Α Ρ Ι Σ Τ Ο Τ Ε Λ Ε Ι Ο Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Θ Ε Σ Σ Α Λ Ο Ν Ι Κ Η Σ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Μέθοδοι επιλογής χαρακτηριστικών για βιολογικά δεδομένα

ΙΝΤΖΕΒΙΔΟΥ ΑΙΚΑΤΕΡΙΝΗ

ΑΕΜ: 2240

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΒΛΑΧΑΒΑΣ ΙΩΑΝΝΗΣ, ΚΑΘΗΓΗΤΗΣ

ΘΕΣΣΑΛΟΝΙΚΗ, ΣΕΠΤΕΜΒΡΙΟΣ 2017

Πρόλογος

Η εξέλιξη της βιοτεχνολογίας έχει οδηγήσει πια στην παραγωγή μεγάλων ποσοτήτων δεδομένων με σημαντικά εύκολο και φθηνό τρόπο. Προκύπτουν, λοιπόν, μεγάλες βάσεις δεδομένων, τις οποίες δε γίνεται να διαχειριστούμε με συμβατικές μεθόδους επεξεργασίας. Αυτό το κενό έρχονται να συμπληρώσουν οι περιοχές της Εξόρυξης Δεδομένων και της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων, οι οποίες περιέχουν τεχνολογίες και αλγορίθμους σχεδιασμένους για τις προκλήσεις των *μεγάλων δεδομένων* (big data).

Το πρόβλημα το οποίο η παρούσα πτυχιακή προσπαθεί να επιλύσει, προέρχεται από τον τομέα της Βιοπληροφορικής και πιο συγκεκριμένα της Πληθυσμιακής Γενετικής. Πρόκειται για την επιλογή των πιο πληροφοριακών δεικτών, από ένα σύνολο Πολυμορφισμών Μονών Νουκλεοτιδίων (Single Nucleotide Polymorphism - SNP).

Για την επιλογή των πιο πληροφοριακών δεικτών, χρησιμοποιούνται μέθοδοι εμπνευσμένες από τη θεωρία κοινωνικής επιλογής, οι οποίες συνδυάζουν τα αποτελέσματα ήδη υπαρχόντων γενετικών μεθόδων, με στόχο, φυσικά, καλύτερα αποτελέσματα. Πέραν αυτού, χρησιμοποιείται και η βαθμολογία του κάθε γενετικού δείκτη ως κριτήριο για την εύρεση των πιο πληροφοριακών.

Με βάση τα αποτελέσματα της επιλογής των δεικτών δημιουργούνται στη συνέχεια μοντέλα ανάθεσης ατόμων σε πληθυσμούς. Από την εκτίμηση της ακρίβειας ανάθεσης των μοντέλων αυτών, μπορεί τελικά να συγκριθεί αν η επιλογή δεικτών οδήγησε σε καλύτερα αποτελέσματα.

Τελικά, στόχο της διπλωματικής αποτελεί η παροχή μεθόδων, με τις οποίες θα επιλέγεται μικρότερος αριθμός δεικτών, διατηρώντας, όμως αρκετή πληροφορία, ώστε τα αποτελέσματα που προκύπτουν από τη χρήση τους να είναι επαρκώς ακριβή.

Feature Selection Methods for Biological Data

Abstract

The era of Big Data is upon us and the field of Biology has not stayed unaffected. Biological Data have reached enormous numbers, mostly due to the advances in biotechnology. As the size of these data make their processing via traditional methods basically impossible, the need for more efficient ways of storing and processing has arisen. Computer science fields like Data Mining and Knowledge Discovery from Databases are based exactly on research and applications that make the processing of these data affordable, both economically and computationally.

This dissertation is aiming to solve a problem coming from the field of Population Genetics, a sub-domain of Biology, from a computer science viewpoint. This problem is about selecting the most informative markers from population genomic data, and in terms of Machine Learning it is a feature selection problem. This is an essential pre-processing step to the assignment of individuals into groups of origin in an accurate and efficient manner. To do that, we are using Single Nucleotide Polymorphisms (SNPs) as the specific markers used, along with a dataset coming from pig populations.

This problem was approached from a scope utilizing rank aggregation methods that originate from social choice and voting theory. We initially use genetic rankings of SNPs, apply some rank aggregation methods on them, therefore creating new rankings, based on two different criteria. These are the rank of each SNP in the base rankers and its score. All rankings are later on used to create prediction models that assign individuals in populations. To test the results in comparison with those of the base rankers, we calculate the assignment accuracy of all models.

Finally, we conclude that results coming from the aggregated rankings are performing somewhat better for some numbers of SNPs, indicating that there still is space for future research on this matter.

Πίνακας Περιεχομένων

Πρόλογος 5
Abstract 7
Πίνακας Περιεχομένων 9
Κεφάλαιο 1: Εισαγωγή 11
Κεφάλαιο 2: Μηχανική Μάθηση και Ανακάλυψη Γνώσης από Βάσεις Δεδομένων 13
2.1 Μηχανική Μάθηση 13
2.1.1. Επιβλεπόμενη Μάθηση 14
2.1.2. Μη επιβλεπόμενη Μάθηση 15
2.1.3. Ενισχυτική Μάθηση 18
2.2 Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων 19
2.2.1. Διαδικασία Ανακάλυψης Γνώσης 20
2.2.2. Προβλήματα στην ανακάλυψη γνώσης 25
Κεφάλαιο 3: Βιοπληροφορική 29
3.1 Ερευνητικοί Τομείς της Βιοπληροφορικής 30
3.1.1. Ανάλυση Βιολογικών Αλληλουχιών (Sequence analysis) 30
3.1.2. Γονιδιακή Έκφραση (Gene expression) 31
3.1.3. Γενετική και Γενωμική (Genetics and genomics) 31
3.1.4. Πρωτεωμική (Proteomics) 32
3.2 Βάσεις δεδομένων στη Βιοπληροφορική 32
3.3 Ανακάλυψη Γνώσης Στη Βιοπληροφορική 33
3.4 Ανάπτυξη εργαλείων στο πλαίσιο της Βιοπληροφορικής 34
3.5 Πληθυσμιακή Γενετική και Γενετικοί Δείκτες 35
3.5.1. Ανάλυση Πολυμορφισμών Μονών Νουκλεοτιδίων 36
Κεφάλαιο 4: Κατατάξεις και Μέθοδοι Συγκερασμού Κατατάξεων 39
4.1. Κατατάξεις και Συγκερασμός Κατατάξεων 39
4.2. Μέθοδοι συγκερασμού κατατάξεων 41
4.2.1. Μέθοδοι Borda 42
4.2.2. Μέθοδοι Markov Chain 45
Κεφάλαιο 5: Υλοποίηση και Πειραματική Διαδικασία 49

5.1. Υλοποίηση 49
5.2. Πειραματική Διαδικασία 59
Κεφάλαιο 6: Επίλογος και Μελλοντική Εργασία 65
6.1. Συμπεράσματα 65
6.2. Μελλοντικές Επεκτάσεις 66
Βιβλιογραφία 69

Κεφάλαιο 1: Εισαγωγή

Με την εξέλιξη των τεχνολογικών μέσων και την πρόοδο στις μεθόδους στον τομέα των βιοεπιστημών, η παραγωγή δεδομένων είναι πια ευκολότερη και λιγότερο ακριβή. Κατά συνέπεια, δημιουργούνται καθημερινά μεγάλα σύνολα βιολογικών δεδομένων, οδηγώντας σε μεγέθη δεδομένων τέτοια, ώστε οι κλασικές μέθοδοι διαχείρισης και επεξεργασίας δεδομένων δεν αρκούν. Για το λόγο αυτό, προκύπτει η ανάγκη για την ανάμιξη του τομέα της Εξόρυξης Δεδομένων και Ανακάλυψης Γνώσης και τη χρήση των μεθόδων που προτείνει σε βιολογικά δεδομένα.

Η παρούσα πτυχιακή, λοιπόν, προσπαθεί να λύσει ένα πρόβλημα της Βιοπληροφορικής. Πιο συγκεκριμένα, μελετάται η ανάθεση ατόμων σε πληθυσμούς βάσει της γενετικής τους σύστασης, η οποία περιγράφεται εδώ με τη μορφή των SNPs. Ελέγχεται, επομένως, αν μπορεί η ανάθεση αυτή να γίνει αποδοτικότερα – με μεγαλύτερη ακρίβεια για μικρό αριθμό SNPs - με τη βοήθεια μεθόδων προ-επεξεργασίας και συνδυασμού επιλογής χαρακτηριστικών. Οι μέθοδοι αυτές χρησιμοποιούνται για τη δημιουργία κατατάξεων των πιο πληροφοριακών SNPs, οι οποίες στη συνέχεια χρησιμοποιούνται στη δημιουργία μοντέλων ανάθεσης ατόμων σε πληθυσμούς. Τελικά, αξιολογείται κατά πόσο οι κατατάξεις που προέκυψαν από το συνδυασμό επιλογής χαρακτηριστικών οδήγησαν σε μοντέλα με καλύτερη επίδοση.

Το παρόν κείμενο είναι δομημένο σε 6 κεφάλαια όπου αρχικά παρουσιάζουν εν συντομία τον ευρύτερο τομέα στον οποίο στηρίζεται η παρούσα πτυχιακή εργασία, στη συνέχεια εμβαθύνουν στις συγκεκριμένες μεθόδους που υλοποιήθηκαν και χρησιμοποιήθηκαν, και δίνουν λεπτομέρειες για την υλοποίηση και την πειραματική διαδικασία της αξιολόγησης των αποτελεσμάτων. Τελικά, παρουσιάζονται τα αποτελέσματα και η ερμηνεία τους, καθώς και μία σύντομη συζήτηση για μελλοντικές εργασίες πάνω στο συγκεκριμένο ζήτημα.

Πιο συγκεκριμένα, στα κεφάλαια 2 και 3 παρουσιάζονται οι περιοχές στις οποίες θεωρητικά στηρίζεται η παρούσα διπλωματική, της Μηχανικής Μάθησης και Ανακάλυψης Δεδομένων, και της Βιοπληροφορικής, αντίστοιχα. Γίνεται, λοιπόν λόγος για τα είδη

μηχανικής μάθησης και τις ομάδες μεθόδων που χρησιμοποιούνται, τα στάδια της διαδικασίας ανακάλυψης γνώσης, και τις προκλήσεις που εμφανίζονται κατά τη διαδικασία αυτή. Στο κεφάλαιο 3 παρουσιάζονται συνοπτικά οι ερευνητικοί τομείς της Βιοπληροφορικής, οι βάσεις δεδομένων και η ανακάλυψη δεδομένων στο πλαίσιό της, καθώς και κατηγορίες εργαλείων. Επίσης γίνεται αναφορά στον τομέα της Πληθυσμιακής Γενετικής και στους γενετικούς δείκτες, από όπου προκύπτει το πρόβλημα το οποίο επιχειρεί να λύσει η διπλωματική εργασία.

Στο κεφάλαιο 4 παρουσιάζονται οι κατατάξεις και ο συγκερασμός των κατατάξεων, που προέρχονται από τον τομέα της θεωρίας κοινωνικής επιλογής. Έμφαση δίνεται στις μεθόδους συγκερασμού που υλοποιήθηκαν, οι οποίες αναλύονται λεπτομερώς.

Στο κεφάλαιο 5 αναλύονται η υλοποίηση και πειραματική διαδικασία που ακολουθήθηκε. Πιο συγκεκριμένα, παρουσιάζεται η υλοποίηση των μεθόδων που περιγράφηκαν στο προηγούμενο κεφάλαιο, καθώς και των μεθόδων με τις οποίες έγινε η αξιολόγηση των αποτελεσμάτων. Εδώ ο αναγνώστης βρίσκει λεπτομέρειες για τη δομή του κώδικα, αλλά και τη λειτουργία του, μέθοδο προς μέθοδο. Επιπλέον, περιγράφεται η διαδικασία δημιουργίας μοντέλων πρόβλεψης με βάση τις κατατάξεις, τόσο τις αρχικές γενετικές, όσο και εκείνες που προέκυψαν από το συγκερασμό, και τον έλεγχο της επιτυχίας των μοντέλων αυτών στη σωστή ανάθεση των ατόμων σε πληθυσμούς.

Τέλος, στο κεφάλαιο 6 εμφανίζονται τα αποτελέσματα, δηλαδή η ακρίβεια ανάθεσης του κάθε μοντέλου πρόβλεψης ανά δεκάδα SNP και ανά μέθοδο, και μελετάται αν τελικά ο συγκερασμός των γενετικών κατατάξεων βοήθησε στη δημιουργία αποδοτικότερων μοντέλων. Επίσης γίνεται συζήτηση για επεκτάσεις και περαιτέρω έρευνα.

Κεφάλαιο 2: Μηχανική Μάθηση και Ανακάλυψη Γνώσης από Βάσεις Δεδομένων

Οι έννοιες της ευφυΐας, της γνώσης και της μάθησης είναι στενά συνδεδεμένες με τον άνθρωπο. Η ικανότητα του να σκέφτεται και να μαθαίνει, καθώς και ο τρόπος με τον οποίο μπορεί πιο αποδοτικά να μάθει, έχουν αποτελέσει θέμα συζήτησης ανά τους αιώνες και έχουν μελετηθεί από διάφορους επιστημονικούς κλάδους συμπεριλαμβανομένων των κλάδων της ψυχολογίας, της παιδαγωγικής, της βιολογίας και της ιατρικής. Με την εξέλιξη της τεχνολογίας, επανεξετάζεται η ικανότητα μάθησης, αυτή τη φορά με τις μηχανές να κατέχουν το ρόλο του υποκειμένου μάθησης. Με αυτόν τον τρόπο εμφανίζεται ο τομέας της *μηχανικής μάθησης* (machine learning), ο οποίος εξετάζει πως μπορούν οι μηχανές να «μάθουν».

Παράλληλα, με τη ραγδαία αύξηση του αριθμού και του μεγέθους των βάσεων δεδομένων, εμφανίζεται η ανάγκη για εργαλεία και τεχνικές επεξεργασίας δεδομένων, ανακάλυψης γνώσης από αυτά και ερμηνείας των αποτελεσμάτων. Αυτά αποτελούν αντικείμενο της *ανακάλυψης γνώσης σε βάσεις δεδομένων* (Knowledge Discovery in Databases - KDD). Στα επόμενα κεφάλαια, θα αναφερθούμε με περισσότερη λεπτομέρεια στη μηχανική μάθηση και την ανακάλυψη γνώσης από βάσεις δεδομένων.

2.1 Μηχανική Μάθηση

Η *μηχανική μάθηση* (machine learning) αποτελεί ερευνητικό τομέα της Τεχνητής Νοημοσύνης. Η ικανότητα μάθησης είναι ένα βασικό στοιχείο της νοημοσύνης. Αυτός είναι ο λόγος που σε πολλές περιπτώσεις αυτές οι δύο έννοιες ταυτίζονται λανθασμένα. Κύριο χαρακτηριστικό της μηχανικής μάθησης είναι ότι δίνει στους υπολογιστές τη δυνατότητα να μάθουν χωρίς να έχουν προγραμματιστεί ρητά (Arthur Samuel, 1959).

Πρόκειται για μία μέθοδο ανάλυσης δεδομένων η οποία αυτοματοποιεί τη δημιουργία *μοντέλων* (models) ή *προτύπων* (patterns). Μοντέλα είναι οι αφαιρετικές, απλοποιημένες περιγραφές του περιβάλλοντος, είτε αυτό αναφέρεται σε ένα σύστημα,

είτε σε κάποια διαδικασία. Πρότυπα ονομάζονται οι νέες δομές που προκύπτουν από τη συσχέτιση και την οργάνωση εμπειριών [Βλαχάβας κ.α., 2006]. Για τη δημιουργία, λοιπόν, μοντέλων και προτύπων, παρέχονται σύνολα δεδομένων στο υπολογιστικό σύστημα, το οποίο με τη χρήση αλγορίθμων της μηχανικής μάθησης, ανακαλύπτει κρυμμένη γνώση, μη γνωρίζοντας εξ αρχής για τι να αναζητήσει. Βασισμένο σε αυτή τη γνώση που έχει εξάγει, το σύστημα μπορεί επιπλέον να κάνει προβλέψεις για άγνωστα δεδομένα.

Ανάλογα με τη φύση του μαθησιακού προβλήματος του συστήματος χρησιμοποιούνται τεχνικές που ανήκουν σε τρεις κύριες περιπτώσεις:

- **Επιβλεπόμενη μάθηση** (Supervised Learning)
- **Μη επιβλεπόμενη μάθηση** (Unsupervised Learning)
- **Ενισχυτική μάθηση** (Reinforcement Learning)

2.1.1. Επιβλεπόμενη Μάθηση

Αυτή η μορφή μάθησης ονομάζεται και *μάθηση με παραδείγματα* (learning from examples). Το σύστημα τροφοδοτείται με ένα σύνολο δεδομένων εκπαίδευσης, που αποτελείται από πιθανές εισόδους και τις αντίστοιχες εξόδους τους. Βάσει των παραδειγμάτων αυτών, το σύστημα καλείται να μάθει επαγωγικά μία συνάρτηση, η οποία εκφράζει το μοντέλο που περιγράφει τα δεδομένα. Η συνάρτηση αυτή είναι γνωστή ως *συνάρτηση - στόχος* (target function). Ως πεδίο ορισμού της έχει το σύνολο των πιθανών τιμών εισόδου του συστήματος. Το σύνολο αυτό ονομάζεται *σύνολο των περιπτώσεων*. Τα *δεδομένα εκπαίδευσης* - ή *παραδείγματα*, όπως αναφέρθηκε παραπάνω - οι περιπτώσεις, δηλαδή, για τις οποίες γνωρίζουμε την τιμή της εξόδου, αποτελούν υποσύνολο του συνόλου των περιπτώσεων.

Η εύρεση της συνάρτησης - στόχου στηρίζεται στην *υπόθεση της επαγωγικής μάθησης*, σύμφωνα με την οποία, κάθε υπόθεση h που έχει βρεθεί να προσεγγίζει καλά τη συνάρτηση στόχο για ένα αρκετά μεγάλο σύνολο παραδειγμάτων, θα προσεγγίζει το ίδιο καλά τη συνάρτηση - στόχο και για περιπτώσεις που δεν έχει εξετάσει.

Ως υπόθεση, ορίζεται κάθε εναλλακτική συνάρτηση που επιχειρεί να προσεγγίσει τη συνάρτηση - στόχο. Υπολογίζονται περισσότερες από μία υποθέσεις, έτσι ώστε να

βρεθεί η καλύτερη δυνατή προσέγγιση της συνάρτησης. Καλή προσέγγιση θεωρείται εκείνη που γενικεύεται σωστά, που μπορεί, δηλαδή να προβλέπει με σωστό τρόπο άγνωστα δεδομένα που δεν έχει ήδη εξετάσει. Μπορεί, φυσικά, να υπάρχουν περισσότερες από μία συνεπείς στη λειτουργία τους υποθέσεις, πράγμα που σημαίνει πως πρέπει να επιλεγεί κάποια από αυτές. Μία λογική επιλογή αποτελεί η απλούστερη δυνατή υπόθεση που περιγράφει ικανοποιητικά τα δεδομένα. (Ξυράφι του Ockham - Ockham's Razor) [Russell et al., 2005].

