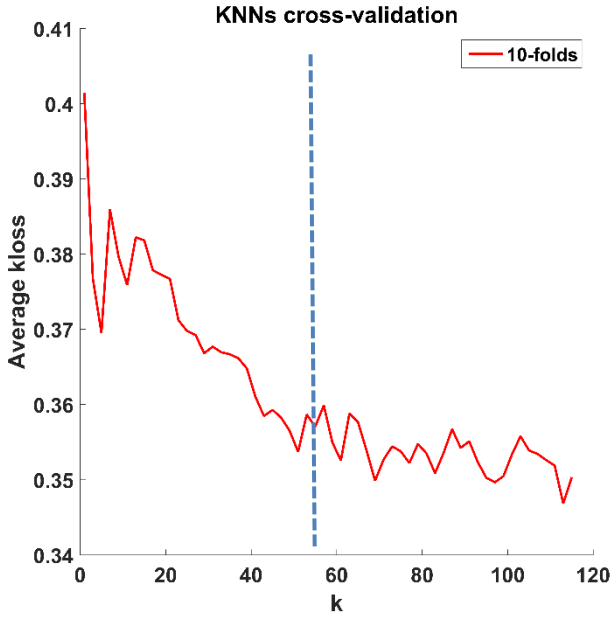
5.1.1. Μέθοδος των k Κοντινότερων Γειτόνων (k nearest neighbors-kNN)

Εδώ προσπαθούμε να βρούμε το βέλτιστο ταξινομητή για τη μέθοδο του κοντινότερου γείτονα δοκιμάζοντας τιμές του k από το 1 μέχρι το 115 (που αποτελεί το μισό του αριθμού των χαρακτηριστικών του δείγματος). Στο Σχ. 5.1, παρουσιάζεται η μετρική kloss για κάθε επανάληψη του 10-fold cross validation.



Σχήμα 5.1 Μετρική kloss του 10-fold cross validation για τις διαφορετικές τιμές του k. Τα διαφορετικά χρώματα απεικονίζουν τις διαφορετικές επαναλήψεις της διαδικασίας.

Για να μπορούμε να ερμηνεύσουμε καλύτερα τα αποτελέσματα του cross validation, όπως αυτά προκύπτουν στο Σχ. 5.1, παίρνουμε τη μέση τιμή του kloss που προέκυψε για κάθε k. Έτσι δημιουργήσαμε το Σχ. 5.2 στο οποίο παρουσιάζεται η μέση τιμή του kloss για κάθε k. Και για τους υπόλοιπους ταξινομητές παρουσιάζουμε μόνο το μέσο kloss που προκύπτει από όλες τις επαναλήψεις του cross validation.

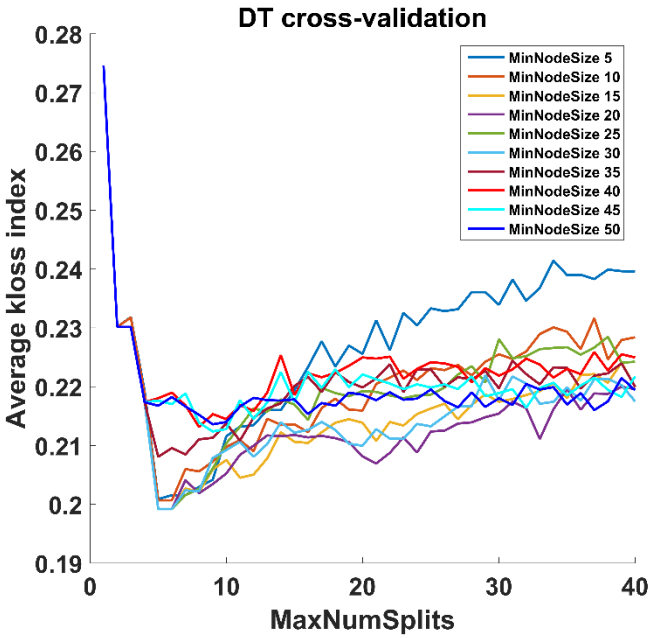


Σχήμα 5.2 Μέσο kloss του 10 fold cross validation για τις διαφορετικές τιμές του k.

Όπως μπορούμε να παρατηρήσουμε από το Σχ. 5.2 ο δείκτης kloss σταθεροποιείται περίπου για k = 61. Έτσι για τον ταξινομητή του κοντινότερου γείτονα διαλέγουμε το συγκεκριμένο k. Θα μπορούσαμε να διαλέξουμε το k στο οποίο ελαχιστοποιείται το kloss, όμως τότε θα υπήρχε η πιθανότητα να καταλήξουμε σε ένα ταξινομητή που να κάνει over fit στο training set και άρα να μη δίνει υψηλά ποσοστά επιτυχίας κατά το testing. Με αυτή την επιλογή k εξασφαλίζουμε ότι ο ταξινομητής μας θα δίνει ισορροπημένα ποσοστά επιτυχίας για κάθε κατηγορία. Τα αποτελέσματα από τη διαδικασία του testing παρουσιάζονται στον Πίνακα 5.1.

