



Πανεπιστήμιο
Ιωαννίνων

Πανεπιστήμιο Ιωαννίνων

ΣΧΟΛΗ: ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ: ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

**ΑΝΑΠΤΥΞΗ ΕΦΑΡΜΟΓΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ
ΔΕΔΟΜΕΝΩΝ ΣΕ ΚΙΝΗΤΑ ΤΗΛΕΦΩΝΑ**

ΖΟΜΠΟΛΟΣ Σ. ΔΗΜΗΤΡΙΟΣ

Επιβλέπων Καθηγητής: Ιωάννης Γ. Τσούλος

Αναπληρωτής καθηγητής

Πανεπιστημίου Ιωαννίνων



Πανεπιστήμιο
Ιωαννίνων

Πανεπιστήμιο Ιωαννίνων

ΣΧΟΛΗ: ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ: ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

**ΑΝΑΠΤΥΞΗ ΕΦΑΡΜΟΓΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ
ΔΕΔΟΜΕΝΩΝ ΣΕ ΚΙΝΗΤΑ ΤΗΛΕΦΩΝΑ**

ΖΟΜΠΟΛΟΣ Σ. ΔΗΜΗΤΡΙΟΣ

Επιβλέπων Καθηγητής: Ιωάννης Γ. Τσούλος

Αναπληρωτής καθηγητής

Πανεπιστημίου Ιωαννίνων

**APPLICATION DEVELOPMENT OF DATA CATEGORY FOR MOBILE
PHONES**

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Τόπος, Ημερομηνία

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Επιβλέπων καθηγητής

Ιωάννης Τσούλος,

Αναπληρωτής Καθηγητής,

Πανεπιστήμιο Ιωαννίνων

2. Μέλος επιτροπής

Όνομα Επίθετο,

τίτλος,

βαθμίδα

3. Μέλος επιτροπής

Όνομα Επίθετο,

τίτλος,

βαθμίδα

Ο/Η Προϊστάμενος/η του Τμήματος

Όνομα Επίθετο,

τίτλος, βαθμίδα

Υπογραφή

© Ζόμπολος Σ.Δημήτριος , 2021.

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα πτυχιακή εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Ζόμπολος Σ. Δημήτριος

Υπογραφή

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά , θα ήθελα να εκφράσω τις ευχαριστίες μου στον Αναπληρωτή Καθηγητή κ. Ιωάννη Γ.Τσούλο για την δυνατότητα που μου έδωσε να πραγματοποιήσω την πτυχιακή μου εργασία. Θα ήθελα να ευχαριστήσω ακόμα, όλους του καθηγητές του Πανεπιστημίου Ιωαννίνων της σχολής Πληροφορικής & Τηλεπικοινωνιών για τις πολύτιμες γνώσεις που μου προσέφεραν όλα αυτά τα χρόνια, αλλά και για την προθυμία τους να στηρίζουν και να βοηθήσουν οποιονδήποτε φοιτητή αναζητά την ακαδημαϊκή ή επαγγελματική του ταυτότητα. Τέλος, θέλω να εκφράσω ένα τεράστιο ευχαριστώ στην οικογένεια μου, για την στήριξη και την εμπιστοσύνη που μου έδειξε όλα αυτά τα χρόνια των σπουδών μου. Πέραν όμως από την πολύτιμη αυτή στήριξη, μου έδωσαν όλα τα εφόδια ώστε να γίνω ένας σωστός άνθρωπος.