Η επιβλεπόμενη μάθηση χρησιμοποιείται σε προβλήματα *ταξινόμησης* (classification) και *παλινδρόμησης* (regression). Στα προβλήματα ταξινόμησης το σύστημα καλείται να δημιουργήσει ένα μοντέλο το οποίο θα προβλέπει σε ποια από κάποιες διακριτές κλάσεις ανήκει μία άγνωστη είσοδος. Ένα παράδειγμα αυτής της κατηγορίας είναι η ταξινόμηση ατόμων σύμφωνα με την ομάδα αίματός τους. Στα προβλήματα παλινδρόμησης το μοντέλο κάνει προβλέψεις συνεχών αριθμητικών τιμών. Ένα τέτοιο πρόβλημα θα ήταν, για παράδειγμα, η πρόβλεψη του βάρους ενός ατόμου γνωρίζοντας το ύψος του.

Κάποιες από τις τεχνικές της μηχανικής μάθησης με επίβλεψη είναι οι παρακάτω [Βλαχάβας κ.α., 2006]:

- Μάθηση εννοιών (Concept Learning)
- Δένδρα απόφασης/ ταξινόμησης (Decision/ Classification Trees)
- Μάθηση κανόνων (Rule Learning)
- Νευρωνικά δίκτυα (Neural Networks)
- Μάθηση κατά Bayes
- Μάθηση με βάση τις περιπτώσεις (Instance Based Learning)
- Μηχανές διανυσμάτων Υποστήριξης (Support Vector Machines) κ.ά.

2.1.2. Μη επιβλεπόμενη Μάθηση

Ονομάζεται και *μάθηση από παρατήρηση* (learning from observation). Σε αυτήν την περίπτωση, δεν παρέχονται συγκεκριμένες τιμές εξόδου στο σύστημα, αλλά το σύστημα καλείται να εξερευνήσει τα δεδομένα και να δημιουργήσει πρότυπα τα οποία προκύπτουν

από τις συσχετίσεις ή τις ομάδες που υπάρχουν στα δεδομένα. Χρησιμοποιείται σε προβλήματα *ανάλυσης συσχετισμών* (association analysis) και *ομαδοποίησης* (clustering).

Η ανάλυση συσχετισμών (ή εξόρυξη κανόνων συσχέτισης - association rule mining) είναι στενά συνδεδεμένη με τις βάσεις δεδομένων, εφόσον πρόκειται για ανακάλυψη συσχετίσεων μεταξύ αντικειμένων μίας βάσης δεδομένων. Οι κανόνες συσχέτισης προτάθηκαν από τον Rakesh Agrawal κ.ά. το 1993 και είχαν ως στόχο αρχικά την εύρεση συσχετίσεων μεταξύ αντικειμένων σε μεγάλες βάσεις δεδομένων με συναλλαγές. Αποτελούσε, δηλαδή, τεχνική ανάλυσης καλαθιού αγορών (market basket analysis). Σύμφωνα με αυτήν την τεχνική, κάθε κανόνας έχει τη μορφή $\{I1, \ldots, In\} \rightarrow Y$και αντιστοιχεί στην ιδέα ότι σύμφωνα με τις συναλλαγές που υπάρχουν στη βάση δεδομένων, αν υπάρχουν σε κάποιο καλάθι τα αντικείμενα $\{I1, \ldots, In\}$, τότε είναι πιθανό να υπάρχει και το αντικείμενο Yστη συναλλαγή. Στο δεξί μέλος του κανόνα μπορούν φυσικά να βρίσκονται και παραπάνω από ένα αντικείμενα. Ως μετρικές για την εκτίμηση της ποιότητας των κανόνων συσχέτισης ορίζονται οι ποσότητες της *υποστήριξης* (support) και της *εμπιστοσύνης* (confidence).

Η υποστήριξη εκφράζει το κατά πόσο είναι πιθανό να βρεθούν όλα τα αντικείμενα του κανόνα, τόσο του αριστερού, όσο και του δεξιού μέλους του, στην ίδια συναλλαγή στη βάση δεδομένων. Η τιμή της, λοιπόν, είναι ίση με το πλήθος των εγγραφών που περιέχουν τα αντικείμενα αυτά, προς το συνολικό αριθμό των εγγραφών.

Η εμπιστοσύνη αντιστοιχεί στην πιθανότητα να περιέχεται τοYσε ένα καλάθι που περιέχει τα αντικείμενα$\{I1, \ldots, In\}$. Πρόκειται, δηλαδή, για το πλήθος των εγγραφών που περιέχουν όλα τα αντικείμενα του κανόνα, προς το πλήθος των εγγραφών που περιέχουν μόνο τα αντικείμενα του αριστερού μέλους του κανόνα συσχέτισης.

Για την επιλογή του υποσυνόλου των κανόνων που έχουν αξία, ορίζονται τιμές κατωφλίου για την υποστήριξη και την εμπιστοσύνη των κανόνων συσχέτισης. Τελικά επιλέγονται οι κανόνες οι τιμές των οποίων ξεπερνούν αυτά τα κατώφλια. Κλασικός αλγόριθμος εύρεσης κανόνων συσχέτισης είναι ο αλγόριθμος *Apriori* [Agrawal et al., 1994], σύμφωνα με τον οποίο προκύπτουν κανόνες συσχέτισης από συχνά σύνολα υποσύνολα αντικειμένων (bottom-up προσέγγιση).

Στα προβλήματα ομαδοποίησης επιζητείται ο διαχωρισμός των δοθέντων δεδομένων σε ομάδες (clusters). Οι ομάδες μπορούν να είναι περισσότερες από δύο και ο διαχωρισμός πρέπει να γίνεται με τρόπο τέτοιο, ώστε σε μία ομάδα να ανήκουν τα στοιχεία εκείνα που μοιάζουν περισσότερο μεταξύ τους, και σε διαφορετικές ομάδες τα στοιχεία που διαφέρουν περισσότερο μεταξύ τους. Πιο συγκεκριμένα, οι ομάδες μπορούν να οριστούν ως σύνολα σημείων στο χώρο, όπου η απόσταση μεταξύ δύο σημείων μίας ομάδας είναι μικρότερη από την απόσταση δύο οποιονδήποτε στοιχείων της ομάδας και οποιουδήποτε σημείου εκτός της ομάδας. Η απόσταση αυτή μπορεί να μετράται με διάφορες μετρικές και να αντιστοιχεί σε ομοιότητα σύμφωνα με διάφορα κριτήρια. Αντίστοιχα και η ποιότητα των αποτελεσμάτων αξιολογείται σύμφωνα με κάποιο κριτήριο ομαδοποίησης. Οι αλγόριθμοι ομαδοποίησης κατατάσσονται σε τρεις γενικές κατηγορίες:

Αλγόριθμοι βασισμένοι σε **διαχωρισμούς** (Partition based)

Δοθέντος ενός συγκεκριμένου αριθμού ομάδων, οι αλγόριθμοι αυτής της κατηγορίας προσπαθούν να διαχωρίσουν όσο δυνατόν καλύτερα τα δεδομένα στις ομάδες αυτές. Γνωστότερος εκπρόσωπος αυτής της κατηγορίας αλγορίθμων είναι ο *αλγόριθμος Κ-μέσων* (K-means algorithm), σύμφωνα με τον οποίο επιλέγονται Κ κέντρα και ανατίθενται σε αυτά τα κοντινότερα - βάσει κάποιου κριτηρίου - σημεία. Τα κέντρα ανανεώνονται και η διαδικασία επαναλαμβάνεται μέχρις ότου είτε να μην αλλάζει ο διαχωρισμός των σημείων, είτε να έχει εκτελεστεί συγκεκριμένος αριθμός επαναλήψεων που εμείς έχουμε ορίσει.

Ιεραρχικοί αλγόριθμοι (Hierarchical algorithms)

Οι ιεραρχικοί αλγόριθμοι βασίζονται στην εύρεση της δομής των ομάδων, χωρίς τη γνώση για το πλήθος τους. Οι αλγόριθμοι χωρίζονται στους *αλγόριθμους συγχώνευσης* (agglomerative ή bottom-up) και στους *αλγόριθμους διαίρεσης* (divisive ή top-down). Οι πρώτοι ξεκινούν θεωρώντας ότι κάθε σημείο αποτελεί μία ομάδα και συνενώνουν τις ομάδες αυτές σύμφωνα με την ομοιότητά τους, ενώ οι δεύτεροι ξεκινούν με όλα τα σημεία να ανήκουν σε μία ομάδα και σταδιακά τη διαιρούν σε υπο - ομάδες. Συχνά οι ιεραρχίες

που προκύπτουν απεικονίζονται με τη χρήση δενδρογραμμάτων. Οι ιεραρχικοί αλγόριθμοι στηρίζονται σε έναν πίνακα αποστάσεων γι' αυτό και απαιτούν πολύ χρόνο και χώρο στον υπολογιστή, καθιστώντας τη χρήση τους σε μεγάλα σύνολα δεδομένων ασύμφορη. Εκτός αυτού, εφόσον οι αλγόριθμοι αυτοί προσφέρουν λύση για διάφορους αριθμούς ομάδων, επαφίεται στον ερευνητή η επιλογή μίας από αυτές.

Πιθανοκρατικοί αλγόριθμοι (Probabilistic algorithms)

Αυτή η κατηγορία αλγορίθμων βασίζεται σε μοντέλα πιθανοτήτων, όπως η θεωρία του Bayes. Πρόκειται για αλγόριθμους οι οποίοι δε βρίσκουν απαραίτητα τη βέλτιστη λύση, μπορούν όμως να βρουν μία ικανοποιητική λύση σε εύλογο σχετικά χρονικό διάστημα.

2.1.3. Ενισχυτική Μάθηση

Στην ενισχυτική μάθηση (reinforcement learning) το σύστημα - στο οποίο αναφερόμαστε και ως πράκτορα - έχει ως στόχο να μάθει μία βέλτιστη συμπεριφορά από την άμεση αλληλεπίδρασή του με το περιβάλλον. Ενεργεί, λοιπόν, σύμφωνα με τις γνώσεις του για το περιβάλλον του και την πολιτική του και δέχεται ανάδραση, την *ενίσχυση* (reinforcement) ή *ανταμοιβή* (reward). Ανάλογα με το είδος αυτής της ανάδρασης, το σύστημα μαθαίνει αν έχει ακολουθήσει μία ικανοποιητική κατάσταση ενεργειών ή όχι και αντίστοιχα ενισχύεται ή αποδυναμώνεται η τάση του να παράγει αυτή την ενέργεια. Εμφανής είναι η αναλογία της λειτουργίας του συστήματος με αντίστοιχες τεχνικές μάθησης με επιβράβευση και τιμωρία που χρησιμοποιούνται για τη μάθηση στα έμβια όντα [Παρτάλας, 2009].

Οι τεχνικές ενισχυτικής μάθησης χρησιμοποιούνται σε προβλήματα *σχεδιασμού ενεργειών* (planning problems), όπου στόχος του συστήματος είναι να βρει μία ακολουθία ενεργειών ώστε να οδηγηθεί σε μία ήδη πλήρως γνωστή τελική κατάσταση - στόχο. Τέτοια προβλήματα απαντώνται στον έλεγχο κίνησης ρομπότ, στη μάθηση επιτραπέζιων παιχνιδιών, κ.ο.κ. Οι τεχνικές της ενισχυτικής μάθησης είναι αρκετά ελκυστικές, καθώς δεν

απαιτείται μεγάλη προγραμματιστική προσπάθεια. Αντιθέτως, είναι αρκετό να οριστεί η ανταμοιβή του συστήματος από τους σχεδιαστές, και το σύστημα στη συνέχεια μαθαίνει από την αλληλεπίδραση με το περιβάλλον, χωρίς να χρειάζεται εκ νέου προγραμματισμός του [Russell et al., 2005].

2.2 Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων

Η ανακάλυψη γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases - KDD) σχετίζεται με τη μηχανική μάθηση, τη συλλογή δεδομένων, τον σχεδιασμό βάσεων δεδομένων, την περιγραφή καταχωρήσεων με την καταλληλότερη αναπαράσταση και την ποιότητα δεδομένων. Ένας ορισμός της θέτει ότι πρόκειται για μία "ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα." [Frawley et al. 1991]. Αναλυτικότερα, η ανακάλυψη γνώσης αναφέρεται στην εξαγωγή χρήσιμης πληροφορίας από ένα σύνολο δεδομένων, το οποίο συνήθως είναι μεγάλο σε μέγεθος και αποτελείται από δεδομένα συγκεντρωμένα για διάφορους σκοπούς. Επειδή, λοιπόν, ο όγκος των δεδομένων είναι μεγάλος και η δομή τους εξυπηρετεί άλλους σκοπούς, είναι απαραίτητος ο μετασχηματισμός της πληροφορίας σε μία μορφή κατάλληλη για περαιτέρω χρήση.

Η ανακάλυψη προτύπων και η δημιουργία μοντέλων από μία βάση δεδομένων που δεν έχει δημιουργηθεί συγκεκριμένα για την εφαρμογή μηχανικής μάθησης στα δεδομένα της μπορεί να αποτελέσει πρόκληση. Αυτό συμβαίνει γιατί είναι υπολογιστικά ακριβή η δημιουργία όλων των δυνατών περιγραφών, οπότε και είναι απαραίτητη η εύρεση της καλύτερης δυνατής περιγραφής δεδομένων.

Οδηγούμαστε από τα δεδομένα μέχρι τη γνώση μέσα από μία διαδικασία που περιέχει την επιλογή (selection), την προ-επεξεργασία (preprocessing), τον μετασχηματισμό (transformation), την εξόρυξη (data mining) και την ερμηνεία/ αξιολόγηση (interpretation/ evaluation) (Εικόνα 2.1).

Πριν ξεκινήσει αυτή η διαδικασία, είναι απαραίτητο να γίνει κατανοητό το πεδίο εφαρμογής, καθώς και οι αναμενόμενοι στόχοι των τελικών χρηστών. Μέσα από την κατανόηση του πεδίου γίνεται ξεκάθαρο τι είδους μετασχηματισμοί πρέπει να

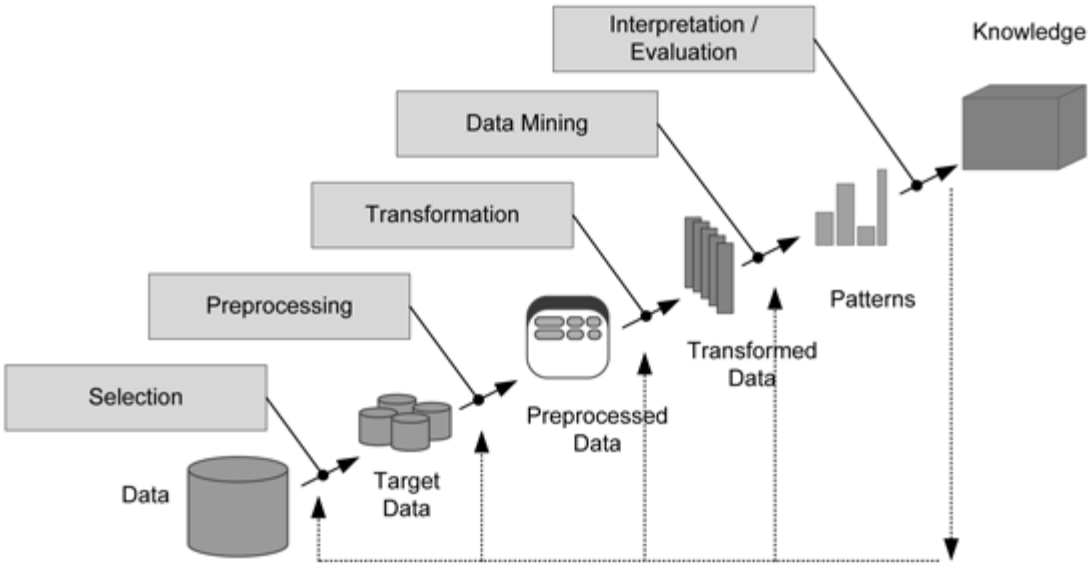
εφαρμοστούν, καθώς και ποιοι αλγόριθμοι και ποιες αναπαραστάσεις μπορούν να χρησιμοποιηθούν. Σε αυτό το βήμα προετοιμασίας είναι σημαντική η συμβολή ενός ειδικού σχετικού με το είδος των δεδομένων της βάσης με την οποία εργαζόμαστε.

2.2.1. Διαδικασία Ανακάλυψης Γνώσης

Στη συνέχεια, αναλύεται κάθε στάδιο της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων ξεχωριστά.

Επιλογή

Όπως αναφέρθηκε παραπάνω, οι βάσεις δεδομένων από τις οποίες καλούμαστε να εξάγουμε γνώση περιέχουν δεδομένα από ετερογενείς πηγές και πολλές φορές βρίσκονται σε μορφή που δεν διευκολύνει την ανακάλυψη γνώσης. Στο στάδιο αυτό, επιλέγονται από τη βάση τα δεδομένα εκείνα που αρχικά είναι σχετικά με τους στόχους της ανακάλυψης γνώσης, καθώς και τα αντίστοιχα χρήσιμα χαρακτηριστικά των εγγραφών. Επίσης σε αυτό το στάδιο οργανώνονται τα επιλεγμένα δεδομένα σε απλούστερες δομές από τις αρχικές, ώστε να γίνει το έργο της ανακάλυψης ευκολότερο.



Εικόνα 2.1: Τα στάδια της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων

Προ-επεξεργασία

Ονομάζεται και στάδιο *καθαρισμού δεδομένων* (data cleaning). Κάποιες φορές τίθενται ζητήματα σχετικά με την αξιοπιστία των δεδομένων. Αυτό συμβαίνει επειδή μπορεί να υπάρχουν ελλιπή ή λανθασμένα δεδομένα που προκύπτουν από ανθρώπινα λάθη, σφάλματα σε μετρήσεις, κ.ά. Για να είναι τα αποτελέσματα της ανακάλυψης πιο αξιόπιστα, είναι σοφό να χρησιμοποιηθούν εξαρχής αξιόπιστα δεδομένα. Για το λόγο αυτό, στο στάδιο της προ-επεξεργασίας γίνονται διορθώσεις στα δεδομένα. Αυτό μπορεί να σημαίνει ότι, για τη βελτίωση της ποιότητάς τους, μπορεί είτε να αφαιρεθούν λανθασμένες και ακραίες τιμές (outliers), είτε να προβλεφθούν τιμές για ελλιπή πεδία. Με αυτόν τον τρόπο γίνεται ο καθαρισμός των δεδομένων.