5.1.2. Δέντρα Απόφασης (Decision Trees-DT)

Εδώ προσπαθούμε να βρούμε το βέλτιστο ταξινομητή για τη μέθοδο των δέντρων απόφασης δοκιμάζοντας διάφορες τιμές για το μέγιστο αριθμό διαιρέσεων (MaxNumSplits) από 1 μέχρι 40 (όπως παρατηρούμε από το Σχ.5.3 ότι το kloss παραμένει περίπου σταθερό μετά το 20) και το ελάχιστο μέγεθος εσωτερικού κόμβου (MinNodeSize) από 5 μέχρι 50 (όπως παρατηρούμε από το Σχ. 5.3 ότι το kloss δεν διαφέρει σημαντικά για ελάχιστο μέγεθος κόμβου μεγαλύτερο του 5).



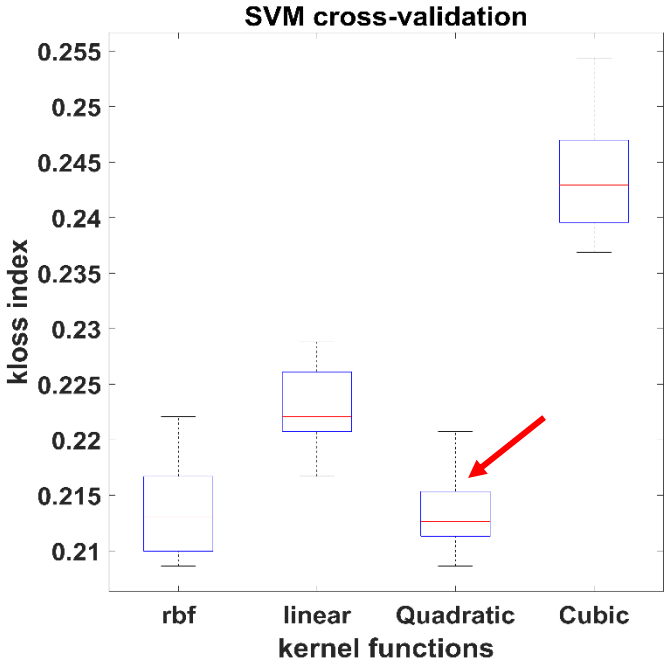
Σχήμα 5.3 Μέσο kloss του 10-fold cross validation για τις διαφορετικές τιμές του MaxNumSplits. Τα διαφορετικά χρώματα απεικονίζουν τις διαφορετικές τιμές του MinNodeSize.

Όπως μπορούμε να παρατηρήσουμε από το Σχ. 5.3 ο δείκτης kloss σταθεροποιείται μετά το MaxNumSplits = 20 για κάθε MinNodeSize εκτός του 5. Έτσι για τον ταξινομητή του δέντρου απόφασης διαλέγουμε το συγκεκριμένο MaxNumSplits και MinNodeSize = 50, διότι η γραφική (μπλε) που προκύπτει παρουσιάζει την ελάχιστη διακύμανση. Θα μπορούσαμε να διαλέξουμε το ζευγάρι παραμέτρων που ελαχιστοποιείται το kloss, όμως τότε θα υπήρχε η πιθανότητα όπως και προηγουμένως να καταλήξουμε σε ένα ταξινομητή που να κάνει over fit στο training set και άρα να μη δίνει υψηλά ποσοστά επιτυχίας κατά το testing. Με αυτή την επιλογή παραμέτρων εξασφαλίζουμε επίσης ότι ο ταξινομητής μας

θα δίνει ισορροπημένα ποσοστά επιτυχίας για κάθε κατηγορία. Τα αποτελέσματα για το testing παρουσιάζονται στον Πίνακα 5.1.

5.1.3. Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines-SVM)

Εδώ προσπαθούμε να βρούμε το βέλτιστο ταξινομητή για τη μέθοδο των SVM δοκιμάζοντας διαφορετικούς τύπους συναρτήσεων πυρήνα (kernel function): Gaussian radial basis function (rbf), linear, quadratic και cubic. Στο Σχ. 5.4, παρουσιάζουμε το kloss που προέκυψε για κάθε συνάρτηση πυρήνα.



Σχήμα 5.4 Μετρική kloss του 10-fold cross validation για τις διαφορετικές συναρτήσεις πυρήνα για τις 10 διαφορετικές επαναλήψεις της διαδικασίας.

Όπως μπορούμε να παρατηρήσουμε από το Σχήμα 5.4 ο δείκτης kloss ελαχιστοποιείται για την τετραγωνική συνάρτηση πυρήνα (βέλος) και μάλιστα παρουσιάζει τη μικρότερη διασπορά για τις 10 φορές που επαναλάβαμε τη διαδικασία. Με αυτή την επιλογή συνάρτηση πυρήνα εξασφαλίζουμε ότι ο ταξινομητής δε θα μας κάνει over fit και θα

δίνει ισορροπημένα ποσοστά επιτυχίας για κάθε κατηγορία. Τα αποτελέσματα για το testing παρουσιάζονται στον Πίνακα 5.1.