ΠΕΡΙΛΗΨΗ

Η ανάπτυξη του διαδικτύου και της τεχνολογίας στις μέρες μας, ωθεί ολοένα και περισσότερο στη καθημερινή χρήση του διαδικτύου τον κόσμο. Η ανάπτυξη του διαδικτύου και της τεχνολογίας έχει ανοίξει το δρόμο στις επιχειρήσεις που βλέπουν μια ευκαιρία προώθησής τους και προβάλλονται όλο και συχνότερα στον κάθε χρήστη που χρησιμοποιεί το διαδίκτυο. Στόχος λοιπόν των επιχειρήσεων είναι να προβληθούν στον κάθε χρήστη. Στην εργασία αυτή, έγινε προσπάθεια να κατασκευαστεί μια εφαρμογή σε ένα κινητό που με τα στοιχεία που λαμβάνει από το διαδίκτυο για τον κάθε χρήστη που το χρησιμοποιεί να μπορεί να αναγνωρίζει με σχετικά μεγάλη ακρίβεια το φύλο του χρήστη. Μια τέτοια εφαρμογή θα μπορούσε να βοηθήσει τις εταιρίες/επιχειρήσεις για πιο ουσιαστικό και στοχοποιημένο μάρκετινγκ αφού αναγνωρίζοντας το φύλο μπορεί να προβάλει διαφημίσεις που αφορούν το φύλο.

Λέξεις κλειδιά: Προβάλλονται όλο και συχνότερα στον κάθε χρήστη που χρησιμοποιεί το διαδίκτυο, αναγνωρίζει με μεγάλη ακρίβεια το φύλο του χρήστη, ουσιαστικό και στοχοποιημένο μάρκετινγκ.

ABSTRACT

The development of the internet and technology nowadays pushes more and more into the daily use of the internet by the world. The development of the internet and technology has paved the way for businesses that see an opportunity to promote themselves and are increasingly displayed to every user who uses the internet. Therefore the goal of a business is to promote this to every user. In this work, an attempt was made to build an application on a mobile phone that with the data it receives from the internet for each user who uses it can be able to identify the user's gender with relatively high accuracy. Such an application could help companies / businesses for more meaningful and targeted marketing since recognizing gender can display ads related to gender.

Keywords: Promote themselves and are increasingly displayed to every user who uses the internet, can be able to identify the user's gender with relatively high accuracy, meaningful and targeted marketing.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΕΡΙΛΗΨΗ.....	viii
ABSTRACT.....	ix
1.Εισαγωγή	2
1.1 Μάθηση με επίβλεψη.....	4
1.2 Μάθηση χωρίς επίβλεψη.....	6
1.3 Προβλήματα μάθησης συναρτήσεων.....	9
1.4 Προβλήματα κατηγοριοποίησης.....	10
2.Τεχνητά νευρωνικά δίκτυα.....	15
2.1 Ιστορική αναδρομή.....	15
2.2 Δίκτυα ενός επιπέδου.....	17
2.3 Δίκτυα πολλών επιπέδων.....	19
2.4 Ο αλγόριθμος Back Propagation.....	21
3. Γενετικοί αλγόριθμοι.....	27
3.1 Ιστορική αναδρομή.....	27
3.2 Κωδικοποιήσεις προβλημάτων.....	30
3.3 Γενετικοί τελεστές.....	32
3.4 Γενετικός προγραμματισμός.....	37
4 Ανάπτυξη εφαρμογής κατηγοριοποίησης δεδομένων σε κινητά τηλέφωνα.....	44
4.1 Τα δεδομένα του προβλήματος.....	44
4.2 Εφαρμογή τεχνητών νευρωνικών δικτύων στα δεδομένα.....	46
4.3 Χρήση γενετικού αλγόριθμου.....	47
4.4 Ενδεικτικά αποτελέσματα.....	49