Μετασχηματισμός

Αφού έχει γίνει ο καθαρισμός των δεδομένων, ακολουθεί ο μετασχηματισμός τους, διαδικασία που διευκολύνει την ανακάλυψη γνώσης. Κατά το μετασχηματισμό ουσιαστικά αλλάζει η μορφή των δεδομένων, πράγμα που μπορεί να γίνει με διαφορετικούς τρόπους.

Μία τέτοια αλλαγή αποτελεί η *μείωση διαστάσεων* (dimensionality reduction). Στη μείωση διαστάσεων, γίνεται επιλογή χαρακτηριστικών (feature selection), μειώνεται, δηλαδή, ο αριθμός των υπό εξέταση χαρακτηριστικών.

Επίσης, μπορεί να γίνει μετασχηματισμός των χαρακτηριστικών, όπως για παράδειγμα η διακριτοποίησή τους, κατά την οποία συνεχόμενες αριθμητικές τιμές μετατρέπονται σε διακριτές τιμές. Αυτό μπορεί να είναι απαραίτητο για αλγόριθμους που δεν μπορούν να υλοποιηθούν ή δε λειτουργούν ικανοποιητικά καλά για συνεχείς τιμές. Ακόμη σε αυτό το στάδιο έχει νόημα να ενοποιηθούν πεδία που αναφέρονται στην ίδια λογική υπόσταση αλλά έχουν καταγραφεί με διαφορετικό τρόπο ή όνομα πεδίου.

Οι αλλαγές αυτές μπορεί να αποδειχθούν πολύ σημαντικές για τη διαδικασία ανακάλυψης γνώσης. Για παράδειγμα, η επιλογή χαρακτηριστικών είναι ιδιαίτερα χρήσιμη γιατί κάποια χαρακτηριστικά μπορεί να είναι άσχετα ή και ακατάλληλα για τη διαδικασία, με αποτέλεσμα να δυσχεραίνουν την ανακάλυψη γνώσης. Παρ’ όλο που πολλοί αλγόριθμοι μπορούν να ελέγξουν την σημαντικότητα ενός κριτηρίου - όπως για παράδειγμα, στα δένδρα απόφασης όπου ο αλγόριθμος μπορεί να επιλέξει το επόμενο σημαντικότερο κριτήριο για διαχωρισμό - η απόδοσή τους τόσο σε ταχύτητα όσο και σε ποιότητα εξαγόμενης γνώσης επηρεάζεται αρνητικά από άσχετα πεδία - χαρακτηριστικά.

Εξόρυξη

Στο στάδιο της εξόρυξης επιλέγεται και εφαρμόζεται ένας συγκεκριμένος αλγόριθμος για εξαγωγή προτύπων από τα - μετασχηματισμένα πλέον - δεδομένα. Ανάλογα με το είδος της γνώσης που αναζητείται, προσδιορίζεται η κατηγορία του αλγόριθμου που θα χρησιμοποιηθεί για την εξόρυξη. Τα αποτελέσματα της εξόρυξης μπορεί να είναι είτε *μοντέλα πρόβλεψης* (predictive models) είτε *πρότυπα πληροφόρησης* (informative patterns). Ανάλογα με αυτό το είδος γνώσης των αποτελεσμάτων που επιστρέφουν, οι εργασίες εξόρυξης δεδομένων χωρίζονται σε δύο βασικές κατηγορίες [Tan et al., 2016]:

- **Προγνωστικές εργασίες** (Predictive tasks)
- **Περιγραφικές εργασίες** (Descriptive tasks)

Οι προγνωστικές εργασίες έχουν ως στόχο την πρόβλεψη της τιμής ενός χαρακτηριστικού, γνωστού και ως στόχος (target) βάσει των τιμών άλλων χαρακτηριστικών, που ονομάζονται επεξηγηματικές μεταβλητές (explanatory variables). Σε αυτήν την κατηγορία ανήκουν εργασίες *κατηγοριοποίησης* (classification), στις οποίες γίνεται πρόβλεψη τιμών για διακριτές τιμές στόχων και *παλινδρόμησης* (regression), στις οποίες οι μεταβλητές στόχοι είναι συνεχείς.

Οι περιγραφικές εργασίες στοχεύουν στην εξαγωγή υποδειγμάτων που περιγράφουν τις σχέσεις που υπάρχουν στα δεδομένα. Τα υποδείγματα αυτά μπορεί να είναι συσχετίσεις, συστάδες, ανωμαλίες κ.ά. Εδώ ανήκουν η *ανάλυση συσχέτισης* (association analysis). η *ανάλυση συστάδων* (cluster analysis), και η *ανίχνευση ανωμαλιών* (anomaly detection), που αντιστοιχούν στα παραπάνω είδη υποδειγμάτων.

Πολλές φορές ο όρος εξόρυξη χρησιμοποιείται για να περιγράψει όλη τη διαδικασία της ανακάλυψης γνώσης σε βάσεις δεδομένων, δε θα έπρεπε όμως να δημιουργείται αυτή η σύγχυση. Η εξόρυξη αποτελεί αναπόσπαστο κομμάτι της KDD, όμως η διαδικασία αυτή περιέχει όλα τα στάδια από τα ακατέργαστα δεδομένα μέχρι τη χρήσιμη πληροφορία.

Ερμηνεία/ Αξιολόγηση

Έπειτα από το βήμα της εξόρυξης δεδομένων είναι σημαντικό να αναλυθούν τα αποτελέσματα της διαδικασίας. Μετά από την ανάλυσή τους, θα χρησιμοποιηθούν μόνο αυτά που θεωρούνται χρήσιμα στα συστήματα στα οποία και θα ενσωματωθούν. Αυτή η εκ των υστέρων επεξεργασία περιέχει συχνά την οπτικοποίηση (visualization), με τη βοήθεια της οποίας γίνεται πιο εύκολα κατανοητή η γνώση που έχει εξαχθεί αλλά και δίνεται η δυνατότητα να διερευνηθούν τα αποτελέσματα της διαδικασίας από διάφορες οπτικές γωνίες. Σε αυτή τη φάση, αν η εξαγόμενη γνώση πρόκειται να προστεθεί σε κάποια ήδη υπάρχουσα βάση γνώσης, θα πρέπει να γίνει διερεύνηση και επίλυση συγκρούσεων.

Πολλές φορές χρειάζεται να επαναληφθούν τμήματα ή ακόμη και ολόκληρη η διαδικασία της ανακάλυψης γνώσης. Επανάληψη της διαδικασίας μπορεί να χρειαστεί επειδή κάποιες πτυχές του προβλήματος γίνονται γνωστές αργότερα στη διαδικασία, ή επειδή δεν έχουν προκύψει τα επιθυμητά αποτελέσματα, είτε ακόμη για να λάβει χώρα η διαδικασία με διαφορετικές παραμέτρους ή τεχνικές.

Η ανακάλυψη γνώσης από δεδομένα βρίσκει εφαρμογή σε πολλούς τομείς, όπως, επί παραδείγματι, τον παγκόσμιο ιστό, τις επιχειρήσεις, τον τομέα των επενδύσεων και την ιατρική. Εμφανές κοινό χαρακτηριστικό των τομέων αυτών αποτελεί το μέγεθος των δεδομένων που παρέχουν και η ανάγκη για εκμαίευση χρήσιμης και εφαρμόσιμης πληροφορίας από αυτά.

2.2.2. Προβλήματα στην ανακάλυψη γνώσης

Κατά τη διάρκεια της ανακάλυψης γνώσης προκύπτουν προβλήματα που σχετίζονται με τα δεδομένα. Είτε πρόκειται για δεδομένα ακατάλληλα για τις μεθόδους που χρησιμοποιούνται, είτε η αναπαράσταση των δεδομένων είναι ακατάλληλη, η διαδικασία της ανακάλυψης γνώσης γίνεται δυσκολότερη. Παρακάτω αναλύονται τα πιο σημαντικά από τα προβλήματα αυτά [Βλαχάβας κ.ά., 2006].

Ακατάλληλα δεδομένα

Όπως αναφέρθηκε παραπάνω, συχνά, οι βάσεις δεδομένων που χρησιμοποιούνται για την ανακάλυψη γνώσης δεν έχουν δημιουργηθεί συγκεκριμένα για το σκοπό αυτό, με αποτέλεσμα κάποιες φορές τα δεδομένα που περιέχουν να είναι ακατάλληλα για την εφαρμογή της. Για παράδειγμα, μπορεί να μην έχουν συγκεντρωθεί τιμές για κάποιο πεδίο που θα διευκόλυνε ιδιαίτερα τη διαδικασία ανακάλυψης, είτε γιατί κατά τη δημιουργία της βάσης δε θεωρήθηκε απαραίτητο, είτε ακόμη επειδή δεν υπάρχει η δυνατότητα να συγκεντρωθούν οι τιμές αυτές.

Ελλιπή δεδομένα (Missing data)

Αυτό το πρόβλημα αναφέρεται σε τιμές πεδίων που δεν έχουν συμπληρωθεί. Είτε πρόκειται για κάποιο αναπάντητο ερώτημα σε φόρμα, είτε για κάποια τιμή που διαγράφηκε ή δε μετρήθηκε εξαρχής, οι ελλιπείς τιμές οδηγούν στη δημιουργία ανακριβών μοντέλων, αποτελώντας με αυτόν τον τρόπο μία δυσκολία στη διαδικασία ανακάλυψης γνώσης.

Θόρυβος (Noise)

Οι τιμές των πεδίων συχνά περιέχουν λάθη τα οποία προκύπτουν από λανθασμένες μετρήσεις ή υποκειμενικές κρίσεις. Τα λάθη αυτά ονομάζονται θόρυβος. Ο θόρυβος, λοιπόν, μπορεί να οδηγήσει σε σημαντική μείωση στην ακρίβεια της γνώσης, καθιστώντας επιτακτική την ανάγκη αντιμετώπισής του.

Αραιά δεδομένα (sparse data)

Οι εγγραφές που υπάρχουν στη βάση δεδομένων μπορεί να αντιστοιχούν σε ένα μικρό μόνο τμήμα των πιθανών εναλλακτικών των πεδίων τους. Αυτό σημαίνει πως τα διαθέσιμα δεδομένα δεν παρουσιάζουν επαρκή ποικιλία, δημιουργώντας προβλήματα σε εφαρμογές όπως ο προσδιορισμός κατηγοριών, όπου έχοντας αραιά δεδομένα είναι ιδιαίτερα δύσκολο να οριστούν τα όρια της εκάστοτε κατηγορίας με ακρίβεια.

Μεγάλο μέγεθος βάσης δεδομένων

Μία βάση δεδομένων μπορεί να περιέχει τεράστια ποσότητα δεδομένων. Είτε το πλήθος των εγγραφών είναι μεγάλο, είτε το πλήθος των πεδίων ανά εγγραφή, το μέγεθος των δεδομένων έχει ως συνέπεια να είναι απαγορευτικά χρονοβόρα η εκτέλεση των επιλεγμένων αλγορίθμων εξόρυξης και ο έλεγχος της ποιότητας της προκύπτουσας γνώσης. Είναι απαραίτητη, λοιπόν, η μείωση του μεγέθους των δεδομένων με κάποια πρακτική, όπως για παράδειγμα με την επιλογή δειγμάτων.

Δείγματα

Η επιλογή δειγμάτων αποτελεί μία δημοφιλή επιλογή για την αντιμετώπιση του μεγάλου αριθμού εγγραφών των βάσεων δεδομένων. Απαιτείται προσοχή και χρήση στατιστικών μεθόδων για τη σωστή δειγματοληψία, καθώς είναι σημαντικό να επιλέγονται τα δείγματα με τρόπο τέτοιο, ώστε η αρχική βάση δεδομένων να αντιπροσωπεύεται σε ικανοποιητικό βαθμό. Σε διαφορετική περίπτωση, η γνώση που θα προκύψει για τις σχέσεις μεταξύ των δεδομένων θα είναι λανθασμένη.

Πρόσφατα δεδομένα

Το περιβάλλον των βάσεων δεδομένων είναι δυναμικό, πράγμα που σημαίνει πως συχνά γίνονται αλλαγές στο περιεχόμενό τους. Τέτοιες αλλαγές είναι η προσθήκη ή αφαίρεση δεδομένων, καθώς και η τροποποίησή τους. Επομένως, προκύπτει το ερώτημα αν τα ανανεωμένα δεδομένα συμβαδίζουν με οποιαδήποτε γνώση προέκυψε πριν από την ενημέρωση της βάσης. Μπορούμε, κατ' επέκταση, να επαναλάβουμε την ανακάλυψη γνώσης με τα νέα δεδομένα, ή να μετατρέψουμε κατάλληλα την ήδη υπάρχουσα γνώση, ώστε να είμαστε βέβαιοι πως δεν υπάρχουν ασυνέπειες.

Κεφάλαιο 3: Βιοπληροφορική

Με το πέρασμα των χρόνων, όπως σε πολλούς άλλους τομείς έτσι και στον τομέα των βιοεπιστημών το μέγεθος των διαθέσιμων δεδομένων συνεχώς αυξάνεται. Με αυτή την έκρηξη δεδομένων -με το μέγεθος τους να φτάνει τα αρκετά petabytes σύμφωνα με το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (EBI) [EBI, 2015]- είναι απαραίτητη η χρήση σύγχρονων τεχνικών τόσο για τη διαχείριση τους, όσο και για την εκμαίευση χρήσιμης πληροφορίας από αυτά. Ακόμη σημαντικότερο φαίνεται αυτό το έργο, αφού η πληροφορία αυτή μπορεί να μας φέρει πιο κοντά σε σημαντικά ευρήματα στον τομέα της ιατρικής και της δημιουργίας φαρμάκων -και όχι μόνο. Αυτή την ανάγκη διαχείρισης και εκμετάλλευσης των βιολογικών δεδομένων προσπαθεί να καλύψει ο τομέας της βιοπληροφορικής.

Η βιοπληροφορική αναφέρεται στην εφαρμογή τεχνολογιών της πληροφορικής για την κατανόηση και διαχείριση βιολογικών δεδομένων. Πιο αναλυτικά, πρόκειται για ένα διατμηματικό επιστημονικό πεδίο που προκύπτει από το συνδυασμό επιστημών σχετικών με την πληροφορία - όπως η πληροφορική, τα μαθηματικά και η στατιστική- και των βιοεπιστημών - όπως η βιολογία, η χημεία και η ιατρική. Ο όρος *βιοπληροφορική* (bioinformatics) εμφανίστηκε στη δεκαετία του 1990, αν και η ανάλυση βιολογικών συστημάτων με υπολογιστικές μεθόδους είχε ξεκινήσει αρκετά νωρίτερα.

Κύριος στόχος του πεδίου αυτού είναι η βαθύτερη κατανόηση για τις βιολογικές διαδικασίες, η οποία προκύπτει από τη συλλογή, αποθήκευση και ανάλυση διάφορων τύπων βιολογικών δεδομένων, όπως αλληλουχιών νουκλεοτιδίων ή αμινοξέων, πρωτεϊνικών δομών κ.λπ., ώστε να προκύψει χρήσιμη βιολογική γνώση. Σε αυτό το πλαίσιο ανήκει και η κατάλληλη οργάνωση των δεδομένων, καθώς και η ανάπτυξη αλγορίθμων και εργαλείων που διευκολύνουν την πρόσβαση και διαχείριση δεδομένων σε διάφορες μορφές. Κάποιες από τις τεχνολογίες της πληροφορικής που χρησιμοποιούνται στη βιοπληροφορική είναι η αναγνώριση προτύπων, η οπτικοποίηση, οι τεχνικές μηχανικής μάθησης και εξόρυξης δεδομένων, κ.ά.

Πολλές φορές οι όροι βιοπληροφορική και *υπολογιστική βιολογία* (computational biology) χρησιμοποιούνται χωρίς διάκριση αν και ο πρώτος όρος αναφέρεται περισσότερο

σε πρακτικές εφαρμογές και την διαχείριση μεγάλου όγκου δεδομένων, ενώ η δεύτερη περισσότερο στη θεωρητική σκοπιά και την ανάπτυξη αλγόριθμων.

Επίσης δε θα πρέπει η βιοπληροφορική να συγχέεται με τη βιολογική υπολογιστική (biological computation), όπου χρησιμοποιείται η βιολογία και η βιολογική μηχανική (bioengineering) για τη δημιουργία βιολογικών υπολογιστών.

3.1 Ερευνητικοί Τομείς της Βιοπληροφορικής

Στο γενικότερο πεδίο της βιοπληροφορικής ανήκουν οι παρακάτω τομείς [Τζανής, 2011].

3.1.1. Ανάλυση Βιολογικών Αλληλουχιών (Sequence analysis)

Οι βιολογικές αλληλουχίες αποτελούν ένα μεγάλο τμήμα των δεδομένων με τα οποία ασχολούνται οι επιστήμονες της βιοπληροφορικής. Μέχρι σήμερα έχει καταγραφεί το γονιδίωμα εκατοντάδων οργανισμών. Τα δεδομένα αυτά κρύβουν ιδιαίτερα χρήσιμη πληροφορία για τον εκάστοτε οργανισμό.

Μία από τις εφαρμογές που στοχεύουν στην εκμαίευση της πληροφορίας αυτής είναι η *πρόβλεψη γονιδίων* (Gene prediction). Στόχος της είναι η εύρεση γονιδίων που έχουν κάποιον ενεργό βιολογικό ρόλο στις αλληλουχίες DNA. Είναι γνωστό ότι το γονιδιώμα των ανώτερων οργανισμών αποτελείται σε μεγάλο βαθμό από γενετικό υλικό, το οποίο δεν κωδικοποιεί γονίδια (Junk DNA) αλλά έπαιξε ρόλο μάλλον κατά την εξέλιξή των οργανισμών. Άλλες εφαρμογές αυτού του τομέα είναι η *αλληλούχιση DNA* (DNA Sequencing), κατά την οποία αναζητάται η αλληλουχία νουκλεοτιδικών βάσεων σε ένα τμήμα DNA και η *στοίχιση αλληλουχιών* (Sequence alignment), που στοχεύει στην εύρεση μίας στοίχισης μεταξύ αλληλουχιών DNA, RNA ή πρωτεϊνικών αλληλουχιών, από την οποία προκύπτουν ομοιότητες ή σχέσεις μεταξύ τους.

3.1.2. Γονιδιακή Έκφραση (Gene expression)

Η γονιδιακή έκφραση αναφέρεται στη διαδικασία κατά την οποία ένα γονίδιο εκφράζεται και οδηγεί στη σύνθεση ενός λειτουργικού γονιδιακού προϊόντος. Οι δύο σημαντικότερες μέθοδοι μέτρησης της γονιδιακής έκφρασης είναι η ανάλυση δεδομένων από *μικροσυστοιχίες γονιδίων* και η *τεχνική SAGE*.