5.1.4. Ταξινομητής Bayes (Naïve Bayes-NB)

Για τον ταξινομητή Bayes διαλέξαμε διαφορετική κατανομή για κάθε χαρακτηριστικό του δείγματός μας. Για τα πρώτα 208 κατηγορικά χαρακτηριστικά επιλέξαμε την πολυωνυμική κατανομή, εφόσον αυτά τα χαρακτηριστικά είναι κατηγορικά. Για τα υπόλοιπα χαρακτηριστικά υποθέτουμε ότι ακολουθούν κανονική κατανομή. Τα αποτελέσματα για το testing της μεθόδου παρουσιάζονται στον Πίνακα 5.1.
Συνοπτικά παρουσιάζονται στον Πίνακα 5.1 τα αποτελέσματα που προκύπτουν κατά την ταξινόμηση των δειγμάτων του test set. Η διαδικασία επαναλήφθηκε 10 φορές και ο αριθμός των δειγμάτων από κάθε κατηγορία επιλέχθηκε να είναι ίσος με το 10% των δειγμάτων της κατηγορίας cis, όπως αναφέραμε και παραπάνω. Για κάθε μετρική παίρνουμε τη μέση τιμή των 10 επαναλήψεων. Όπως μπορούμε να παρατηρήσουμε τα καλύτερα αποτελέσματα επιτυγχάνονται από τον ταξινομητή NB. Όμως και ο ταξινομητής DT παρά το γεγονός ότι δίνει χαμηλότερο f measure σε σχέση με το NB, είναι αρκετά αποτελεσματικός εφόσον μπορεί και προβλέπει σχεδόν με το ίδιο ποσοστό επιτυχίας (sensitivity 0.8 και specificity 0.78) και τις δύο κατηγορίες. Επίσης, αξιοσημείωτο είναι ότι ο ταξινομητής SVM έχει υψηλότερο ποσοστό επιτυχίας στην κλάση trans, πράγμα που υποδεικνύει ότι πάρα το γεγονός ότι εκπαιδεύεται σε ένα δείγμα στο οποίο δεν περιλαμβάνεται η πληροφορία όλης της κατηγορίας trans, λόγω του under sampling, καταφέρνει και μαθαίνει να την προβλέπει σωστά πιο αποδοτικά (sensitivity 0.75 και specificity 0.78). Αυτό το πρόβλημα, θα το αντιμετωπίσουμε μέσω της μεθόδου ensemble για κάθε ταξινομητή, εκπαιδεύοντας το βέλτιστο ταξινομητή κάθε ομάδας με διαφορετικό training set.

Πίνακας 5.1 Αποτελέσματα αξιολόγησης των επιλεγμένων ταξινομητών.

Ταξινομητές	Sensitivity	Precision	Specificity	F-measure	Accuracy
kNN	0.69	0.68	0.67	0.68	0.68
DT	0.8	0.79	0.78	0.79	0.79
SVM	0.75	0.77	0.78	0.76	0.76
NB	0.86	0.79	0.77	0.81	0.82

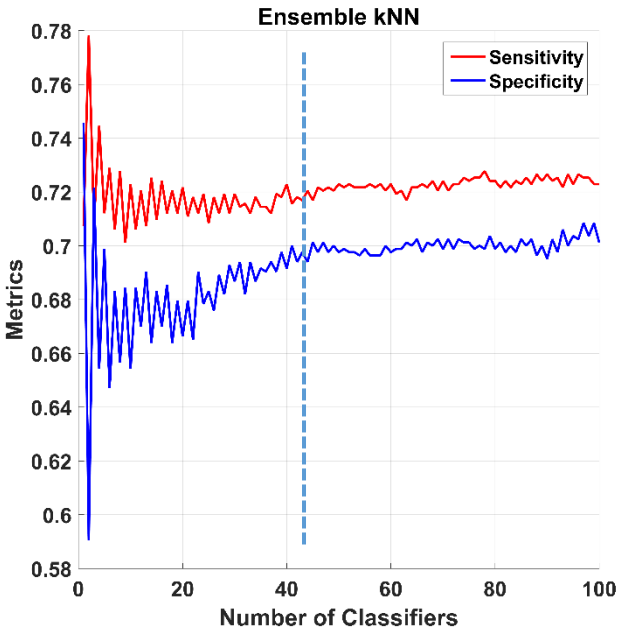
5.2. Μέθοδος Ensemble Ταξινομητών

Αρχικά διαχωρίσαμε το δείγμα μας σε training set και test set, όπως και προηγουμένως. Η κατηγορία trans αποτελεί το 94.7% (14875) ενώ η κατηγορία cis αποτελεί το 5.3% (826) του δείγματος μας. Συγκεκριμένα, διαχωρίσαμε το δείγμα μας σε 90% για training (743 δείγματα cis και 13388 δείγματα trans) και σε 10% για testing για κάθε κατηγορία. Έπειτα για να αντιμετωπίσουμε την ανισορροπία του δείγματος, πήραμε τυχαία δείγματα από την κατηγορία trans, ίσα με τον αριθμό των δειγμάτων του training set της κατηγοράις cis, δηλαδή 743. Αυτή τη διαδικασία την επαναλαμβάνουμε 100 φορές για να δημιουργήσουμε το training set για το ensemble κάθε ταξινομητή (knn, naïve bayes και support vector machine). Επομένως, για την εκπαίδευση του ensemble αυτών των ταξινομητών χρησιμοποιήσαμε τη μέθοδο Bootstrap Aggregating (bagging). Και έπειτα κατά την πρόβλεψη του κάθε ensemble κάθε ταξινομητής ψήφιζε ισότιμα σε σχέση με τους υπολοίπους για το αν ένα δείγμα είναι cis ή trans και το δείγμα κατηγοριοποιούνταν σύμφωνα με τον κανόνα της πλειοψηφίας (majority rule). Για την εκπαίδευση του ensemble δέντρων απόφασης χρησιμοποιήσαμε τη μέθοδο boosting. Και έπειτα κατά την πρόβλεψη ακολουθήσαμε την ίδια τεχνική. Για να αξιολογήσουμε αξιόπιστα τους ταξινομητές επαναλάβαμε τη διαδικάσία 10 φορές δημιουργώντας τυχαία 10 διαφορετικά training και test sets, κρατώντας σταθερό το training set της κατηγορίαςς cis και κάνοντας under sampling από τα δείγματα της κατηγορίας trans.