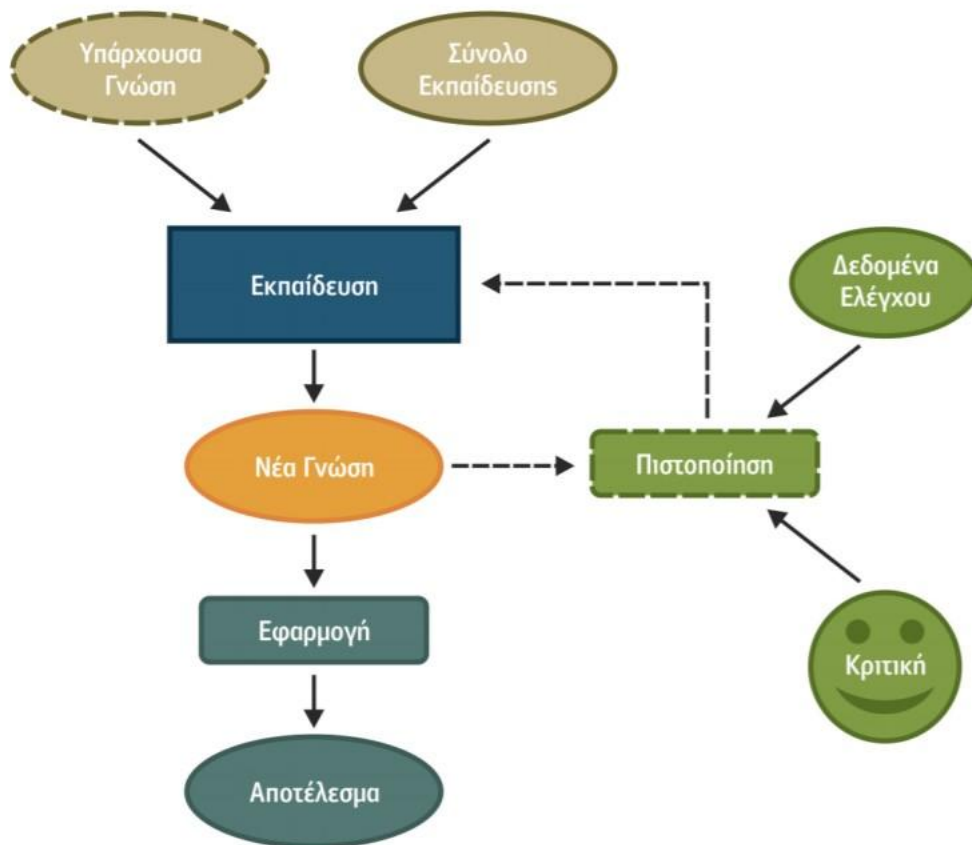
4.5 Παράδειγμα από τον κώδικα.....	52
5. Συμπεράσματα-μελλοντικές κατευθύνσεις.....	61
Βιβλιογραφία.....	63

1 Εισαγωγή.

Μηχανική μάθηση είναι υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Το 1959, ο Άρθουρ Σάμουελ ορίζει τη μηχανική μάθηση ως "Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μάθαινουν, χωρίς να έχουν ρητά προγραμματιστεί. Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μάθαινουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Η μηχανική μάθηση είναι στενά συνδεδεμένη και συχνά συγχέεται με υπολογιστική στατιστική, ένας κλάδος, που επίσης επικεντρώνεται στην πρόβλεψη μέσω της χρήσης των υπολογιστών. Έχει ισχυρούς δεσμούς με την μαθηματική βελτιστοποίηση, η οποία παρέχει μεθόδους, τη θεωρία και τομείς εφαρμογής. Η Μηχανική μάθηση εφαρμόζεται σε μια σειρά από υπολογιστικές εργασίες, όπου τόσο ο σχεδιασμός όσο και ο ρητός προγραμματισμός των αλγορίθμων είναι ανέφικτος. Παραδείγματα εφαρμογών αποτελούν τα φίλτρα spam (spam filtering), η οπτική αναγνώριση χαρακτήρων (OCR), οι μηχανές αναζήτησης και η υπολογιστική όραση. Η Μηχανική μάθηση μερικές φορές συγχέεται με την εξόρυξη δεδομένων, όπου η τελευταία επικεντρώνεται περισσότερο στην εξερευνητική ανάλυση των δεδομένων, γνωστή και ως μη επιτηρούμενη μάθηση.