Οι μικροσυστοιχίες είναι μία διάταξη μικροσκοπικών σημείων που αντιπροσωπεύουν μοναδικά γονίδια. Στόχος τους είναι να προσδιορίσουν, για κάποιο τύπο κυττάρου ενός οργανισμού, ποιά γονίδια ενεργοποιούνται υπό συγκεκριμένες συνθήκες και σε συγκεκριμένη χρονική στιγμή.

Η τεχνική SAGE (Serial Analysis of Gene Expression) είναι μέθοδος σύμφωνα με την οποία αναγνωρίζεται πόσα διαφορετικά είδη κλώνων mRNA υπάρχουν μία συγκεκριμένη χρονική στιγμή σε ένα κύτταρο. Πρόκειται για μία ισχυρή μέθοδο που επιτρέπει την ανάλυση του συνόλου της γενετικής έκφρασης σε ένα κύτταρο.

3.1.3. Γενετική και Γενωμική (Genetics and genomics)

Ο τομέας αυτός αναφέρεται στη μελέτη γονιδιωμάτων. Η μελέτη αυτή περιλαμβάνει τον προσδιορισμό του πλήρους γονιδιώματος ενός οργανισμού (χαρτογράφηση γονιδιώματος), τον προσδιορισμό της λειτουργίας των γονιδίων και την αλληλούχιση του DNA.

Οι ερευνητικές περιοχές της γενωμικής συμπεριλαμβάνουν την *συγκριτική γενωμική* (comparative genomics) και τη *λειτουργική γενωμική* (functional genomics). Η συγκριτική γενωμική ασχολείται με τη σύγκριση των γονιδιακών χαρακτηριστικών - όπως η αλληλουχία DNA, η αλληλουχία γονιδίων και οι ρυθμιστικές αλληλουχίες - διαφορετικών οργανισμών που αποσκοπεί στην εύρεση των μεταξύ τους ομοιοτήτων ή εξελικτικών σχέσεων. Στόχος της λειτουργικής γενωμικής είναι η κατανόηση του ρόλου των γονιδίων και του τρόπου με τον οποίο αυτά οδηγούν σε συγκεκριμένη λειτουργία των οργανισμών.

3.1.4. Πρωτεωμική (Proteomics)

Η πρωτεωμική ασχολείται με τη μελέτη της δομής και της λειτουργίας των πρωτεϊνών που προκύπτουν από την έκφραση ενός γονιδιώματος. Ο αριθμός των πρωτεϊνών σε έναν οργανισμό είναι πολύ μεγαλύτερος από τον αριθμό των γονιδίων του. Υπάρχουν, όμως, κοινά χαρακτηριστικά μεταξύ τους βάσει των οποίων μπορούν να ταξινομηθούν σε ομάδες και να διευκολυνθεί με τον τρόπο αυτό η μελέτη τους. Ιδιαίτερο ενδιαφέρον εδώ, δίνεται στην τρισδιάστατη μορφή των πρωτεϊνών, η οποία φαίνεται να παίζει σημαντικό ρόλο στη λειτουργία τους.

3.2 Βάσεις δεδομένων στη Βιοπληροφορική

Οι βιολογικές βάσεις δεδομένων αποτελούν αποθήκες βιολογικής πληροφορίας, και είναι απαραίτητο κομμάτι της έρευνας και των εφαρμογών της βιοπληροφορικής. Στόχος τους είναι η αποθήκευση, διαχείριση και χρήση των δεδομένων, ώστε να διευκολυνθεί η κατανόηση βιολογικών φαινομένων. Μπορεί να είναι περιέχουν διαφόρων ειδών πληροφορίες, όπως, για παράδειγμα, πρωτεϊνικές αλληλουχίες ή αλληλουχίες DNA, μακρομοριακές δομές, κ.ο.κ.

Οι βάσεις δεδομένων μπορούν να κατηγοριοποιηθούν με διάφορους τρόπους [Παπανικολάου κ.ά., 2015]. Για παράδειγμα, ανάλογα με τον τύπο δεδομένων που αποθηκεύουν μπορούν να ταξινομηθούν σε *βάσεις δεδομένων αλληλουχιών* (sequence databases), που περιέχουν αλληλουχίες DNA και πρωτεϊνών ή άλλων πολυμερών, *βάσεις* που περιέχουν *δομές* (structure databases), στις οποίες αποθηκεύονται τρισδιάστατες δομές DNA, RNA και πρωτεϊνών και λειτουργικές βάσεις δεδομένων (functional databases), οι οποίες περιέχουν πληροφορία σχετική με τη λειτουργία των γονιδίων και των γονιδιακών προϊόντων.

Επίσης μπορούν να ταξινομηθούν σε πρωτογενείς και δευτερογενείς βάσεις δεδομένων. Οι *πρωτογενείς βάσεις δεδομένων* (primary databases) περιέχουν εμπειρικά δεδομένα τα οποία έχουν προκύψει από πειράματα. Οι *δευτερογενείς βάσεις δεδομένων* (secondary databases) έχουν ως πηγές τους άλλες βάσεις δεδομένων και περιέχουν τιμές

που έχουν προκύψει από αναλύσεις στα πρωτογενή δεδομένα. Άλλες ταξινομήσεις μπορούν να γίνουν σύμφωνα με τα τεχνικά χαρακτηριστικά και την αρχιτεκτονική τους, αν είναι γενικές ή ειδικές για κάποιον οργανισμό κ.ά. [Μπάγκος, 2015].

Κάποια παραδείγματα πρωτογενών βάσεων βιολογικών δεδομένων είναι η EMBL (Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής - EBI), η GenBank (Εθνικό Κέντρο Βιοτεχνολογικής Πληροφορίας - NCBI) και η DNA Data Bank of Japan (Εθνικό Ινστιτούτο Γενετικής - NIG), που περιέχουν δεδομένα βιολογικών αλληλουχιών, ενώ δευτερογενείς βάσεις είναι η RefSeq, που περιέχει αλληλουχίες DNA, RNA και πρωτεϊνών, η BLOCKS, που περιέχει πρωτεϊνικές αλληλουχίες, κ.ά.

3.3 Ανακάλυψη Γνώσης Στη Βιοπληροφορική

Αναφερθήκαμε σε προηγούμενο υποκεφάλαιο στην ανακάλυψη γνώσης από βάσεις δεδομένων. Συγκεκριμένα, η ανακάλυψη γνώσης στη βιοπληροφορική σχετίζεται με την ανάλυση βιολογικών δεδομένων διαφόρων ειδών.

Μία από τις εφαρμογές της είναι η ανακάλυψη γνώσης από βιολογικές αλληλουχίες, σύμφωνα με την οποία χρησιμοποιούνται αλγόριθμοι για την επίλυση ζητημάτων όπως η πρόβλεψη ρυθμιστικών περιοχών και των σημείων έναρξης μεταγραφής και μετάφρασης και η σύγκριση μίας άγνωστης ακολουθίας με γνωστές αλληλουχίες μίας βάσης δεδομένων.

Μία ακόμη εφαρμογή είναι η ανακάλυψη γνώσης στη δομική βιοπληροφορική, στην οποία επιλύονται προβλήματα όπως η πρόβλεψη διάφορων χαρακτηριστικών των πρωτεϊνών και η εύρεση της λειτουργίας μίας πρωτεΐνης από τη δομή της.

Άλλες σημαντικές εφαρμογές της ανακάλυψης γνώσης στη βιοπληροφορική είναι ακόμη η ανακάλυψη γνώσης από δεδομένα γονιδιακής έκφρασης και η ανακάλυψη γνώσης από βιολογικά κείμενα.

3.4 Ανάπτυξη εργαλείων στο πλαίσιο της Βιοπληροφορικής

Για την εξαγωγή γνώσης από βιολογικά δεδομένα έχουν δημιουργηθεί και δημιουργούνται ειδικά σχεδιασμένα προγράμματα λογισμικού. Αυτά τα εργαλεία μπορεί να είναι σχεδιασμένα για μία συγκεκριμένη εργασία είτε να είναι γενικής χρήσης, και να είναι περισσότερο ή λιγότερο πολύπλοκα. Από εργαλειοθήκες για διάφορες γλώσσες προγραμματισμού - όπως η Java, η Perl, η Python ή η R - σε πλατφόρμες και ολοκληρωμένα συστήματα, τα προγράμματα αυτά αποτελούν τα εργαλεία της βιοπληροφορικής για την ανάλυση προβλημάτων στους διάφορους τομείς της.

Τα εργαλεία μπορούν να ταξινομηθούν ως εξής [Τζανής, 2011]:

Εργαλεία Ομολογίας και Ομοιότητας (Homology and Similarity tools)

Πρόκειται για εργαλεία που διερευνούν αν δύο αλληλουχίες είναι ομόλογες - αν έχουν, δηλαδή προκύψει από έναν κοινό πρόγονο - ή το βαθμό ομοιότητας μεταξύ δύο αλληλουχιών. Κατά κανόνα η σύγκριση αυτή γίνεται μεταξύ νέων αλληλουχιών, οι ιδιότητες των οποίων δεν είναι γνωστές, με αποθηκευμένες αλληλουχίες για τις οποίες η λειτουργία και η δομή τους είναι γνωστές.

Κάποιο από τα γνωστά εργαλεία σε αυτήν την κατηγορία είναι το BLAST (Basic Local Alignment Search Tool), ένας αλγόριθμος που συγκρίνει πρωτογενή βιολογικά δεδομένα αλληλουχίας. Ουσιαστικά, για μία αλληλουχία αμινοξέων ή νουκλεοτιδίων ως είσοδο, το σύστημα κάνει συγκρίσεις με ήδη γνωστές αλληλουχίες και επιστρέφει ως έξοδο εκείνες με το μεγαλύτερο βαθμό ομοιότητας. Ένα ακόμη εργαλείο είναι ο αλγόριθμος FASTA (FAST All). Πρόκειται για έναν ευριστικό αλγόριθμο που, όπως και ο BLAST, συγκρίνει νουκλεοτιδικές και πρωτεϊνικές αλληλουχίες. Γρήγορα αναγνωρίζει περιοχές με υψηλή ομοιότητα, και σε δεύτερο στάδιο γίνεται πιο λεπτομερής σύγκριση στα επιμέρους τμήματα.

Εργαλεία Ανάλυσης Δομών (Structural Analysis tools)

Με τα εργαλεία ανάλυσης δομών γίνονται συγκρίσεις μεταξύ της δομής άγνωστων πρωτεϊνών και πρωτεϊνών που είναι ήδη γνωστές. Συγκεκριμένα για τις πρωτεϊνικές δομές οι συγκρίσεις αυτές είναι ιδιαίτερα χρήσιμες, μιας και η δισδιάστατη και η τρισδιάστατη δομή των πρωτεϊνών κατέχει σημαντικότερο ρόλο στη λειτουργία τους απ' ότι η σειρά των αμινοξέων που τις συνθέτουν. Ένα εργαλείο αυτής της κατηγορίας είναι το OpenStructure, μία πλατφόρμα ανοιχτού κώδικα, σχεδιασμένη για εφαρμογές δομικής βιολογίας.

Εργαλεία Ανάλυσης Αλληλουχίας (Sequence Analysis tools)

Με αυτά τα εργαλεία υποστηρίζεται η σε βάθος ανάλυση των βιολογικών αλληλουχιών. Η ανάλυση βοηθά στην διευκρίνιση της λειτουργίας της υπό ανάλυση αλληλουχίας, και μπορεί να παρέχει λειτουργικότητα όπως η αναγνώριση μεταλλάξεων και η εξελικτική ανάλυση. Ένα τέτοιο εργαλείο είναι το EMBOSS (European Molecular Biology Open Softerware Suite). Πρόκειται για ένα πακέτο λογισμικού ανοικτού κώδικα που προσφέρει εφαρμογές όπως τη στοίχιση αλληλουχιών, τη γρήγορη αναζήτηση σε βάσεις δεδομένων με πρότυπα αλληλουχιών, την αναγνώριση μοτίβων πρωτεϊνών και πολλές ακόμη.

Εργαλεία Ανάλυσης Λειτουργίας Πρωτεϊνών (Protein Function Tools)

Πρόκειται για εργαλεία που συνεισφέρουν στη σύγκριση πρωτεϊνών με πληροφορίες σχετικά με μοτίβα και πρωτεϊνικές περιοχές, που είναι διαθέσιμες σε δευτερογενείς βάσεις δεδομένων, στις οποίες αναφερθήκαμε νωρίτερα. Η διαδικασία αυτή επιτρέπει την προσέγγιση της βιοχημικής λειτουργίας των πρωτεϊνών που εξετάζονται.

3.5 Πληθυσμιακή Γενετική και Γενετικοί Δείκτες

Η πληθυσμιακή γενετική (population genetics) ή γενετική των πληθυσμών αποτελεί έναν τομέα της εφαρμοσμένης βιολογίας και της γενετικής. Στο πλαίσιό της μελετάται η

γενετική σύνθεση των πληθυσμών, καθώς και η ανάλυση των διεργασιών που επιφέρουν αλλαγές στη γενετική δομή τους κατά τη μεταβίβαση γονιδίων από τα άτομα του πληθυσμού στα άτομα της επόμενης γενιάς. Πιο συγκεκριμένα, ερευνάται το πως η συχνότητα εμφάνισης ενός γονιδίου επηρεάζει τα χαρακτηριστικά των ατόμων κατά το πέρασμα των ετών, τους λόγους για τους οποίους αλλάζει η συχνότητα αυτή, και πως μπορούν οι αλλαγές αυτές να επηρεάσουν την γενετική εξέλιξη των ατόμων.

Με την εξέταση της γενετικής δομής των ατόμων και την παρατήρηση της εμφάνισης ποικιλομορφίας σε συγκεκριμένα τμήματα του γονιδιώματος τους, είναι δυνατός ο χαρακτηρισμός των ατόμων και η ανάθεσή τους σε κάποιον πληθυσμό. Τα τμήματα αυτά είναι αλληλουχίες DNA ή γονίδια που έχουν γνωστή θέση σε ένα χρωμόσωμα και ονομάζονται γενετικοί δείκτες (genetic markers). Τέτοιοι γενετικοί δείκτες είναι οι μικροδορυφόροι (microsatellites) και οι πολυμορφισμοί μονών νουκλεοτιδίων (Single Nucleotide Polymorphisms).

3.5.1. Ανάλυση Πολυμορφισμών Μονών Νουκλεοτιδίων

Οι πολυμορφισμοί μονών νουκλεοτιδίων (SNP - προφέρεται “σνιπ”) αποτελούν γενετικούς δείκτες ενός νουκλεοτιδίου. Αυτό σημαίνει, πως σε αυτές τις αλληλουχίες DNA εμφανίζεται κάποια μετάλλαξη σε μία και μόνο βάση. Τα SNPs έχουν δύο αλληλόμορφα, δηλαδή η βάση στη συγκεκριμένη θέση του γονιδιώματος μπορεί να είναι μία από δύο πιθανές. Στην Εικόνα 3.1 οι βάσεις αυτές είναι οι αδενίνη (A) και κυτοσίνη (C) για το πρώτο χρωματισμένο SNP και οι αδενίνη και γουανίνη (G) για το δεύτερο. Για να θεωρηθεί μία τέτοια διαφορά βάσης ως SNP, θα πρέπει το λιγότερο συχνό αλληλόμορφο να εμφανίζεται με ποσοστό τουλάχιστον 1% [Vignal et al., 2002].

Οι μεταλλάξεις αυτές μπορεί να βρίσκονται μέσα σε γονίδια ή έξω από αυτά (κωδικές ή μη κωδικές περιοχές) και μπορεί να ευθύνονται για διάφορες ασθένειες ή να είναι αβλαβείς.

Εικόνα 3.1: Χρωμοσώματα και SNP

Τα SNPs αποτελούν χρήσιμους γενετικούς δείκτες, εφόσον ένα μικρό σύνολο από SNP περιέχει μεγάλη ποσότητα πληροφορίας. Η πληροφορία αυτή σχετικά με τη γενετική δομή των οργανισμών μπορεί να δώσει απαντήσεις σε διάφορα ενδιαφέροντα ζητήματα, όπως για παράδειγμα - στο ανθρώπινο γονιδίωμα - τον εντοπισμό γονιδίων που σχετίζονται με ασθένειες - όπως ο καρκίνος, ο διαβήτης ή ασθένειες της καρδιάς- , την πιθανή αντίδραση ενός ατόμου σε συγκεκριμένα φάρμακα, την ανίχνευση κληρονομικότητας ασθενειών σε μία οικογένεια ή την ευαισθησία του στους περιβαλλοντικούς παράγοντες. Τα SNPs έχουν επιπλέον εμπορικό ενδιαφέρον στην κτηνοτροφία, όπου μπορούν να δημιουργηθούν προφίλ για πληθυσμούς ζώων, αναγνωρίζοντας σε αυτά επιθυμητούς φαινοτύπους - όπως η τρυφερότητα του κρέατος ή η ανθεκτικότητα σε κάποιο παράσιτο - ή να ανατεθούν άτομα στον πληθυσμό στον οποίο ανήκουν.

Η ανάθεση ατόμων σε πληθυσμούς με βάση τους γενοτύπους τους αποτελεί και το ζήτημα το οποίο πραγματεύεται η παρούσα πτυχιακή εργασία. Πιο συγκεκριμένα, λόγω του μεγάλου αριθμού SNPs είναι απαραίτητο να βρεθούν τρόποι ώστε να αναγνωριστούν ποια είναι τα πιο πληροφοριακά SNPs, με σκοπό την ανάθεση των ατόμων στον πληθυσμό στον οποίο ανήκουν με ικανοποιητική ακρίβεια αλλά και χαμηλό κόστος. Ένας τρόπος επιλογής των πιο πληροφοριακών SNPs είναι με την αξιολόγηση και κατάταξή τους βάσει κάποιας μετρικής, που διαφέρει ανάλογα με τη μέθοδο που χρησιμοποιείται. Αυτές οι γενετικές μέθοδοι φιλτραρίσματος αποτελούν μία γρήγορη, μη μεροληπτική επιλογή για την αξιολόγηση των γενετικών δεικτών. Οι πιο διαδεδομένες από αυτές τις μεθόδους είναι η Wright's FST, η Delta και η Informativeness for Assignment (In) [Καβακιώτης, 2015]

Σε αυτήν την εργασία μελετώνται κάποιες μέθοδοι συγκερασμού κατατάξεων και αν η εφαρμογή τους σε κατατάξεις που προκύπτουν από τις γενετικές μεθόδους που αναφέρθηκαν παραπάνω δίνουν καλύτερα αποτελέσματα στην ανάθεση ατόμων σε πληθυσμούς. Ως καλύτερο αποτέλεσμα ορίζεται το να δίνει ο συγκερασμός κατατάξεις στις οποίες τα πρώτα SNPs είναι περισσότερο πληροφοριακά συγκριτικά με αυτά των γενετικών πρωτευόντων κατατάξεων, άρα και αναθέτουν άτομα σε πληθυσμούς με μεγάλο ποσοστό ακρίβειας χρησιμοποιώντας σχετικά μικρό αριθμό SNP. Επίσης εξετάζεται και αν ο συγκερασμός προσφέρει καλύτερα αποτελέσματα όταν γίνεται με βάση τη βαθμολογία των SNPs αντί για τη σειρά κατάταξής τους. Στην επόμενη παράγραφο γίνεται, λοιπόν, λόγος για τις κατατάξεις και τον συγκερασμό τους.