Έτσι δημιουργήσαμε τα ensemble για κάθε είδος ταξινομητή χρησιμοποιώντας την παραμετροποίηση των ταξινομητών στην οποία καταλήξαμε στο προηγούμενο βήμα.

5.2.1. KNN Ensemble

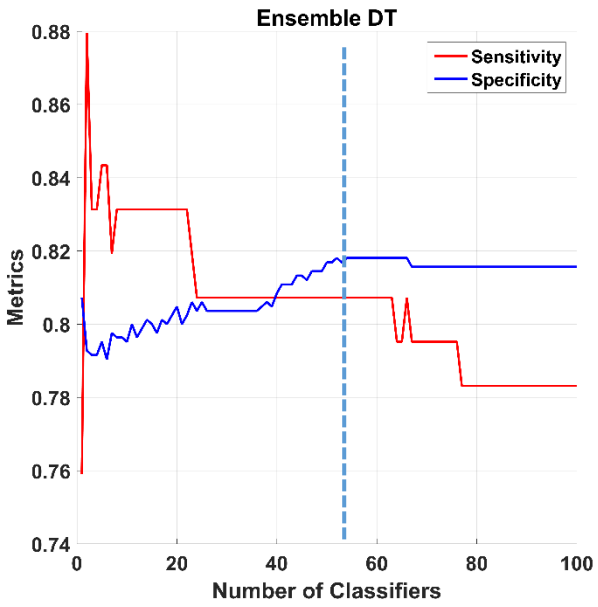
Εδώ δημιουργήσαμε ensemble kNN με μέγεθος από 1 μέχρι το 100. Από το Σχ.5.5 μπορούμε να παρατηρήσουμε ότι το ensemble, από τους 50 ταξινομητές και μετά έχει περίπου σταθερά ποσοστά επιτυχίας, και υψηλότερα σε σχέση με τον απλό ταξινομητή kNN που παρουσιάσαμε παραπάνω και για τις δύο κατηγορίες. Παρόλα αυτά το ensemble αυτό δεν καταφέρνει να προβλέψει με υψηλότερο ποσοστό την κατηγορία trans όπως ήταν αναμενόμενο σύμφωνα με τη φύση των δεδομένων μας. Τα αποτελέσματα συνοψίζονται στον Πίνακα 5.2.



Σχήμα 5.5 Οι μετρικές sensitivity και specificity για τα ensemble του kNN.

5.2.2. Decision Tree Ensemble

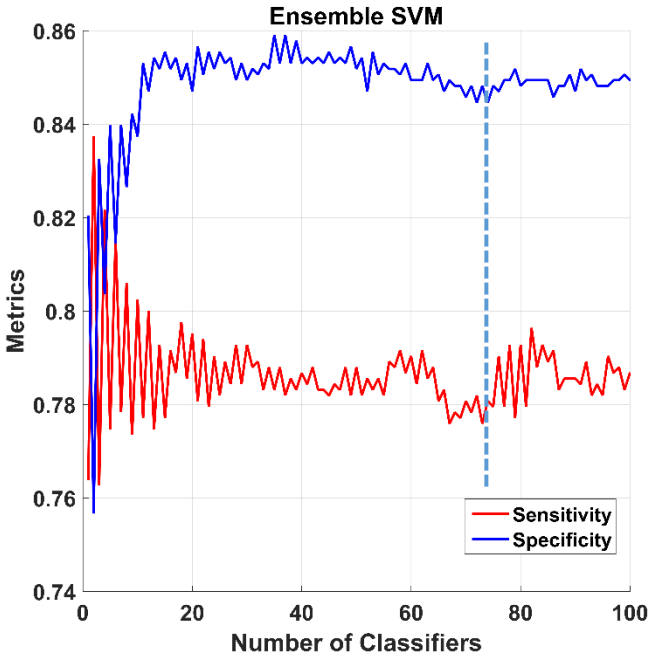
Εδώ δημιουργήσαμε ensemble από δέντρα απόφασης με μέγεθος από 1 μέχρι το 100. Από το Σχ. 5.6 μπορούμε να παρατηρήσουμε ότι το ensemble, από τους 40 ταξινομητές και μετά καταφέρνει να προβλέψει την κατηγορία trans με μεγαλύτερο ποσοστό επιτυχίας, πράγμα που είναι το επιθυμητό με βάση τα δεδομένα μας. Επίσης τα ποσοστά πρόβλεψης και των δύο κατηγοριών βελτιώνονται σε σχέση με τον απλό ταξινομητή δέντρου απόφασης που παρουσιάσαμε παραπάνω. Τα αποτελέσματα αυτά συνοψίζονται στον Πίνακα 5.2.



Σχήμα 5.6 Οι μετρικές sensitivity και specificity για τα ensemble των δέντρων απόφασης.