Κεφάλαιο 4: Κατατάξεις και Μέθοδοι Συγκερασμού Κατατάξεων

Οι μέθοδοι συγκερασμού κατατάξεων χρησιμοποιούνται εδώ και χρόνια σε τομείς όπως αυτοί της θεωρίας κοινωνικής επιλογής (social choice theory), του μάρκετινγκ και της εφαρμοσμένης ψυχολογίας γενικότερα. Τα τελευταία χρόνια, επιπλέον, εμφανίζεται αντίστοιχο ενδιαφέρον και στην εξόρυξη δεδομένων και σε εφαρμογές ανάκτησης πληροφοριών και ακόμη, στον τομέα της βιολογίας. Μερικά παραδείγματα χρήσης του συγκερασμού περιλαμβάνουν την κατάταξη των αθλητών με βάση τις επιδόσεις τους, ή τον ορισμό ενός νικητή σε μία σειρά από παιχνίδια, στον αθλητισμό, το συνδυασμό αποτελεσμάτων από πολλές βάσεις δεδομένων με διαφανή τρόπο από τα ενδιάμεσα λογισμικά (middleware) στην πληροφορική, και την επιλογή του καλύτερου υποψηφίου σύμφωνα με την πλειοψηφία σε μία ψηφοφορία, στη θεωρία κοινωνικής επιλογής [Dwork et al., 2001], [Kolde et al., 2012].

Ο συγκερασμός κατατάξεων με αποδοτικό τρόπο δεν είναι εύκολος σε πραγματικά προβλήματα, όπου τα δεδομένα τα οποία απαιτείται να συνδυαστούν, εκτός από πολυπληθή, είναι και θορυβώδη ή ατελή, δυσχεραίνοντας την διεξαγωγή του συγκερασμού, και απαιτώντας επέκταση των γνωστών μεθόδων.

Στο παρόν κεφάλαιο γίνεται αναφορά στον ορισμό των κατατάξεων και των top-k λιστών, καθώς και του συγκερασμού κατατάξεων. Επιπλέον αναφέρεται ο τρόπος με τον οποίο μπορούν να χρησιμοποιηθούν οι μέθοδοι συγκερασμού στην εύρεση των πιο πληροφοριακών SNPs για την ανάθεση ατόμων σε πληθυσμούς. Τέλος, παρουσιάζονται οι μέθοδοι συγκερασμού που υλοποιήθηκαν στο πλαίσιο της παρούσας πτυχιακής.

4.1. Κατατάξεις και Συγκερασμός Κατατάξεων

Κατάταξη (ranking) ονομάζεται μία λίστα στοιχείων στην οποία τα στοιχεία βρίσκονται τοποθετημένα με σειρά προτίμησης, με το στοιχείο που προτιμάται περισσότερο, συγκριτικά με τα άλλα, στην μικρότερη θέση. Μία τέτοια λίστα αποτελεί

υποσύνολο ενός συνόλου αντικειμένων, μπορεί λοιπόν, να περιέχει κάποια ή και όλα τα δυνατά στοιχεία του συνόλου. Στην πρώτη περίπτωση η κατάταξη ονομάζεται μερική (incomplete ranking), ενώ στην δεύτερη περίπτωση ονομάζεται πλήρης (complete ranking). Σε μία κατάταξη μπορεί επίσης να επιτρέπεται η ισοβαθμία αντικειμένων ή και όχι.
Τα αντικείμενα που περιέχει μία κατάταξη μπορεί να είναι, για παράδειγμα, διαφορετικά προϊόντα, ονόματα υποψηφίων εκλογών ή ακόμη τα αποτελέσματα μίας αναζήτησης στο διαδίκτυο.

Η ύπαρξη και χρήση μίας ολόκληρης λίστας κατάταξης μπορεί να μην είναι επιθυμητή για κάποια εφαρμογή, ή ακόμη και να μην είναι δυνατό να αποκτηθεί. Σε αυτές τις περιπτώσεις, χρησιμοποιούνται λίστες οι οποίες περιέχουν μόνο τα στοιχεία που βρίσκονται υψηλότερα σε μία κατάταξη. Οι λίστες αυτές ονομάζονται *top-k λίστες* και περιέχουν τα k καλύτερα στοιχεία, σύμφωνα με την κατάταξη.

Τα στοιχεία μίας λίστας μπορούν να τοποθετηθούν σε διαφορετική σειρά προτίμησης, από άτομα με διαφορετικές προτιμήσεις, ή ανάλογα με τη μέθοδο ή τα κριτήρια με τα οποία κατατάσσονται, παράγοντας περισσότερες από μία διαφορετικές κατατάξεις. Απαιτείται, λοιπόν, να συνδυαστούν οι κατατάξεις με τρόπο τέτοιο, ώστε στην τελική κατάταξη που προκύπτει, να λαμβάνονται με δίκαιο τρόπο όλες οι επιμέρους υπ' όψιν. Η διαδικασία συνδυασμού των κατατάξεων σε μία ονομάζεται *συγκερασμός κατατάξεων* (rank aggregation) και υπάρχουν πολλές μέθοδοι σύμφωνα με τις οποίες μπορεί να διεκπεραιωθεί. Αξίζει να σημειωθεί πως οι αρχικές κατατάξεις μπορεί να έχουν κάποια ή όλα τα στοιχεία τους κοινά, να περιέχουν θόρυβο, ή ακόμη και να μην περιέχουν κανένα κοινό στοιχείο [Sculley, 2007].

4.2. Μέθοδοι συγκερασμού κατατάξεων

Όπως αναφέρθηκε παραπάνω, υπάρχει μία πληθώρα μεθόδων για την υλοποίηση του συγκερασμού κατατάξεων. Μία δυνατή κατηγοριοποίηση τους είναι ο διαχωρισμός τους σε μεθόδους βασισμένες σε κατανομές (distributional based methods), ευριστικές μεθόδους (heuristic methods) και μεθόδους στοχαστικής αναζήτησης (stochastic search methods) [Lin, 2010].

Μέθοδοι βασισμένες σε κατανομές

Στις μεθόδους αυτές, για κάθε στοιχείο υπάρχει ένα διάνυσμα αριθμών που ακολουθούν κάποιου είδους κατανομή. Με βάση το διάνυσμα αυτό και τις ανά ζεύγη συγκρίσεις των αντικειμένων - σχετικά με το ποιο από τα δύο αντικείμενα βρίσκεται υψηλότερα σε κατάταξη συχνότερα - προκύπτει η τελική κατάταξη. Ένα γνωστό παράδειγμα αυτού του τύπου μεθόδων είναι το μοντέλο του Thurstone (Thurstone's model ή scaling), το οποίο θεωρεί πως οι αριθμοί αυτοί ακολουθούν κανονική κατανομή. Τέτοιες μέθοδοι λειτουργούν συνήθως καλά για το συγκερασμό πολλών λιστών με μικρό αριθμό στοιχείων.

Ευριστικές μέθοδοι

Οι ευριστικές μέθοδοι συγκερασμού κατατάξεων είναι ντετερμινιστικής φύσης και προσφέρουν όχι ακριβείς αλλά προσεγγιστικές λύσεις. Λειτουργούν καλύτερα για μικρό αριθμό λιστών με πολλά στοιχεία, όπως για παράδειγμα για την κατάταξη των πρώτων αποτελεσμάτων που επιστρέφουν κάποιες μηχανές αναζήτησης, ή την μετα-ανάλυση στις ερευνητικές εργασίες. Σε αυτήν την κατηγορία μεθόδων κατατάσσονται και οι μέθοδοι συγκερασμού τύπου Borda και Markov Chain στις οποίες γίνεται εκτενής αναφορά αργότερα.

Μέθοδοι στοχαστικής αναζήτησης

Οι μέθοδοι στοχαστικής αναζήτησης στοχεύουν στην εύρεση μίας βέλτιστης λύσης, παρακάμπτοντας τη συνδυαστική φύση του προβλήματος του συγκερασμού κατατάξεων. Η λύση αυτή είναι βέλτιστη σύμφωνα με κάποιο συγκεκριμένο κριτήριο, στο οποίο εστιάζει η μέθοδος. Όπως και οι ευριστικές μέθοδοι, ενδείκνυνται για το συγκερασμό μικρού αριθμού λιστών με πολλά στοιχεία.

4.2.1. Μέθοδοι Borda

Πρόκειται για μία ευρεία κατηγορία μεθόδων που λειτουργούν με εύκολα κατανοητό τρόπο. Έχουν λάβει το όνομά τους από το Γάλλο μαθηματικό, φυσικό, πολιτικό επιστήμονα και ναυτικό Jean-Charles de Borda, έπειτα από μία μέθοδο την οποία πρότεινε.

Στις μεθόδους αυτές, κάθε ψηφοφόρος θέτει τις επιλογές του -είτε πρόκειται για πολιτικούς, αθλητές, SNPs κ.λπ.- σε σειρά προτίμησης. Με τον τρόπο αυτό, προκύπτει, για κάθε μία από αυτές τις επιλογές, ένα σύνολο από βαθμολογίες- όπου ως βαθμολογία ορίζεται ο αριθμός της θέσης στην οποία βρίσκεται αυτό το στοιχείο στην εκάστοτε κατάταξη. Στη συνέχεια, χρησιμοποιείται μία συνάρτηση ώστε να συνδυαστούν αυτές οι βαθμολογίες και να δώσουν μία τελική βαθμολογία (Borda score) για κάθε επιλογή. Σύμφωνα με αυτήν την καινούρια βαθμολογία δημιουργείται η τελική κατάταξη.

Η μέθοδος που πρότεινε ο Borda χρησιμοποιούσε τον αριθμητικό μέσο ως συνάρτηση συγκερασμού και έκτοτε έχουν δοκιμαστεί διάφορες συναρτήσεις. Κάποιες από αυτές είναι η *διάμεσος* (median), ο *γεωμετρικός μέσος* (geometric mean) και η *p-norm*, οι οποίες και υλοποιήθηκαν για αυτή την εργασία.

Διάμεσος

Σε αυτή τη μέθοδο τοποθετούνται αρχικά όλες οι βαθμολογίες για ένα στοιχείο σε αύξουσα σειρά. Στη συνέχεια, ως το τελικό αποτέλεσμα, επιλέγεται η διάμεσος - το κεντρικό στοιχείο αν ο αριθμός των στοιχείων είναι περιττός ή ο μέσος όρος των δύο κεντρικών τιμών αν ο αριθμός των στοιχείων είναι άρτιος.

$$f(x_1, \dots, x_L) = median\{|x_1|, \dots, |x_L|\} \qquad (1)$$

Για το παράδειγμα του πίνακα 4.1, οι επιλογές 1 και 2 ισοβαθμούν και προηγούνται της τρίτης επιλογής. Τέτοιες περιπτώσεις ισοβαθμίας, μπορούν να λυθούν με διάφορους τρόπους, όπως για παράδειγμα η κατάταξη υψηλότερα της επιλογής που έχει συχνότερα καλύτερη βαθμολογία. Εδώ επιλέγεται τυχαία η επιλογή 1 στην πρώτη θέση και η επιλογή 2 στη δεύτερη.

Πίνακας 4.1: Τρεις διαφορετικές κατατάξεις τριών επιλογών

	Κατάταξη 1	Κατάταξη 2	Κατάταξη 3
Επιλογή 1	2	1	3
Επιλογή 2	1	2	2
Επιλογή 3	3	3	1

Γεωμετρικός Μέσος

Μία ακόμη μέθοδος με την οποία μπορούν να συνδυαστούν κατατάξεις είναι ο γεωμετρικός μέσος, το αποτέλεσμα του οποίου προκύπτει σύμφωνα με τον παρακάτω τύπο.

$$f(x_1, \dots, x_L) = (\prod_{l=1}^{L} |x_L|)^{1/L} \qquad (2)$$

Αυτό σημαίνει, πως υπολογίζεται το γινόμενο των επιμέρους κατατάξεων/βαθμολογιών για ένα στοιχείο, υψώνεται σε μία τιμή - το λόγο ένα προς το πλήθος των κατατάξεων για το στοιχείο αυτό- και αυτή η τιμή αποτελεί τελικά τη θέση του στοιχείου στη νέα κατάταξη.

Για τα στοιχεία του πίνακα 4.1 η κατάταξη που προκύπτει είναι η: 2, 1 και 3, με τιμές που φαίνονται στον πίνακα 4.2.

P-norm

Στην περίπτωση αυτή, για κάθε στοιχείο, αρχικά υπολογίζεται το άθροισμα του κάθε αριθμού υψωμένου σε μία παράμετρο p και έπειτα, το άθροισμα αυτό διαιρείται με το συνολικό πλήθος των κατατάξεων για αυτό το στοιχείο, L. Ο τύπος που συνοψίζει τον υπολογισμό αυτό είναι ο παρακάτω:

$$f(x_1, \dots, x_L) = \sum_{l=1}^{L} {|x_l|^p}/{L} \qquad (3)$$

Για τιμή p = 1, η σχέση αυτή είναι ο αριθμητικός μέσος, και πρόκειται για την μέθοδο που πρότεινε αρχικά ο Borda το 1781.

Πίνακας 4.2: Βαθμολογία επιλογών του πίνακα 4.1 έπειτα από εφαρμογή των μεθόδων διαμέσου, γεωμετρικού μέσου και p-norm με παράμετρο p = 0.5

	Διάμεσος	Γεωμετρικός Μέσος	P-norm (p = 0.5)
Επιλογή 1	2	1,817	1,382
Επιλογή 2	2	1,587	1,276
Επιλογή 3	3	2,080	1,488

Και πάλι στο παράδειγμα του πίνακα 4.1, για p = 0,5, η κατάταξη των στοιχείων συμπίπτει με εκείνη του γεωμετρικού μέσου, με βαθμολογίες που φαίνονται στον πίνακα 4.2.

4.2.2. Μέθοδοι Markov Chain

Οι μέθοδοι που ανήκουν σε αυτή την κατηγορία βασίζονται στις *αλυσίδες Markov* (Markov chains). Πιο συγκεκριμένα, στοχεύουν στη δημιουργία ενός τετραγωνικού, στοχαστικού πίνακα με μη αρνητικές τιμές και μέγεθος ίσο με το συνολικό πλήθος των στοιχείων που πρόκειται να τοποθετηθούν σε κατάταξη. Θεωρώντας πως κάθε ένα από αυτά τα στοιχεία αντιστοιχεί σε μία κατάσταση, ο πίνακας περιέχει την πιθανότητα μετάβασης από κάθε κατάσταση σε κάθε άλλη. Αυτή η πιθανότητα υπολογίζεται με βάση την κατά ζεύγη σύγκριση της θέσης των δύο στοιχείων στις αρχικές κατατάξεις.

Η πιθανότητα μετάβασης που υπολογίζεται για κάθε ζεύγος καταστάσεων στον πίνακα μπορεί να υπολογιστεί με διάφορους τρόπους. Παρακάτω, παρατίθενται τρεις από αυτούς τους τρόπους, στις μεθόδους MC1, MC2 και MC3 [Lin, 2010].

MC1

Σε αυτή την πρώτη μέθοδο, η πιθανότητα μετάβασης από το ένα στοιχείο στο άλλο υπολογίζεται σύμφωνα με το αν υπάρχει έστω και μία αρχική κατάταξη, στην οποία το πρώτο στοιχείο βρίσκεται χαμηλότερα σε προτίμηση από το δεύτερο. Αν αυτό ισχύει, η πιθανότητα μετάβασης από την πρώτη κατάσταση στην άλλη ισούται με ένα προς το συνολικό μέγεθος του χώρου καταστάσεων (S). Διαφορετικά, το αντίστοιχο κελί παίρνει την τιμή 0. Αυτό είναι λογικό, εφόσον αν το πρώτο στοιχείο προτιμάται σε όλες τις αρχικές κατατάξεις, δεν υπάρχει λόγος να γίνει μετάβαση από αυτό στο δεύτερο στοιχείο.

Πίνακας 4.3: Αποτέλεσμα μεθόδου MC1 στο παράδειγμα του πίνακα 4.1

Καταστάσεις/ Αντικείμενα	1	2	3
1	1/3	1/3	1/3
2	1/3	1/3	1/3
3	1/3	1/3	1/3

Επιπλέον, η πιθανότητα μετάβασης από μία κατάσταση στον εαυτό της ισούται με τη διαφορά μεταξύ της μονάδας –πρόκειται για τη μέγιστη δυνατή τιμή, εφόσον

αναφερόμαστε σε πιθανότητες - και του αθροίσματος των πιθανοτήτων μετάβασης από αυτήν την κατάσταση σε όλες τις άλλες. Στον πίνακα 4.3 παρουσιάζονται οι πιθανότητες μετάβασης μεταξύ των τριών καταστάσεων του παραδείγματος του πίνακα 4.1

MC2

Σε αυτή τη μέθοδο, η πιθανότητα μετάβασης από μία κατάσταση σε μία άλλη είναι και πάλι ίση με ένα προς το πλήθος των στοιχείων, δεδομένου, αυτή τη φορά, ότι η πρώτη κατάσταση βρίσκεται σε χαμηλότερη προτίμηση όχι μόνο σε μία αρχική κατάταξη, αλλά στην πλειοψηφία τους. Και πάλι, αν αυτό δεν ισχύει, η αντίστοιχη θέση στον πίνακα συμπληρώνεται με μηδέν. Η πιθανότητα μετάβασης από μία κατάσταση στον εαυτό της υπολογίζεται όπως ακριβώς και στη μέθοδο MC1.

Πίνακας 4.4: Αποτέλεσμα μεθόδου MC2 στο παράδειγμα του πίνακα 4.1

Καταστάσεις/ Αντικείμενα	1	2	3
1	2/3	1/3	0
2	0	1	0
3	1/3	1/3	1/3

MC3

Στη μέθοδο MC3, η πιθανότητα μετάβασης υπολογίζεται αναλογικά με την προτίμηση μεταξύ των δύο υπό εξέταση καταστάσεων στις αρχικές κατατάξεις. Πιο συγκεκριμένα, για τη μετάβαση από το πρώτο στοιχείο στο δεύτερο, η πιθανότητα αντιστοιχεί στο πλήθος των κατατάξεων στις οποίες το δεύτερο στοιχείο προτιμάται από το πρώτο, προς το γινόμενο του αριθμού τον κατατάξεων στις οποίες εμφανίζονται και οι δύο επιλογές που συγκρίνονται, επί το συνολικό πλήθος των αρχικών κατατάξεων.

Τελικά, η νέα κατάταξη προκύπτει από τον υπολογισμό της στάσιμης κατανομής της εργοδικής αλυσίδας Markov που περιγράφει το σύστημα. Υπάρχουν διάφοροι τρόποι υπολογισμού της κατανομής αυτής, όπως ο υπολογισμός του ιδιοδιανύσματος που αντιστοιχεί σε ιδιοτιμή 1 του πίνακα μεταβάσεων, ή παρόμοια η λύση της εξίσωσης $Px = x$, όπου x η στάσιμη κατανομή. Αυτό είναι λογικό, εφόσον εξ' ορισμού η *στάσιμη* κατανομή αντιστοιχεί σε μία μεταβολή στον πίνακα μεταβάσεων που οδηγεί, όμως, στον ίδιο πίνακα.

Πίνακας 4.5: Αποτέλεσμα μεθόδου MC3 στο παράδειγμα του πίνακα 4.1

Καταστάσεις/ Αντικείμενα	1	2	3
1	6/9	2/9	1/9
2	1/9	7/9	1/9
3	2/9	2/9	5/9

Η στάσιμη κατανομή αντιστοιχεί μεγαλύτερη πιθανότητα σε μία κατάσταση με μεγαλύτερη προτίμηση και χαρακτηρίζει τη συμπεριφορά του συστήματος έπειτα από επαρκή αριθμό επαναλήψεων, ανεξάρτητα από την αρχική κατάσταση. Για να επιτευχθεί αυτή τη κατανομή, γίνεται αρχικά μετασχηματισμός του πίνακα μετάβασης, αντικαθιστώντας τις τιμές του σύμφωνα με τον τύπο:

$$P'(u \to v) = (1 - a)P(u \to v) + a/S \qquad (4)$$

όπου α μία παράμετρος, συνήθως μικρή και S το πλήθος των στοιχείων του πίνακα μετάβασης. [Lin, 2010]. Στη συνέχεια υπολογίζεται η στάσιμη κατανομή, δίνοντας ως αποτέλεσμα ένα διάνυσμα με τις τιμές που αντιστοιχούν σε κάθε κατάσταση-στοιχείο του προβλήματος. Τοποθετώντας τα στοιχεία με αύξουσα σειρά των τιμών δημιουργείται η τελική κατάταξη.

Κεφάλαιο 5: Υλοποίηση και Πειραματική Διαδικασία

Σε αυτό το κεφάλαιο γίνεται λεπτομερής αναφορά στην υλοποίηση των μεθόδων συγκερασμού κατατάξεων στις οποίες εστιάζει η παρούσα πτυχιακή εργασία. Οι μέθοδοι αυτές - Borda και Markov Chain - παρουσιάστηκαν στο κεφάλαιο 5. Για την υλοποίησή τους χρησιμοποιήθηκε η γλώσσα προγραμματισμού Java και η εξωτερική βιβλιοθήκη JAMA (JAva MAtrix). Επίσης περιγράφεται με λεπτομέρεια η διαδικασία από τη δημιουργία των νέων κατατάξεων, μέχρι την εκτίμηση των αποτελεσμάτων των μοντέλων, δηλαδή η πειραματική διαδικασία που ακολουθήθηκε.

5.1. Υλοποίηση

Η υλοποίηση χωρίζεται σε δύο κύρια μέρη, που αντιστοιχούν στα πακέτα preprocessing και rankAggregationMethods. Το δεύτερο πακέτο περιέχει δύο επιμέρους πακέτα, το BordaMethods και το MarkovChainMethods, στα οποία βρίσκονται οι κλάσεις που υλοποιούν τις αντίστοιχες μεθόδους συγκερασμού.

Preprocessing

Περιέχει κλάσεις που δέχονται ως είσοδο μία κατάταξη από συγκεκριμένου τύπου αρχεία (προκύπτουν από την εφαρμογή TRES που αναφέρεται σε παρακάτω κεφάλαιο), και κάνουν τις απαραίτητες μετατροπές ώστε να προσφέρουν τα δεδομένα στις μεθόδους συγκερασμού με ομοιόμορφο τρόπο. Πιο συγκεκριμένα, έπειτα από την επεξεργασία σε αυτό το στάδιο προκύπτει μία κατάταξη σε μορφή Map, που περιέχει μόνο τα id και την τιμή του SNP στο κριτήριο που έχει επιλεγεί, ταξινομημένη. Παρακάτω δίνονται περισσότερες λεπτομέρειες για τις επιμέρους κλάσεις αυτού του πακέτου.

Κλάση **SNP**: Περιέχει πεδία με τις πληροφορίες του κάθε SNP, και μεθόδους για την επεξεργασία και προσπέλασή τους.

Πεδία

- private String SNPid: Μοναδικό id που χαρακτηρίζει κάθε SNP.
- private Double rank: Τιμή που αντιστοιχεί στη θέση του κάθε SNP σε μία κατάταξη.
- private Double score: Βαθμολογία του SNP σε μία κατάταξη.

- private Double normalizedScore: Βαθμολογία του SNP έπειτα από κανονικοποίηση και αναστροφή – τιμή στο διάστημα [0.0, 1.0], όπου μικρότερος αριθμός αντιστοιχεί σε καλύτερη βαθμολογία.

Επίσης σε αυτή την κλάση περιέχονται μέθοδοι προσπέλασης και μετάλλαξης, καθώς και βοηθητικές μέθοδοι για τον υπολογισμό της κανονικοποιημένης βαθμολογίας.

Μέθοδοι

- protected static void normalizeSNPScoreValues(List<SNP> SNPList): Μέθοδος για την κανονικοποίηση των βαθμολογιών από μία λίστα SNP. Αξίζει να σημειωθεί πως για να γίνει η κανονικοποίηση μίας τιμής είναι απαραίτητο να είναι γνωστές η μέγιστη και η ελάχιστη τιμή που μπορεί να λάβει αυτή η τιμή. Για αυτό το λόγο η κανονικοποίηση γίνεται με τη χρήση μίας λίστας με όλα τα SNPs μίας κατάταξης και όχι ανά SNP.
- protected static void normalizeAndReverseSNPScoreValues(List<SNP> SNPList): Κανονικοποίηση και αναστροφή των βαθμολογιών μίας λίστας SNP, ώστε ο τρόπος διαχείρισης να συμβαδίζει με εκείνον των κατατάξεων – μικρότερη βαθμολογία συνεπάγεται καλύτερη θέση.
- private static Double normaliseAValue(Double currentValue, Double minValue, Double maxValue): Επιστρέφει την κανονικοποιημένη τιμή της currentValue.
- private static Double reverseScore(Double value): Αναστρέφει μία τιμή. Θεωρείται ότι η τιμή αυτή βρίσκεται στο διάστημα [0.0, 1.0].
- private static Pair<Double, Double> findRange(List<Double> list): Εύρεση της μικρότερης και μεγαλύτερης βαθμολογίας από τα SNPs της λίστας – παραμέτρου.
- public static Map<String, Double> formatOutput(List<SNP> SNPList, int parameter): Μέθοδος που μετατρέπει τη μορφή των δεδομένων, από μία λίστα από SNPs σε ένα

Map που περιέχει το μοναδικό id του κάθε SNP ως κλειδί και το κριτήριο βάσει του οποίου δημιουργήθηκε η κατάταξη ως τιμή. Το κριτήριο αυτό μπορεί να είναι η θέση στην κατάταξη, η βαθμολογία ή, μελλοντικά, κάποιο άλλο. Ποιο από αυτά θα χρησιμοποιηθεί για τη νέα κατάταξη ορίζεται με την παράμετρο parameter, που παίρνει ακέραιες τιμές. Εδώ, για τιμή ίση με 0 χρησιμοποιείται η θέση του SNP στην κατάταξη και 1 η κανονικοποιημένη και ανεστραμμένη βαθμολογία του.

Κλάση **FileOperations**: Σε αυτή την κλάση υλοποιούνται μέθοδοι για την ανάγνωση των αρχείων που περιέχουν τις πληροφορίες των SNP, και την επεξεργασία των δεδομένων αυτών, ώστε να δημιουργηθεί η κατάταξη στην κατάλληλη μορφή. Επίσης υλοποιείται μία μέθοδος για την εγγραφή της κατάταξης σε αρχείο.

Πεδία

- private String inputFilename: Το όνομα του αρχείου από το οποίο θα διαβαστεί η κατάταξη με τις πληροφορίες για το κάθε SNP. Πρόκειται για σχετική διαδρομή στο αρχείο αυτό.
- private String outputFilename: Το όνομα του τελικού αρχείου στο οποίο εγγράφεται η κατάταξη στην τελική της μορφή.

Μέθοδοι

- private List<String> readRawDataFromFile(): Άνοιγμα του αρχείου με τα δεδομένα και εισαγωγή κάθε γραμμής του αρχείου σε μία λίστα, για ευκολότερη διαχείριση και για να μη χρησιμοποιούνται περισσότεροι πόροι από το σύστημα (κλείσιμο αρχείου).
- private List<String> clearInputList(): Απομάκρυνση των γραμμών που αντιστοιχούν σε γραμμές του αρχείου εισόδου που δεν περιέχουν χρήσιμη πληροφορία. Αυτές οι γραμμές περιέχουν τίτλους ή βρίσκονται εκεί για οπτικούς λόγους.
- public List<SNP> fileLinesListToSNPList(): Σε αυτή τη μέθοδο γίνεται επεξεργασία κάθε εγγραφής της λίστας, ώστε να αποσπάσουμε τις πληροφορίες για τα SNPs. Αφού καταγραφεί το μοναδικό id και η βαθμολογία για κάθε SNP και δημιουργηθεί η λίστα-

κατάταξη των SNPs, γίνεται κανονικοποίηση και αναστροφή όλων των βαθμολογιών και αποθηκεύεται επιπλέον αυτή η τιμή σε διαφορετικό πεδίο του κάθε SNP.

- public void writeToFile(Map<String, Double> map): Δημιουργία ενός φακέλου στο project και ενός αρχείου για την εγγραφή των δεδομένων στην κατάλληλη μορφή. Η μορφή αυτή είναι μόνο το id για κάθε SNP, στη σωστή σειρά κατάταξης (μικρότερη τιμή κριτηρίου- value – προς μεγαλύτερη). Η μορφή αυτή απαιτείται για τη χρήση του αρχείου από την εφαρμογή που χρησιμοποιήθηκε για την εκτίμηση των μεθόδων συγκερασμού.

RankAggregationMethods

Σε αυτό το πακέτο βρίσκονται μία κλάση με γενικές, βοηθητικές συναρτήσεις και δύο υπό-πακέτα που περιέχουν τις δύο ομάδες μεθόδων συγκερασμού που υλοποιήθηκαν στην παρούσα εργασία.

Κλάση **RankAggregationDataTransformation**: Περιέχονται μέθοδοι οι οποίες αντιστοιχούν στο ενδιάμεσο βήμα μεταξύ του αποτελέσματος της προ-επεξεργασίας και του συγκερασμού κατατάξεων.

Μέθοδοι

- public static List<Double> getRankingsOfAnElement (List<Map<String, Double>> inputListOfMaps, String element): Σε αυτή τη μέθοδο λαμβάνεται μία λίστα από κατατάξεις στη μορφή id, κριτήριο κατάταξης (η κάθε κατάταξη αποτελεί αποτέλεσμα της προ-επεξεργασίας) και για κάθε SNP συγκεντρώνεται μία λίστα με την τιμή του σε κάθε μία από τις κατατάξεις. Αυτή η συγκεντρωτική λίστα θα χρησιμοποιηθεί κατά τον συγκερασμό. Αξίζει να σημειωθεί πως ένα SNP δε βρίσκεται απαραίτητα σε όλες τις κατατάξεις, άρα και οι λίστες αυτές μπορεί να έχουν διαφορετικό μήκος για κάθε SNP.
- public static Map<String, List<Double>> getRankingsOfAllElements(List<Map<String, Double>> inputListOfMaps, List<String> idList): Βρίσκει τη λίστα με όλες τις κατατάξεις για κάθε SNP και δημιουργεί ένα Map που περιέχει το ID ως κλειδί και τη λίστα αυτή ως τιμή. Με τον τρόπο αυτό έχουμε τελικά κάθε τιμή που χρειάζεται ώστε να υπολογιστεί

η καινούρια κατάταξη για όλα τα SNP. Και πάλι σημειώνεται πως σε αυτή τη δομή βρίσκονται τα SNP από όλες τις κατατάξεις, ανεξάρτητα του αν υπάρχουν στην ίδια κατάταξη.

- public static Map<String, Double> createSortedOutput (Map<String, Double> unsortedRankingMap): Δέχεται ως είσοδο μία κατάταξη (με το ID και την αντίστοιχη τιμή του), ταξινομεί ως προς την τιμή και επιστρέφει την – ταξινομημένη πλέον – κατάταξη.

BordaMethods

Σε αυτό το πακέτο περιέχονται τρεις κλάσεις που αντιστοιχούν σε κάθε Borda μέθοδο που υλοποιείται και μία υπερκλάση τους.

Κλάση **BordaMethod**: Πρόκειται για μία abstract κλάση που περιέχει μεθόδους που χρησιμοποιούνται για την εκτέλεση του συγκερασμού κατατάξεων στις Borda υπό-κλάσεις.

Μέθοδοι

- private Map<String, Double> doTheAggregation (List<Map<String, Double>> listOfRankings): Μέθοδος που συνοψίζει τη διαδικασία του συγκερασμού κατατάξεων. Δέχεται ως είσοδο μία λίστα με τις αρχικές κατατάξεις, κάνει τις κατάλληλες μετατροπές – χρησιμοποιώντας βοηθητικές συναρτήσεις – και δημιουργεί μία καινούρια κατάταξη, με το ID κάθε SNP ως κλειδί και το αποτέλεσμα του συγκερασμού ως τιμή. Το αποτέλεσμα αυτό υπολογίζεται από τη συνάρτηση *computeAggregation*, η οποία υλοποιείται σε κάθε υποκλάση.
- protected static Map<String, List<Double>> transformRankingsForAggregation(List<Map<String, Double>> listOfRankings): Επίσης μία μέθοδος που συνοψίζει διαδικασίες. Πιο συγκεκριμένα, με είσοδο τη λίστα με τις αρχικές κατατάξεις, καλούνται μέθοδοι της κλάσης RankAggregationDataTransformation, ώστε να δημιουργηθεί μία λίστα με τα IDs όλων των SNPs που περιέχονται στις κατατάξεις και στη συνέχεια να δημιουργηθεί λίστα με όλα τα SNPs και τις τιμές όλων των κατατάξεων στις οποίες περιέχονται.

- protected abstract Double computeAggregation(List<Double> numbersToBeAggregated): Abstract μέθοδος. Υλοποιείται στις υπό-κλάσεις και περιέχει την ουσία του συγκερασμού, την υλοποίηση, δηλαδή της εκάστοτε συνάρτησης που χρησιμοποιείται για το συνδυασμό των τιμών που αντιστοιχούν σε κάθε SNP.
- public Map<String, Double> getBordaMethodRanking(List<Map<String, Double>> listOfRankings): Μέθοδος που καλεί τη συνάρτηση *doTheAggregation*, ταξινομεί το αποτέλεσμά της και το δίνει ως έξοδο. Από αυτή τη συνάρτηση προκύπτει η τελική κατάταξη, η οποία χρησιμοποιείται από τη μέθοδο *formatOutput* της κλάσης *FileOperations* που παρουσιάστηκε νωρίτερα.

Κλάση **BordaMedian**: Υποκλάση της BordaMethod που χρησιμοποιεί τη διάμεσο για το συγκερασμό των κατατάξεων.

Μέθοδοι

- protected Double computeAggregation(List<Double> numbersToBeAggregated): Υλοποίηση της abstract μεθόδου της υπερκλάσης, ταξινομεί τη λίστα με τις τιμές ενός SNP και δίνει ως έξοδό της –άρα και ως τελική τιμή του στοιχείου στη νέα κατάταξη - τη διάμεσο. Όπως είναι εμφανές, περισσότερα από ένα στοιχεία μπορεί να λαμβάνουν την ίδια τιμή. Στην παρούσα υλοποίηση, τα στοιχεία με την ίδια τιμή διατηρούν τη σειρά εμφάνισης των SNPs στις αρχικές λίστες.

Κλάση **BordaGeometricMean**: Υποκλάση της BordaMethod που χρησιμοποιεί το γεωμετρικό μέσο για το συγκερασμό των κατατάξεων.

Μέθοδοι

- protected Double computeAggregation(List<Double> numbersToBeAggregated): Αντίστοιχα με την προηγούμενη μέθοδο, εδώ υλοποιείται ο γεωμετρικός μέσος και χρησιμοποιείται το αποτέλεσμά του ως η νέα τιμή του κάθε SNP.

Κλάση **BordaPNorm**: Υποκλάση της BordaMethod που χρησιμοποιεί το p-Norm για το συγκερασμό των κατατάξεων.

Πεδία

- private Double p: Μία μη αρνητική παράμετρος p. Για p=1.0 η συνάρτηση αντιστοιχεί στον αριθμητικό μέσο.

Μέθοδοι

- protected Double computeAggregation(List<Double> numbersToBeAggregated): Η μέθοδος αυτή χρησιμοποιεί την λίστα των αριθμών προς συγκερασμό και την τιμή της παραμέτρου p που τίθεται κατά τη δημιουργία ενός αντικειμένου αυτής της κλάσης και επιστρέφει το αποτέλεσμα της p-Norm για τη συγκεκριμένη λίστα.

Λεπτομερής περιγραφή των παραπάνω συναρτήσεων υπάρχει στο Κεφάλαιο 5.

MarkovChainMethods

Και σε αυτό το πακέτο περιέχονται τρεις κλάσεις, κάθε μία εκ των οποίων υλοποιεί μία διαφορετική Markov Chain μέθοδο για το συγκερασμό των κατατάξεων, καθώς και μία υπερκλάση τους.

Κλάση **MarkovChain**: Πρόκειται για μία abstract κλάση που περιέχει μεθόδους κοινές για όλες τις υποκλάσεις. Επιπλέον περιέχει και μία abstract μέθοδο που αντιστοιχεί στον τρόπο δημιουργίας του πίνακα μεταβάσεων που διαφέρει σε κάθε μέθοδο, άρα και υλοποιείται ξεχωριστά σε κάθε μία από αυτές. Λεπτομερής περιγραφή τους βρίσκεται στο Κεφάλαιο 5.

Πεδία

- protected Matrix transitionMatrix: Πρόκειται για έναν πίνακα (μεταβλητή τύπου Matrix από τη βιβλιοθήκη JAMA) που υλοποιεί τον πίνακα μεταβάσεων των μεθόδων Markov Chain. Ο πίνακας αυτός περιέχει μη αρνητικές τιμές που αντιστοιχούν σε πιθανότητες

μετάβασης από μία κατάσταση σε μία άλλη (Θεωρούμε πως κάθε SNP αντιστοιχεί σε μία κατάσταση). Αυτό σημαίνει πως κάθε γραμμή του πίνακα αυτού έχει άθροισμα τη μονάδα, καθώς και πως η μεγαλύτερη δυνατή τιμή για ένα κελί είναι η τιμή 1.