5.2.3. Support Vector Machine Ensemble

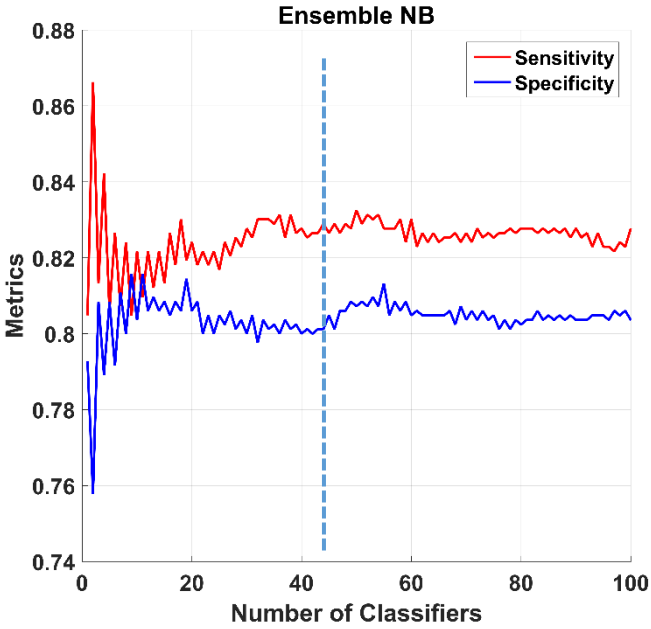
Εδώ δημιουργήσαμε ensemble από μηχανές διανυσμάτων υποστήριξης με μέγεθος από 1 μέχρι το 100. Από το Σχ. 5.7 μπορούμε να παρατηρήσουμε ότι το ensemble, από τους 80 ταξινομητές και μετά έχει περίπου σταθερά ποσοστά επιτυχίας, και υψηλότερα σε σχέση με τον απλό ταξινομητή svm που παρουσιάσαμε παραπάνω και για τις δύο κατηγορίες. Είναι αξιοσημείωτο ότι το ensemble αυτό συνεχίζει να προβλέπει με υψηλότερο ποσοστό την κατηγορία trans όπως ήταν αναμενόμενο σύμφωνα με τη φύση των δεδομένων μας, αλλά και με τον τρόπο λειτουργίας του svm. Τα αποτελέσματα συνοψίζονται στον Πίνακα 5.2.



Σχήμα 5.7 Οι μετρικές sensitivity και specificity για τα ensemble των μηχανών διανυσμάτων υποστήριξης.

5.2.4. Naïve Bayes Ensemble

Εδώ δημιουργήσαμε ensemble naïve Bayes με μέγεθος από 1 μέχρι το 100. Από το Σχ. 5.8 μπορούμε να παρατηρήσουμε ότι το ensemble, από τους 50 ταξινομητές και μετά έχει περίπου σταθερά ποσοστά επιτυχίας, και υψηλότερα σε σχέση με τον απλό ταξινομητή naïve Bayes που παρουσιάσαμε παραπάνω και για τις δύο κατηγορίες. Παρόλα αυτά το ensemble αυτό δεν καταφέρνει να προβλέψει με υψηλότερο ποσοστό την κατηγορία trans η οποία θα έπρεπε να προβλεπόταν με υψηλότερη ακρίβεια εφόσον αποτελεί το 95% των δεδομένων μας . Τα αποτελέσματα συνοψίζονται στον Πίνακα 5.2.



Σχήμα 5.8 Οι μετρικές sensitivity και specificity για τα ensemble των ταξινομητών Bayes.

Συνοπτικά παρουσιάζονται στον Πίνακα 5.2 τα αποτελέσματα που προκύπτουν κατά την ταξινόμηση των δειγμάτων του test set από κάθε ensemble. Η διαδικασία επαναλήφθηκε 10 φορές και ο αριθμός των δειγμάτων από κάθε κατηγορία επιλέχθηκε να είναι ίσος με το 10% των δειγμάτων της κατηγορίας cis, όπως αναφέραμε και παραπάνω. Για κάθε μετρική παίρνουμε τη μέση τιμή των 10 επαναλήψεων. Από τον πίνακα 5.2, μπορούμε εύκολα να συμπεράνουμε ότι τα ποσοστά επιτυχίας των ταξινομητών μας έχουν βελτιωθεί με το σχήμα ensemble σε σχέση με τους μεμονωμένους ταξινομητές και το under sampling. Συγκεκριμένα, όπως μπορούμε να παρατηρήσουμε τα καλύτερα αποτελέσματα επιτυγχάνονται από το ensemble του ταξινομητή NB και εδώ. Όμως το ensemble των SVM και των δέντρων απόφασης είναι πιο αξιόπιστα γιατί αποτυπώνουν καλύτερα την εικόνα των δεδομένων μας. Δηλαδή εφόσον έχουμε την μια κατηγορία σε μεγάλη περίσσεια θα θέλαμε αυτή να προβλέπεται με μεγαλύτερα επιτυχία αν όχι ίση με την κατηγορία που βρίσκεται σε έλλειμμα. Αυτοί οι δύο ταξινομητές το καταφέρνουν αυτό. Επομένως, θα ξεχωρίζαμε αυτούς τους δύο σε σχέση με τους υπολοίπους. Ίσως το ensemble των δέντρων απόφασης να είναι ο ιδανικός ταξινομητής για αυτό το πρόβλημα γιατί καταφέρνει να προβλέπει περίπου ισάξια και τις δύο κατηγορίες (sensitivity 0.81 και specificity 0.82). Συμπερασματικά, βλέπουμε ότι το σχήμα ensemble δε βοηθάει μόνο να βελτιώσουμε τα ποσοστά προβλέψεις μεμονωμένων αλγορίθμων, αλλά καταφέρνει να έχει

μια ολοκληρωμένη εικόνα του δείγματος μας εφόσον έρχεται σε επαφή με περισσότερα δείγματα σε σχέση με την πρώτη μέθοδο που ακολουθήσαμε.