Μέθοδοι

- protected abstract Matrix createTransitionProbabilityMatrix(List<Map<String, Double>> listOfRankings): Μέθοδος για τη δημιουργία του πίνακα μετάβασης. Υλοποιείται ξεχωριστά σε κάθε υποκλάση.
- private static void transformMCMatrix(Matrix A, Double a): Μετατρέπει τον πίνακα μετάβασης ώστε να μπορεί να προκύψει μία στατική κατανομή από αυτόν.
- private static Matrix stationaryDistribution(Matrix transitionMatrix): Υπολογισμός της στατικής κατανομής του πίνακα μετάβασης με χρήση της μεθόδου δυνάμεων (power method). Από αυτήν την κατανομή προκύπτουν οι τελικές τιμές της κατάταξης για τα στοιχεία, αφαιρώντας την τιμή που αντιστοιχεί σε κάθε SNP από τη μέγιστη – δηλαδή τη μονάδα.
- public Map<String, Double> getMCMethodRanking(List<Map<String, Double>> listOfRankings, Double a): Συγκεντρωτική μέθοδος που μετατρέπει τον πίνακα μετάβασης, δημιουργεί τη στατική κατανομή και ταξινομεί τα αποτελέσματα, επιστρέφοντας, τελικά την τελική κατάταξη.

Κλάση **MC1**: Υποκλάση της κλάσης MarkovChain. Χρησιμοποιεί τη μέθοδο MC1 για τη δημιουργία του πίνακα μετάβασης.

Μέθοδοι

- protected Matrix createTransitionProbabilityMatrix(List<Map<String, Double>> listOfRankings): Γίνεται προσπέλαση του πίνακα μετάβασης, ο οποίος συμπληρώνεται με τιμή διαφορετική του μηδέν αν το στοιχείο στήλης βρίσκεται υψηλότερα από το στοιχείο γραμμής τουλάχιστον σε μία από τις αρχικές κατατάξεις.

Κλάση **MC2**: Υποκλάση της κλάσης MarkovChain. Χρησιμοποιεί τη μέθοδο MC2 για τη δημιουργία του πίνακα μετάβασης.

Μέθοδοι

- protected Matrix createTransitionProbabilityMatrix(List<Map<String, Double>> listOfRankings): Συμπλήρωση του πίνακα μετάβασης, με τιμή διαφορετική του μηδέν αν το στοιχείο στήλης βρίσκεται υψηλότερα από το στοιχείο γραμμής στην πλειοψηφία των αρχικών κατατάξεων.

Κλάση **MC3**: Υποκλάση της κλάσης MarkovChain. Αντίστοιχα με τις άλλες δύο, χρησιμοποιεί τη μέθοδο MC3 για τη δημιουργία του πίνακα μετάβασης.

Μέθοδοι

- protected Matrix createTransitionProbabilityMatrix(List<Map<String, Double>> listOfRankings): Αυτή τη φορά, το κάθε κελί του πίνακα συμπληρώνεται με τιμή ανάλογη του αριθμού των κατατάξεων στις οποίες το στοιχείο στήλης υπερέχει του στοιχείου γραμμής, στις κατατάξεις στις οποίες υπάρχουν και τα δύο στοιχεία.

Επιπλέον υπάρχουν οι κλάσεις Main και Evaluation. Στη Main καλούνται οι κατάλληλες συναρτήσεις για τη δημιουργία νέων κατατάξεων από ήδη υπάρχουσες.

Κλάση **Evaluation**: Σε αυτήν την κλάση περιέχονται συναρτήσεις για την αξιολόγηση των τελικών αποτελεσμάτων του πειραματικού τμήματος της παρούσας εργασίας, το οποίο παρουσιάζεται σε επόμενο κεφάλαιο. Ως είσοδο δέχεται ένα φάκελο που περιέχει υπο-φακέλους για κάθε μέθοδο. Σε κάθε έναν από αυτούς τους φακέλους περιέχονται .csv αρχεία με τα αποτελέσματα της ανάθεσης ατόμων.

Μέθοδοι

- private static List<String> getRawLinesFromFile(String inputFilename): Μέθοδος που διαβάζει από ένα αρχείο εισόδου και αποθηκεύει τα δεδομένα σε μία λίστα από αλφαριθμητικά για ευκολότερη διαχείριση, παρόμοια με την κλάση *FileOperations*.
- private static List<Pair<String, String>> lineListToListOfPairs(List<String> inputListOfLines): Η μέθοδος αυτή χρησιμοποιεί το αποτέλεσμα της *getRawLinesFromFile*, αφαιρεί τις πρώτες γραμμές που δεν περιέχουν ουσιαστικά δεδομένα και δημιουργεί μία λίστα από ζεύγη αλφαριθμητικών. Αυτά τα αλφαριθμητικά αντιστοιχούν, το πρώτο στον πληθυσμό στον οποίο τοποθετήθηκε ένα SNP, και το δεύτερο στον πληθυσμό στον οποίο είναι πιθανότερο να ανήκει το άτομο στην πραγματικότητα. Όλες οι υπόλοιπες πληροφορίες του αρχείου παραλείπονται, καθώς δεν παρουσιάζουν ενδιαφέρον για την εφαρμογή μας.
- private static Double computeEvaluation (List<Pair<String, String>> listOfPairs): Σε αυτή τη μέθοδο γίνεται προσπέλαση της λίστας με τα ζεύγη αλφαριθμητικών και υπολογίζεται σε πόσα από τα συνολικά ζεύγη, τα δύο τμήματα έχουν την ίδια τιμή. Αυτό σημαίνει, πως το μοντέλο έχει τοποθετήσει το άτομο στον πιθανότατα σωστό πληθυσμό. Το ποσοστό των ατόμων που έχουν ανατεθεί σωστά αποτελεί την απόδοση της μεθόδου, και είναι η τιμή η οποία επιστρέφεται από αυτήν την συνάρτηση.
- public static void computeEvaluationForAllValues (File folder, String outFile): Μέθοδος που δέχεται ως είσοδο ένα φάκελο που περιέχει όλα τα αρχεία στα οποία απαιτείται να γίνει αξιολόγηση, και ένα απόλυτο μονοπάτι για την έξοδο των αποτελεσμάτων.
- public static void computeEvaluationPerMethod(File folder): Σε αυτή τη μέθοδο δίνεται ως είσοδος ένας φάκελος που περιέχει αρχεία με την ανάθεση ατόμων σε πληθυσμούς σύμφωνα με μία συγκεκριμένη μέθοδο. Υπάρχουν, λοιπόν, αρχεία που αντιστοιχούν στα αποτελέσματα για τα top-10, top-20, …, top-100 SNPs, όπως αυτά έχουν καταταχθεί από μία μέθοδο, είτε πρόκειται για τις γενετικές μεθόδους, είτε για τις μεθόδους συγκερασμού.
- public static void main(String[] args): Η συνάρτηση main η οποία «τρέχει» για την αξιολόγηση των αποτελεσμάτων. Εδώ ορίζεται ο φάκελος από τον οποίο θα

διαβαστούν τα αρχεία με τα αποτελέσματα και καλείται η μέθοδος που αξιολογεί ανά μέθοδο.

Σε αυτήν την παράγραφο παρουσιάστηκε με λεπτομέρεια το τμήμα της υλοποίησης της παρούσας πτυχιακής. Πιο συγκεκριμένα, αναλύθηκε η δομή του κώδικα, οι μέθοδοι που υλοποιήθηκαν για την επεξεργασία των ακατέργαστων δεδομένων, για τη δημιουργία των συγκερασμένων κατατάξεων, και ακόμη για την αξιολόγηση των μεθόδων βάσει των μοντέλων πρόβλεψης που προέκυψαν από τα αποτελέσματά τους. Στη συνέχεια παρουσιάζεται η πειραματική διαδικασία που ακολουθήθηκε για τη δημιουργία των μοντέλων αυτών.

5.2. Πειραματική Διαδικασία

Για την εκτίμηση της απόδοσης των κατατάξεων που προκύπτουν από το συγκερασμό σε σχέση με τις αρχικές γενετικές κατατάξεις ελέγχεται η επιτυχία του μοντέλου πρόβλεψης που προκύπτει από κάθε κατάταξη. Συγκρίνεται, δηλαδή το ποσοστό ανάθεσης ατόμων στο σωστό πληθυσμό, μεταξύ των μοντέλων που προκύπτουν από τις αρχικές και των μοντέλων που προκύπτουν από τις συγκερασμένες κατατάξεις.

Τα δεδομένα στα οποία βασίστηκε η πειραματική διαδικασία αποτελούν ένα σύνολο από SNP διαφόρων ατόμων από πληθυσμούς χοίρων. Πέραν των δεδομένων, χρησιμοποιήθηκαν οι εφαρμογές TRES [Καβακιώτης, 2015] και GeneClass2 [Piry et al., 2004] για τη δημιουργία γενετικών κατατάξεων, μετατροπές αρχείων και διαχωρισμό των ατόμων για την εκπαίδευση και τη δοκιμή του μοντέλου (Train και Test sets).

Αρχικά, γίνεται μετατροπή του αρχείου με τα δεδομένα από .ped σε .arff στην εφαρμογή TRES, ώστε να βρίσκονται σε μορφή που μπορεί η εφαρμογή να χειριστεί. Στη συνέχεια, γίνεται διαχωρισμός του αρχείου σε train και test για την εκπαίδευση και τη δοκιμή του μοντέλου αντίστοιχα. Προκύπτουν λοιπόν δύο αρχεία, ένα που περιέχει – στην περίπτωσή μας - το 70% των αρχικών δεδομένων, και ένα που περιέχει το εναπομείναν 30%.

Αφού γίνει ο παραπάνω διαχωρισμός των δεδομένων, χρησιμοποιείται το αρχείο με τα δεδομένα εκπαίδευσης για τη δημιουργία των γενετικών κατατάξεων. Μέσα από την εφαρμογή TRES, λαμβάνουμε τις κατατάξεις από τις μεθόδους Delta, Pairwise Wright's Fst και Informativeness for Assignment. Πιο συγκεκριμένα, δημιουργούμε αρχεία, καθ' ένα εκ των οποίων περιέχει μία top-k κατάταξη από κάθε μέθοδο, για αριθμό από δέκα μέχρι εκατό SNPs, με βήμα δέκα.

Τα αρχεία αυτά χρησιμοποιούνται στη συνέχεια από τον κώδικα που έχει αναπτυχθεί, ώστε να γίνει συγκερασμός των κατατάξεων. Για τα αρχεία με τον ίδιο αριθμό SNPs, προκύπτουν έξι νέα αρχεία, καθένα εκ των οποίων περιέχει τα top-k SNPs από κάθε μέθοδο συγκερασμού. Αξίζει να σημειωθεί πως σε αυτό το σημείο το πλήθος των SNPs που προκύπτουν από το συγκερασμό δεν είναι απαραίτητο να είναι ίδιο με το πλήθος των SNPs των αρχικών κατατάξεων, αλλά μπορεί να είναι μεγαλύτερο. Αυτό συμβαίνει διότι είναι πιθανό οι αρχικές κατατάξεις να μην περιέχουν μόνο κοινά SNPs. Κατά το συγκερασμό όμως, γίνεται κατάταξη στο σύνολο των SNPs από όλες τις κατατάξεις, οδηγώντας στη δημιουργία μίας top-k λίστας με περισσότερα στοιχεία.

Μόλις δημιουργηθούν όλα τα αρχεία με τις κατατάξεις, χρησιμοποιούνται για την παραγωγή μειωμένων συνόλων δεδομένων, τη δημιουργία, δηλαδή, αρχείων τύπου GENEPOP, που αποτελούν είσοδο για την εφαρμογή GeneClass2. Για τη δημιουργία τους χρησιμοποιούμε τα αρχεία train και test, καθώς και το αρχείο με την κατάταξη του εκάστοτε αριθμού SNPs. Επίσης ορίζεται και ο αριθμός των SNP της κατάταξης που θα ληφθεί υπ' όψιν για τη δημιουργία του αρχείου - χρήσιμο ιδιαίτερα για τις κατατάξεις που προέκυψαν από το συγκερασμό, ώστε να χρησιμοποιούμε τον ίδιο αριθμό SNPs όπως και με τις αρχικές κατατάξεις. Με βάση αυτά παράγει ως έξοδο ένα αρχείο πληθυσμού αναφοράς και ένα αρχείο με άτομα προς ανάθεση – που προκύπτουν με βάση το train και test αρχείο, αντίστοιχα. Η διαδικασία αυτή ακολουθείται για κάθε διαφορετική κατάταξη και κάθε δεκάδα SNP.

Σε αυτό το σημείο χρησιμοποιούμε τα αρχεία με τον πληθυσμό αναφοράς και τα άτομα προς ανάθεση στην εφαρμογή GeneClass2. Η εφαρμογή αυτή παρέχει έναν αριθμό γενετικών μεθόδων ανάθεσης, οι οποίες αναθέτουν τα άτομα στους πληθυσμούς στους

οποίους πιθανόν ανήκουν με βάσει τη γενετική τους σύσταση. Προκύπτει, λοιπόν ένα αρχείο μορφής .csv, το οποίο περιέχει γραμμές που αντιστοιχούν σε κάθε άτομο, με την πρώτη στήλη να περιέχει το όνομα του πληθυσμού στον οποίο ανατέθηκε το άτομο, και στις επόμενες –ανά ζεύγη- την πιθανότητα να ανήκει σε κάποιον πληθυσμό και το όνομα αυτού του πληθυσμού.

Η παραπάνω διαδικασία ακολουθήθηκε για το συγκερασμό με βάση τη θέση, αλλά και με βάσει τη βαθμολογία του κάθε SNP στις αρχικές κατατάξεις, ώστε να ελεγχθεί αν η διαφορά στη βαθμολογία φέρει κάποια επιπλέον πληροφορία που βοηθά στη δημιουργία καλύτερων αποτελεσμάτων.

Αποτελέσματα

Για την αξιολόγηση των αποτελεσμάτων της παραπάνω διαδικασίας, υπολογίζεται το ποσοστό ακρίβειας ανάθεσης των ατόμων σε πληθυσμούς για τα δεδομένα της κάθε μεθόδου. Στους παρακάτω πίνακες – Πίνακες 5.1 και 5.2 – καταγράφονται αυτά τα ποσοστά ακρίβειας, ανά δέκα SNPs. Ο Πίνακας 5.1 περιέχει τα αποτελέσματα όταν ο συγκερασμός των κατατάξεων έγινε με βάση τη θέση των SNPs στις αρχικές κατατάξεις, ενώ ο Πίνακας 5.2 όταν έγινε με βάση τη βαθμολογία τους.

Πίνακας 5.1: Ποσοστά απόδοσης όταν η συγκερασμένη κατάταξη προκύπτει βάσει της θέσης του κάθε SNP στις αρχικές κατατάξεις (rank)

#SNP	Delta	Wright's Fst	In	Borda GMean	Borda Median	Borda PNorm (p = 2)	MC1	MC2	MC3
10	44.20	46.37	62.31	52.17	44.20	59.42	50.72	44.20	50.72
20	62.31	84.05	78.26	76.08	77.53	76.08	76.08	84.78	76.08
30	89.85	92.02	89.13	90.57	86.23	93.47	90.57	87.68	87.68
40	93.47	91.30	92.75	93.47	93.47	92.02	92.02	94.20	92.02
50	94.20	90.57	93.47	92.75	92.75	92.75	92.75	93.47	92.75
60	93.47	92.02	94.20	94.92	94.92	93.47	92.75	97.10	92.75
70	96.37	94.92	93.47	94.92	95.65	96.37	94.20	99.27	94.20
80	96.37	95.65	94.20	97.10	96.37	95.65	94.20	99.27	94.20
90	96.37	94.92	94.20	98.55	97.10	97.82	94.92	100.0	94.92
100	97.10	96.37	96.37	99.27	98.55	99.27	95.65	100.0	96.37

Σημειώνεται πως για τη δημιουργία των νέων κατατάξεων από τις Markov Chain μεθόδους χρησιμοποιήθηκαν μόνο τα 500 πρώτα SNPs από τις αρχικές κατατάξεις, εφόσον η δημιουργία τους με βάση όλο το σύνολο των SNPs ήταν αδύνατη υπολογιστικά στο μηχάνημα στο οποίο εκτελέσθηκε η πειραματική διαδικασία.

Με μπλε χρώμα σημειώνεται η κατάταξη - ή οι κατατάξεις - με την καλύτερη απόδοση ανά αριθμό SNPs, ενώ με πράσινο οι αποδόσεις των συγκερασμένων κατατάξεων, οι οποίες ξεπερνούν αυστηρά την απόδοση τουλάχιστον μίας από των αρχικών κατατάξεων, και πάλι ανά αριθμό SNPs.

Πίνακας 5.2: Ποσοστά απόδοσης όταν η συγκερασμένη κατάταξη προκύπτει βάσει της βαθμολογίας του κάθε SNP στις αρχικές κατατάξεις (score)

#SNP	Delta	Wright's Fst	In	Borda GMean	Borda Median	Borda PNorm (p = 2)	MC1	MC2	MC3
10	44.20	46.37	62.31	50.72	56.52	62.31	50.72	44.20	50.72
20	62.31	84.05	78.26	77.53	84.05	77.53	76.08	77.53	76.08
30	89.85	92.02	89.13	85.50	92.02	87.68	90.57	87.68	87.68
40	93.47	91.30	92.75	93.47	91.30	92.75	92.02	94.20	92.02
50	94.20	90.57	93.47	92.75	91.30	92.75	92.75	93.47	92.75
60	93.47	93.47	94.20	94.92	92.02	93.47	92.75	97.10	92.75
70	96.37	94.92	93.47	94.92	94.20	96.37	94.20	99.27	94.20
80	96.37	95.65	94.20	97.82	96.37	97.10	94.20	100.0	94.20
90	96.37	94.92	94.20	97.82	96.37	97.10	94.92	100.0	94.92
100	97.10	96.37	96.37	99.27	96.37	97.10	95.65	100.0	96.37

Παρατηρούμε ότι ενώ για πολύ μικρό αριθμό SNPs, οι γενετικές κατατάξεις φαίνεται να έχουν καλύτερο ποσοστό ακρίβειας ανάθεσης, γρήγορα οι συγκερασμένες κατατάξεις καλύπτουν το έδαφος δίνοντας ικανοποιητικά ποσοστά ακρίβειας. Παράλληλα, όμως, η διαφορά μεταξύ των αποτελεσμάτων Φαίνεται όμως ότι μετά από τα 90 – 100 SNPs τόσο οι αρχικές όσο και οι συγκερασμένες κατατάξεις καταλήγουν να έχουν αντίστοιχα υψηλό ποσοστό ακρίβειας, με τις συγκερασμένες κατατάξεις να έχουν ένα μικρό προβάδισμα.