Πίνακας 5.2 Αποτελέσματα αξιολόγησης των ensemble ταξινομητών.

Ensemble	Sensitivity	Precision	Specificity	F-measure	Accuracy
kNN (100)	0.72	0.71	0.70	0.72	0.71
DT (60)	0.81	0.82	0.82	0.81	0.81
SVM (80)	0.79	0.84	0.85	0.82	0.82
NB (50)	0.83	0.81	0.81	0.82	0.82

ΚΕΦΑΛΑΙΟ 6. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

Στην εργασία αυτή διερευνήσαμε την απόδοση διαφόρων ταξινομητών με σκοπό την πρόβλεψη του ισομερισμού της προλίνης σε cis ή trans σε πολυπεπτιδικές αλυσίδες πρωτεϊνών.

Οι ταξινομητές που χρησιμοποιήθηκαν ήταν ο ταξινομητής των k κοντινότερων γειτόνων, τα δέντρα απόφασης, οι μηχανές διανυσμάτων υποστήριξης, ο αφελής ταξινομητής του Bayes και οι συλλογικοί ταξινομητές που στηρίζονται στους προηγούμενους. Η μοντελοποίηση του προβλήματος έγινε με τη χρήση διαφόρων χαρακτηριστικών που τα εξάγαμε από τις πεπτιδικές αλυσίδες των πρωτεϊνών. Συγκεκριμένα, εξάγαμε ένα μοτίβο 11 αμινοξέων κάθε προλίνης (5 αμινοξέα πριν από μία προλίνη και 5 αμινοξέα μετά), την απόλυτη συχνότητα που εμφανίζονται τα 20 αμινοξέα και των αριθμό των νερών στη σφαίρα κάθε προλίνης (η οποία ορίζεται από το κέντρο προλίνης και συγκεκριμένη ακτίνα 5Α), τη δευτεροταγή δομή της αλυσίδας στο σημείο που βρίσκεται κάθε προλίνη. Επιπλέον υπολογίσαμε τη δίεδρη γωνία μεταξύ της προλίνης και του προηγούμενο αμινοξέος για να μπορούμε να χαρακτηρίσουμε αν μια προλίνη είναι ισομερισμένη σε cis ή trans.

Οι ταξινομητές που χρησιμοποιήθηκαν έδωσαν υψηλά ποσοστά επιτυχία για τα συγκεκριμένα χαρακτηριστικά. Αρχικά το πρόβλημα του μη ισορροπημένου δείγματος δεν επέτρεπε στους ταξινομητές μας να δώσει ικανοποιητικά αποτελέσματα. Αφού λύσαμε αυτό το πρόβλημα μέσω της υποδειγματοληψίας, για την οποία μιλήσαμε εκτενώς σε προηγούμενο κεφάλαιο, καταφέραμε να αυξήσουμε περισσότερο τα ποσοστά επιτυχίας χρησιμοποιώντας συλλογικούς ταξινομητές. Συγκεκριμένα, το ensemble των SVM και των δέντρων απόφασης δίνουν τα υψηλότερα αποτελέσματα. Όσον αφορά την πρώτη μέθοδο, δίνει υψηλά αποτελέσματα λόγω της εγγενούς ικανότητας των SVM να ξεχωρίζουν δύο

κατηγορίες και να μην εξαρτώνται τόσο από τον αριθμό των δειγμάτων κάθε μιας. Έτσι μέσω του ensemble, ο ταξινομητής αυτός κατάφερε να μάθει σε ικανοποιητικό βαθμό και τις δύο κατηγορίες αφού εκπαιδεύεται με μεγαλύτερο αριθμό δειγμάτων της κατηγορίας trans και να δίνει υψηλά ποσοστά επιτυχίας, της τάξης του 82% (f-measure). Όσον αφορά τη δεύτερη μέθοδο, αυτό συμβαίνει γιατί το ensemble των δέντρων απόφασης μέσω της μεθόδου boosting καταφέρνει και μαθαίνει εξίσου και τις δύο κατηγορίες παρόλο που το δείγμα μας δεν είναι ισορροπημένο, διότι σε κάθε επανάληψη εκμάθησης των δειγμάτων, λαμβάνει υπόψη τα δείγματα που ταξινομήθηκαν λανθασμένα στην προηγούμενη. Επομένως αυτή τη μέθοδο θα τη χαρακτηρίζαμε την πιο ισορροπημένη για την ασφαλή πρόβλεψη των δύο κατηγοριών. Το ensemble των δέντρων απόφασης είναι ο καταλληλότερος ταξινομητής για αυτό το πρόβλημα γιατί καταφέρνει να προβλέπει περίπου ισάξια και τις δύο κατηγορίες (sensitivity 0.81 και specificity 0.82) με λίγο χαμηλότερο ποσοστό f-measure σε σχέση με το ensemble των SVM 81%). Αυτοί οι δύο ταξινομητές εκτός από τα υψηλά ποσοστά επιτυχίας ικανοποιούν και την διαίσθηση μας όσον αφορά το πρόβλημα αυτό, διότι προβλέπουν με υψηλότερο ποσοστό επιτυχίας την κατηγορία που βρίσκεται σε πλεόνασμα στα δεδομένα μας.