Αντίστοιχα είναι και τα αποτελέσματα που προέκυψαν από τη χρήση της βαθμολογίας των SNPs για τη δημιουργία των νέων κατατάξεων. Παρατηρείται ότι τα αποτελέσματα δεν είναι πολύ διαφορετικά από ότι τα προηγούμενα σε σύγκριση με τις

αρχικές κατατάξεις, εκτός από το γεγονός ότι για μικρό αριθμό SNPs, δύο επιπλέον ποσοστά ισοβαθμούν στην πρώτη θέση με εκείνα των συγκερασμένων κατατάξεων.

Από τη σύγκριση μεταξύ των αποτελεσμάτων των κατατάξεων με βάση τη θέση των στοιχείων και των αποτελεσμάτων των κατατάξεων που βασίστηκαν σε βαθμολογίες, παρατηρούνται κάποιες επιπλέον διαφορές. Αρχικά, για μικρό αριθμό SNPs, φαίνεται κάποιες μέθοδοι να δίνουν αρκετά καλύτερα ποσοστά (Borda Median, Borda p-Norm), συνολικά όμως, τα ποσοστά ακρίβειας δεν είναι υψηλότερα σε σχέση με εκείνα των αρχικών κατατάξεων, πράγμα το οποίο γίνεται αρκετά εμφανές και για μεγαλύτερο αριθμό SNPs. Ξεκάθαρο παράδειγμα αποτελούν τα αποτελέσματα για τη μέθοδο Borda Median, όπου τα ποσοστά για αριθμό SNPs μεγαλύτερο των 40 είναι όλα μικρότερα σε σχέση με τα αποτελέσματα της ίδιας μεθόδου στον Πίνακα 5.1.

Φαίνεται, συνολικά, πως η χρήση των συγκερασμένων κατατάξεων προσφέρει οριακά καλύτερα αποτελέσματα, χωρίς όμως να ξεπερνά σημαντικά τις κατατάξεις που προκύπτουν από τις γενετικές μετρικές. Αυτό σημαίνει πως ενώ οι συγκεκριμένες συγκερασμένες κατατάξεις δεν είναι επαρκείς να αντικαταστήσουν τις αρχικές, μπορεί με τη χρήση διαφορετικών κριτηρίων ή την εστίαση στη σχετική διαφορά των SNPs (σε τι βαθμό το ένα είναι καλύτερο από το άλλο) τα αποτελέσματα να είναι καλύτερα. Υπάρχει, λοιπόν, χώρος για έρευνα πάνω σε αυτό το ζήτημα, στο οποίο γίνεται αναφορά στο κεφάλαιο.

Κεφάλαιο 6: Επίλογος και Μελλοντική Εργασία

Στο παρόν κεφάλαιο γίνεται μία ανασκόπηση των αποτελεσμάτων της πτυχιακής εργασίας και συζήτηση για πιθανές βελτιώσεις και περαιτέρω έρευνα στο ζήτημα της επιλογής χαρακτηριστικών για δεδομένα βιοπληροφορικής.

6.1. Συμπεράσματα

Στην παρούσα εργασία μελετήθηκε το ζήτημα της επιλογής χαρακτηριστικών σε γενετικούς δείκτες, και πιο συγκεκριμένα σε σύνολα από SNPs, με στόχο τη δημιουργία μοντέλων τα οποία χρησιμοποιώντας μικρότερο αριθμό SNPs προσφέρουν αποδοτικότερα μοντέλα ανάθεσης ατόμων σε πληθυσμούς. Σε μία προσπάθεια βελτίωσης των μοντέλων αυτών, χρησιμοποιήθηκαν top-k κατατάξεις των πληροφοριακότερων SNPs για τη δημιουργία τους. Τα top-k στοιχεία προέκυψαν από συστήματα γενετικών κατατάξεων (base rankers) και επιπλέον από συνδυασμό αυτών των αρχικών κατατάξεων με τη χρήση αλγορίθμων συγκερασμού κατατάξεων.

Υλοποιήθηκαν, λοιπόν, σε αυτό το πρώτο μέρος της εργασίας, έξι μέθοδοι για το συγκερασμό κατατάξεων με βάση τη θέση των στοιχείων στις αρχικές κατατάξεις (order-based aggregation), καθώς και βάσει της βαθμολογίας των SNPs. Η πληροφορία αυτή βρίσκεται στις αρχικές γενετικές κατατάξεις και μελετήθηκε ώστε να διευκρινιστεί αν η γνώση της σχετικής θέσης των SNPs στις top-k καταστάσεις μπορεί να οδηγήσει σε χρησιμότερες, για τη δημιουργία των μοντέλων, κατατάξεις.

Έπειτα από τη δημιουργία των top-k κατατάξεων, για ένα συγκεκριμένο σύνολο δεδομένων, ακολούθησε το πειραματικό τμήμα της εργασίας, στο οποίο έγινε αξιολόγηση των νέων, συνδυασμένων κατατάξεων, σε σχέση με τις αρχικές. Για να επιτευχθεί αυτό, δημιουργήθηκαν μοντέλα πρόβλεψης για την ανάθεση ατόμων σε πληθυσμούς. Το κάθε μοντέλο προέκυψε από μία κατάταξη, και για διαφορετικό αριθμό SNPs ανά κατάσταση.

Επίσης, χωριστά αξιολογήθηκαν τα μοντέλα που προέκυψαν από κατατάξεις βασισμένες στη θέση των στοιχείων και εκείνα που βασίστηκαν στη βαθμολογία του κάθε SNP.

Τελικά, προέκυψαν καλύτερα αποτελέσματα από τα μοντέλα που δημιουργήθηκαν έπειτα από το συγκερασμό των κατατάξεων; Όπως φαίνεται και στους Πίνακες 5.1 και 5.2 του προηγούμενου κεφαλαίου, υπήρξε κάποια αύξηση στο ποσοστό επιτυχημένων αναθέσεων από τις αρχικές κατατάξεις στις συγκερασμένες για κάποιο αριθμό SNPs. Παρατηρήθηκε, όμως, ταυτόχρονα, ότι τα αποτελέσματα των καινούριων κατατάξεων δεν είναι σημαντικά καλύτερα από εκείνα των αρχικών, ακόμη και για μεγαλύτερο αριθμό SNPs. Παρομοίως, για μικρό αριθμό SNPs, η χρήση της βαθμολογίας για τη δημιουργία της κατάταξης, φαίνεται να δίνει χαμηλότερα ποσοστά για κάποιες μεθόδους – σε σημείο ώστε να αλλάζει η μέθοδος που κρατά τα καλύτερα αποτελέσματα – και υψηλότερα για κάποιες άλλες. Αυτό σημαίνει πως δεν υπάρχει ένα καθολικό συμπέρασμα από την σύγκριση των αποτελεσμάτων μεταξύ των συγκερασμένων κατατάξεων που προκύπτουν από τα δύο κριτήρια. Τελικά, και τα δύο κριτήρια – για μεγαλύτερο αριθμό SNP – έχουν τα καλύτερα αποτελέσματα από τη μέθοδο MC2.

Οδηγούμαστε, τελικά, στο συμπέρασμα ότι οι μέθοδοι συγκερασμού, με τα κριτήρια που υλοποιήθηκαν στην παρούσα εργασία, δεν είναι ικανές να αντικαταστήσουν εξ ολοκλήρου τις αρχικές, γενετικές κατατάξεις. Είναι, όμως, εμφανές, πως υπάρχει έδαφος για περαιτέρω αναζητήσεις σε αυτόν τον τομέα, με στόχο τη βελτίωση των αποτελεσμάτων.

6.2. Μελλοντικές Επεκτάσεις

Από την πλευρά της δημιουργίας συγκερασμένων κατατάξεων, μία πιθανή βελτίωση έρχεται από τη μετατροπή των μεθόδων συγκερασμού, με τρόπο τέτοιο, ώστε να δίνεται διαφορετική βαρύτητα στις επιμέρους κατατάξεις. Στην παρούσα εργασία οι αρχικές κατατάξεις συμμετέχουν εξίσου στο τελικό αποτέλεσμα. Θα μπορούσε, λοιπόν, μία επέκτασή της να περιέχει ένα διάνυσμα βαρών, το οποίο θα δίνει μεγαλύτερη σημαντικότητα, για συγκεκριμένο αριθμό SNPs, στην κατάταξη που οδηγεί σε καλύτερα

αποτελέσματα. Επιπλέον, τα βάρη θα μπορούσαν να βελτιώνονται σε κάθε επανάληψη της δημιουργίας μοντέλου για τον ίδιο πληθυσμό.

Όπως φάνηκε στα αποτελέσματα της εργασίας, η χρήση των βαθμολογιών των SNPs απέφερε αλλαγές στην ποιότητα των μοντέλων. Θα μπορούσε, κατ' επέκταση, να δοθεί βαρύτητα σε αυτήν την πληροφορία, και να γίνει έρευνα σχετικά με το πώς μπορεί να χρησιμοποιηθεί καλύτερα η σχετική θέση των SNPs στην κάθε κατάταξη.

Σημαντική θα ήταν, επιπλέον, η λήψη περισσότερων πληροφοριών για τα SNPs από την σκοπιά της βιολογίας. Με τον τρόπο αυτό, μπορούν να γίνουν μετατροπές που στηρίζονται στη φύση τους και οι οποίες δε λαμβάνονται υπ' όψιν από γενικού σκοπού μεθόδους και πρακτικές, όπως οι μέθοδοι που υλοποιήθηκαν. Με αυτή τη γνώση, μπορούν να γίνουν στοχευμένες υποθέσεις για τα SNPs, που θα βοηθήσουν στη δημιουργία κατάταξης που περιέχει τα SNP με την περισσότερη και ποιοτικότερη πληροφορία. Ένα παράδειγμα γι' αυτό, θα ήταν η αφαίρεση των ισχυρά συνδεδεμένων SNPs από τη λίστα, εφόσον η πληροφορία που παρέχουν είναι παρόμοια.

Παρατηρούμε συνολικά, πως υπάρχει χώρος για βελτιώσεις, και πως το ζήτημα της επιλογής χαρακτηριστικών στον τομέα της Βιοπληροφορικής μπορεί να οδηγήσει σε σημαντικά αποτελέσματα. Είναι ιδιαίτερα εμφανές πως με την αύξηση της διαθέσιμης πληροφορίας - λόγω της προόδου της βιοτεχνολογίας – κρίνεται πλέον απαραίτητη η επεξεργασία των δεδομένων με τρόπο χρονικά και υπολογιστικά συμφέρον. Φυσικά, πρέπει και τα αποτελέσματα της επεξεργασίας αυτής να είναι ποιοτικά, να δίνουν, δηλαδή, ακριβή και χρήσιμα αποτελέσματα, στοχεύοντας, τελικά, στη βελτίωση της ποιότητας ζωής των ανθρώπων.

Βιβλιογραφία

[Agrawal et al., 1994] Agrawal R., Srikant R. (1994), *Fast algorithms for mining association rules between sets of items in large databases*, IBM Almaden Research Center, California

[Arora et al., 2005], Arora Sanjeev, Nabieva Elena, *Lecture 7: Markov Chains and Random Walks,* Princeton University, 2005

[Äyrämö et al., 2006] Äyrämö S., Kärkkäinen T. (2006), *Introduction to partitioning-based clustering methods with a robust example*, University of Jyväskylä, Department of Mathematical Information Technology, Finland

[Brancotte et al., 2015], Brancotte Bryan, Yang Bo, Blin Guillaume, Cohen-boulakia Sarah et al., *Rank aggregation with ties : Experiments and Analysis Rank Aggregation problem*, Proceedings of the 41st International Conference on Very Large Data Bases, 2015

[Dwork et al., 2001], Dwork Cynthia, Kumar Ravi, Naor Moni, Sivakumar D., *Rank Aggregation Revisited*, Issue 2, Vol. 13, Systems Research Journal, 2001

[EBI, 2015] European Bioinformatics Institute (2015), *Annual Scientific Report 2015* [Digital Edition], European Molecular Biology Laboratory, ανασύρθηκε στις 30 Νοεμβρίου

[Frawley et al. 1991] Frawley W.J., Piatesky-Shaphiro G., Matheus C.J. (1991), *Knowledge Discovery in Databases*, AAAI/MIT Press

[Henikoff et al., 2000] Henikoff G. J., Pietrokovski S., McCallum M. C., Henikoff S. (2000), *Blocks-based methods for detecting protein homology*, Electrophoesis

[Kolde et al., 2012], Kolde R., Laur S., Adler P. et al, *Robust rank aggregation for gene list integration and meta-analysis*, Bioinformatics (Journal), Volume 28, Issue 4, 2012

[Lin, 2010] Lin Shili, *Rank aggregation methods*, WIREs Comp Stat, 2010

[Liu et al., 2007], Liu Yu-Ting, Liu Tie-Yan, Qin Tao, Ma Zhi-Ming, Li Hang, *Supervised rank aggregation*, Proceedings of the 16th international conference on World Wide Web - WWW '07, 2007

[Piry et al., 2004] Piry S, Alapetite A, Cornuet, J.-M., Paetkau D, Baudouin, L., Estoup, A. (2004), *GeneClass2: A Software for Genetic Assignment and First-Generation Migrant Detection*, *Journal of Heredity* **95**:536-539

[Qin et al., 1999] Qin J., Norton M. J. (1999), *Knowledge Discovery in Bibliographic Databases*, Library Trends, Summer, University of Illinois, Graduate School of Library and Information Science

[Russell et al., 2005] Russell S., Norvig P. (2005), *Τεχνητή Νοημοσύνη: Μία σύγχρονη προσέγγιση*, Κλειδάριθμος

[Sculley, 2007] Sculley D., *Rank Aggregation for Similar Items*, Proceedings of the 2007 SIAM International Conference on Data Mining, Pages: 587-592, 2007

[Tan et al., 2016] Tan Pang-Ning, Steinbach M., Kumar V. (2016), *Εισαγωγή στην Εξόρυξη Δεδομένων*, Εκδόσεις Τζιόλα

[Vignal et al., 2002] Vignal A., Milan D., SanCristobal M., Eggem A., *A review on SNP and other types of molecular markers and their use in animal genetics*, Genetics Selection Evolution, Pages: 275–305, 2002

[Βλαχάβας κ.ά., 2006] Βλαχάβας Ι., Κεφαλάς Π., Βασιλειάδης Ν., Κόκκορας Φ., Σακελλαρίου Η. (2006), *Τεχνητή Νοημοσύνη* (3η έκδοση), Εκδόσεις Πανεπιστημίου Μακεδονίας

[Γεωγούλη, 2015] Γεωργούλη Α. (2015), *Τεχνητή νοημοσύνη* [Ηλεκτρ. βιβλίο], Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, Αθήνα, ανασύρθηκε στις 6 Νοεμβρίου

[Διπλάρης, 2010] Διπλάρης, Σ. (2010), *Προηγμένες τεχνικές εξόρυξης δεδομένων και γνώσης σε βάσεις βιολογικών δεδομένων*, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Θεσσαλονίκη, ανασύρθηκε στις 29 Νοεμβρίου

[Καβακιώτης, 2015] Καβακιώτης Ι., *Ανάπτυξη βιοπληροφορικών εργαλείων με εφαρμογές στη γενετική πληθυσμών*, Τμήμα Βιολογίας, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2015

[Μπάγκος, 2015] Μπάγκος, Π. (2015), *Βιοπληροφορική* [Ηλεκτρ. βιβλίο], Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, Αθήνα, ανασύρθηκε στις 25 Νοεμβρίου

[Μπουλούγαρης, 2008] Μπουλούγαρης Ι. Γεώργιος (Νοέμβριος 2008), *Τεχνικές Ενισχυτικής Μάθησης σε Πολυπρακτορικά Συστήματα* [Διπλωματική εργασία], Αθήνα

[Παπανικολάου κ.ά., 2015] Παπανικολάου, Γ., Παλαιολόγου, Δ., Κατσαρέλη, Ε., Κατσίλα, Θ., Τσαρουχά, Χ., Τζέτη, Μ., Λιλάκος, Κ., Δούκισσας, Λ. (2015), *Εργαστηριακές ασκήσεις γενετικής του ανθρώπου*. [ηλεκτρ. Βιβλ.], Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, Αθήνα

[Παρτάλας, 2009] Παρτάλας Ιωάννης (2009), *Μέθοδοι ενισχυτικής μάθησης σε συστήματα πρακτόρων*, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (ΑΠΘ)

[Ρήγας, 2014] Ρήγας, Λ. (2014), *Χρήση τεχνικών εξόρυξης γνώσης σε ιατρικά δεδομένα* [Διπλωματική εργασία], Πανεπιστήμιο Πατρών, Τμήμα Διοίκησης Επιχειρήσεων, Πάτρα, ανασύρθηκε στις 29 Νοεμβρίου

[Τζανής, 2011] Τζανής, Γ. (2011), *Ανακάλυψη Γνώσης από Βιολογικά Δεδομένα* [Διδακτορική διατριβή], Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Τμήμα Πληροφορικής, Θεσσαλονίκη, ανασύρθηκε στις 29 Νοεμβρίου

[Marsh, 2014] Marsh Jennifer, Απρίλιος 2014, *Knowledge Discovery in Databases: 9 Steps to Success*, ανασύρθηκε στις 16 Νοεμβρίου 2016 από: https://blog.udemy.com/knowledge-discovery-in-databases/

Bioinformatics organization, *Sequence Alignment*, ανασύρθηκε στις 2 Δεκεμβρίου, από: http://www.bioinformatics.org/wiki/sequence_alignment

bioplanet.com, , *What is bioinformatics*, ανασύρθηκε στις 16 Νοεμβρίου 2016 από: http://www.bioplanet.com/what-is-bioinformatics/

EMBOSS tool, ανασύρθηκε στις 10 Δεκεμβρίου από: http://emboss.sourceforge.net/what/

Genetics Home Reference, *What are single nucleotide polymorphisms (SNPs)?*, ανασύρθηκε στις 25 Μαρτίου από: https://ghr.nlm.nih.gov/primer/genomicresearch/snp

Institute for Systems Biology, *What Is Systems Biology*, ανασύρθηκε στις 2 Δεκεμβρίου από: https://www.systemsbiology.org/about/what-is-systems-biology/

National Cancer Institute, *NCI Dictionary of Cancer Terms: genetic marker*, ανασύρθηκε στις 26 Μαρτίου από:
https://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=46129

NCBI, *Bioinformatics*, ανασύρθηκε στις 28 Νοεμβρίου από: https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/bioinformatics.html

OpenStructure tool, ανασύρθηκε στις 10 Δεκεμβρίου από:
http://www.openstructure.org/

SAS(2016), *Machine Learning: What it is and why it matters*, ανασύρθηκε στις 30 Οκτωβρίου 2016 από:
http://www.sas.com/en_us/insights/analytics/machine-learning.html

Swiss Institute of Bioinformatics, *What is bioinformatics*, ανασύρθηκε στις 28 Νοεμβρίου από:
http://www.sib.swiss/bioinformatics-for-all/what-is-bioinformatics