Σε αυτή την εργασία διερευνήσαμε το πρόβλημα της πρόβλεψης του ισομερισμού των αμινοξέων προλίνης σε cis και trans με διάφορα ήδη ταξινομητών. Προσπαθώντας να διερευνήσουμε παραπάνω το ήδη υπάρχον πρόβλημα ταξινόμησης θα μπορούσαμε να εξετάσουμε την ποιότητα των χαρακτηριστικών που χρησιμοποιήσαμε ή και ακόμα να διερευνήσουμε καινούρια χαρακτηριστικά, όπως για παράδειγμα τη μέση απόσταση των αμινοξέων ή των νερών (από το κεντροειδές της προλίνης) που ανήκουν στην σφαίρα της προλίνης. Επίσης θα μπορούσαμε να διερευνήσουμε την απόδοση του ταξινομητή μας είτε χρησιμοποιώντας μοτίβα προλίνης με μεγαλύτερο μήκος είτε αυξάνοντας την ακτίνα στη σφαίρα της προλίνης, μέσα στον όγκο της οποίας ψάχναμε αμινοξέα και νερά. Επομένως συνδυάζοντας τα χαρακτηριστικά αυτά θα ήταν εφικτό να βελτιώσουμε επιπλέον τα ποσοστά επιτυχίας των ταξινομητών μας.

Επιπλέον, όπως είναι γνωστό από τη βιβλιογραφία πολλές προλίνες με συγκεκριμένα μοτίβα αμινοξέων μπορούν είτε να παρουσιάζουν ισομερισμό cis είτε ισομερισμό trans ανάλογα με πρωτεΐνη στην οποία βρίσκονται [2]. Αυτές οι προλίνες χαρακτηρίζονται ως switch γιατί στην πραγματικότητα περνάνε από το ένα είδος ισομερισμού στο άλλο. Έτσι θα ήταν χρήσιμο να εκμεταλλευτούμε τη μέθοδο που αναπτύξαμε για τον προσδιορισμό

προλινών switch και έπειτα την εκπαίδευση ταξινομητών που προβλέπουν αν μια προλίνη είναι switch ή όχι. Αυτό το πρόβλημα είναι πολύ ενδιαφέρον δεδομένης της πολύ μικρής συχνότητας με την οποία συμβαίνει ένα τέτοιο γεγονός. Αυτό οφείλεται φυσικά στο χαμηλό ποσοστό διαμορφώσεων trans στην προλίνη παρόλα αυτά αποτελεί μια πρόκληση που θα μπορούσε να εξεταστεί μελλοντικά.

ΑΝΑΦΟΡΕΣ

[1] Andreotti AH. "Native state proline isomerization: an intrinsic molecular switch", Biochemistry, 42:9515-9524, 2003.

[2] Berg JM., Tymoczko JL., Stryer L. «Βιοχημεία», Παν.Εκδόσεις Κρήτης, http://www.chem.uoa.gr/courses/undergraduate/biochem/mavri/und_stryer.htm

[3] Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N, Weissig H., Shindyalov I.N, Bourne B.E. " The Protein Data Bank Nucleic Acids Research", 28: 235-242, 2000.

[4] Bishop C.M. "Pattern Recognition and Machine Learning (Information Science and Statistics)", Springer, 2006.

[5] Branden C., Tooze J. "Introduction to Protein Structure (2ded.), Garland,1999.

[6] Burnham K., Anderson D. "Model Selection and Inference: A Practical Information Theoretic Approach", second eds, New York: Springer-Verlag, 2002.

[7] Chothia C., Finkelstein A. V. "The classification andorigin of protein folding patterns", Annu. Rev. Biochem, 59:1007-1039, 1990.

[8] Dugave C., Demange L. "Cis - trans isomerization of organic molecules and biomolecules: implications and applications", Chem Rev, 103:2475-2532, 2003.

[9] Eckert B., Martin A., Balbach J., Schmid FX. "Prolyl isomerization as a molecular timer in phage infection", Nat Struct Mol Biol , 12:619-623, 2005.

[10] Frömmel C., Preissner R. "Prediction of prolyl residues in cis-conformation in protein structures on the basis of the amino acid sequence", FEBS Lett , 277:159-163, 1990.

[11] Gutierrez Juliette, Leroy Gondy. "Using Decision Trees to Predict Crime Reporting", Claremont Graduate University.

[12] Jabs A, Weiss M.S., Hilgenfeld R., J Mol Biol. "Non-proline cis peptide bonds in protein", 286:291-304, 1999

[13] Kang YK., Choi H.Y. “Cis - trans isomerization and puckering of proline residue”, Biophys Chem , 111:135-142, 2004.

[14] Kabsch W., Sander C. "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features", Biopolymers **22** (12): 2577–637, 1983.

[15] Lorenzen S., Peters B., Goede A., Preissner R., Frömmel C. “Conservation of cis prolyl bonds in proteins during evolution”, Proteins, 58:589-595, 2005.

[16] Mitchell T.M. “Machine Learning”, McGraw-Hill Higher Education, Boston, 1997.

[17] Opitz David, Richard Maclin. “Popular Ensemble Methods: An Empirical Study, Journal of Artificial Intelligence Research”, Al Access Foundation and Morgan Kaufmann, 11: 169-198, 1999.

[18] Pahlke D., Freund C., Leitner D., Labudde D. “ Statistically significant dependence of the Xaa-Pro peptide bond conformation on secondary structure and amino acid sequence”, BMC Struct Biol, 5:1-8, 2005.

[19] Pahlke D., Leitner D., Wiedemann U., Labudde D. “COPS- cis/trans peptide bond conformation prediction of amino acids on the basis of secondary structure information”, Bioinformatics , 21:685-686, 2005.

[20] Pall D., Chakraabarti P. “Cis peptide bonds in proteins: residues involved, their conformation, interaction and locations”, J Mol Biol , 294:271-288, 1999.

[21] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. “Intoduction to Data Mining”, Addison Wesley, 2006.

[22] Qian J., Lin J., Luscombe N.M., Yu H., Gerstein M. “Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data”, Bioinformatics , 19:1917–1926, 2003.

[23] Regan L. “Protein structure: Born to be beta”, Curr. Biol, 4:656-658, 1994.

[24] Reimer U., Fischer G. “Local structural changes caused by peptidyl-prolyl cis/trans isomerization in the native state of proteins”, Biophys Chem, 96:203-212, 2002.

[25] Reimer U., Scherer G., Drewello M., Kruber S., Schutkowski M., Fischer G. "Side-chain effects on peptidyl-prolyl cis/trans isomerization", J. Mol. Biol., 279:449-460, 1998.

[26] Richards F. M. "The protein folding problem", Sci. Am, 264(1): 54-57, 1991.

[27] Richardson J. S., Richardson D. C., Tweedy N. B., Gernert K.M., Quinn T. P., Hecht M. H., Erickson B. W., Yan Y., McClain R. D., Donlan M. E. and Suries M. C. "Looking at proteins: Representations, folding, packing, and design", Biophys J., 63: 1186-1220, 1992.

[28] Ripley B.D. "Pattern Recognition and Neural Networks", New York: Cambridge University Press, 1996.

[29] Rokach Lior. "Ensemble – based classifiers", Artif Intell Rev., 33:1 – 39, Springer Science & Business Media, 2010.

[30] Rong She, Je_rey Shih-Chieh Chu, Ke Wang, Nansheng Chen. "Fast and Accurate Gene Prediction by Decision Tree Classification", School of Computing Science, Department of Molecular Biology and Biochemistry, Simon Fraser University, Canada.

[31] Schmid F.X., Mayr L.M., Mücke M., Schönbrunner E.R. "Prolyl isomerases: role in protein folding", Advan Protein Chem , 44:25-66,1993.

[32] Schultz, G. E. and Schirmer, R. H. "Principles of Protein Structure. Springer-Verlag, 1979.

[33] Srinivasan R. and Rose G. D. "A physical basis for protein secondary structure", Proc. Natl. Acad. Sci. USA, 96: 14258-14263, 1999.

[34] Stewart D.E., Sarkar A., Wampler J.E. "Occurrence and role of cis peptide bonds in protein structures", J. Mol. Biol., 214:253-260,1990.

[35] Theodoridis S. and Koutroumbas K. "Pattern Recognition", 2nd ed., Academic Press, 2006.

[36] Wang Guoli and Roland L. Dunbrack, Jr. "PISCES: a protein sequence culling server Bioinformatics", 19 (12): 1589-1591, 2003.

[37] Wang M.L., Li W.J, Xu W.B. "Support vector machines for prediction of peptidyl prolyl cis/trans isomerization", J. Peptide Res., 63:23-28, 2004.

[38] Weber A. L. and Miller S. L. “Reasons for the occurrence of the twenty coded protein amino acids”, J. Mol. Evol., 17:273-284, 1981.

[39] Wedemeyer W.J., Welker E., Scheraga H.A. “Proline cis-trans isomerization and protein folding”, Biochemistry, 41:14637-14644, 2002.

[40] Weiss M.S., Jabs A., Hilgenfeld R. “Peptide bonds revisited”, Nat. Struct. Biol., 5:676, 1998.

[41] Weiss S. I. and Kulikowski, C. “Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning and Expert Systems”, San Francisco, Calif.: Morgan Kaufmann, 1991.

[42] Witten I.H. and Frank E. “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”, second eds., Morgan Kaufmann, 2005.

[43] Wu Y., Matthews C.A. “Cis-prolyl peptide bond isomerization dominates the folding of the alpha subunit of trp synthase”, a TIM barrel protein, J. Mol. Biol., 322:7-13, 2002.

ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ

Η Παρασκευή Πασχάλη γεννήθηκε το 1987 στην Πρέβεζα. Το 2011 απέκτησε Πτυχίο από το Τμήμα Πληροφορικής του Ιονίου Πανεπιστημίου. Από το 2012 παρακολουθεί το μεταπτυχιακό πρόγραμμα του τμήματος Μηχανικών Η/Υ και Πληροφορικής της Σχολής Θετικών Επιστημών του Πανεπιστημίου Ιωαννίνων με κατεύθυνση Τεχνολογίες και Εφαρμογές.