Στο πεδίο της ανάλυσης δεδομένων, η μηχανική μάθηση είναι μια μέθοδος που χρησιμοποιείται για την επινόηση πολύπλοκων μοντέλων και αλγορίθμων που οδηγούν στην πρόβλεψη. Τα αναλυτικά μοντέλα επιτρέπουν στους ερευνητές, τους επιστήμονες δεδομένων, τους μηχανικούς και τους αναλυτές να παράγουν αξιόπιστες αποφάσεις και αποτελέσματα και να αναδείξουν αλληλοσυσχετίσεις μέσω της μάθησης από ιστορικές σχέσεις και τάσεις στα δεδομένα.



Εικόνα 1.1 Αποτυπώνεται ο γενικός τρόπος λειτουργίας των αλγορίθμων Μηχανικής Μάθησης. Η βασικότερη φάση κάθε αλγόριθμου είναι η εκπαίδευση, όπου χρησιμοποιεί ως είσοδο ένα σύνολο δεδομένων εκπαίδευσης προς επίτευξη του σκοπού του, τη δημιουργία νέας γνώσης. Επιπλέον, μπορεί είτε να χρησιμοποιήσει λιγότερο ή περισσότερο την υπάρχουσα γνώση είτε να μην τη χρησιμοποιήσει καθόλου. Την εκπαίδευση ακολουθεί η φάση της πιστοποίησης της παραγόμενης νέας γνώσης. Συνήθως, η πιστοποίηση πραγματοποιείται καταρχάς από τον ίδιο τον αλγόριθμο μέσω διαδικασιών ανάκλησης (recall) με τη βοήθεια δεδομένων ελέγχου (test data) και, στη συνέχεια, μέσω κριτικής που κάνει ο χρήστης βάσει των γνώσεων που διαθέτει για το πρόβλημα που επιχειρεί να λύσει ο αλγόριθμος. Τέλος, η νέα γνώση δίνεται προς χρήση σε εφαρμογές στις οποίες είναι απαραίτητη, για να λυθούν πραγματικά προβλήματα.

Είδη μηχανικής μάθησης

Έχουν αναπτυχθεί πολλές τεχνικές μάθησης που χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος και εμπίπτουν σε ένα από τα παρακάτω δύο είδη:

- Μάθηση με επίβλεψη (Supervised Learning).
- Μάθηση χωρίς επίβλεψη (Unsupervised Learning).

Στη μάθηση με επίβλεψη το σύστημα καλείται να μάθει μια έννοια ή συνάρτηση από ένα σύνολο δεδομένων, η οποία αποτελεί περιγραφή ενός μοντέλου.

Στη μάθηση χωρίς επίβλεψη το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποιά είναι.

1.1 Μάθηση με επίβλεψη

Ο όρος μάθηση με επίβλεψη αναφέρεται στη διαδικασία εξαγωγής μίας συνάρτησης ή ενός μοντέλου που χαρτογραφεί τη σχέση μεταξύ ενός συνόλου περιγραφικών γνωρισμάτων εισόδου και ενός (ή περισσότερων στην περίπτωση της πολυκατηγορικής ταξινόμησης) γνωρίσματος εξόδου. Ανάλογα με το είδος του γνωρίσματος εξόδου, η μάθηση με επίβλεψη μπορεί να κατηγοριοποιηθεί περαιτέρω ως παλινδρόμηση όταν η έξοδος είναι συνεχής, ή κατηγοριοποίηση δεδομένων στην περίπτωση διακριτών εξόδων. Γενικά, ένας αλγόριθμος μάθησης με επίβλεψη στοχεύει στη δημιουργία ενός μοντέλου που προβλέπει την τιμή εξόδου για κάθε έγκυρο αντικείμενο εισόδου, γενικεύοντας από τα δεδομένα εκπαίδευσης σε άγνωστες καταστάσεις, με κάποιον “λογικό” τρόπο.

Σε αυτήν την κατηγορία Μηχανικής Μάθησης το σύστημα πρέπει να “μάθει” με επαγωγική μέθοδο μια συνάρτηση που ονομάζεται “συνάρτηση στόχος” (target function) και που αποτελεί έκφραση του μοντέλου που περιγράφει τα δεδομένα. Η

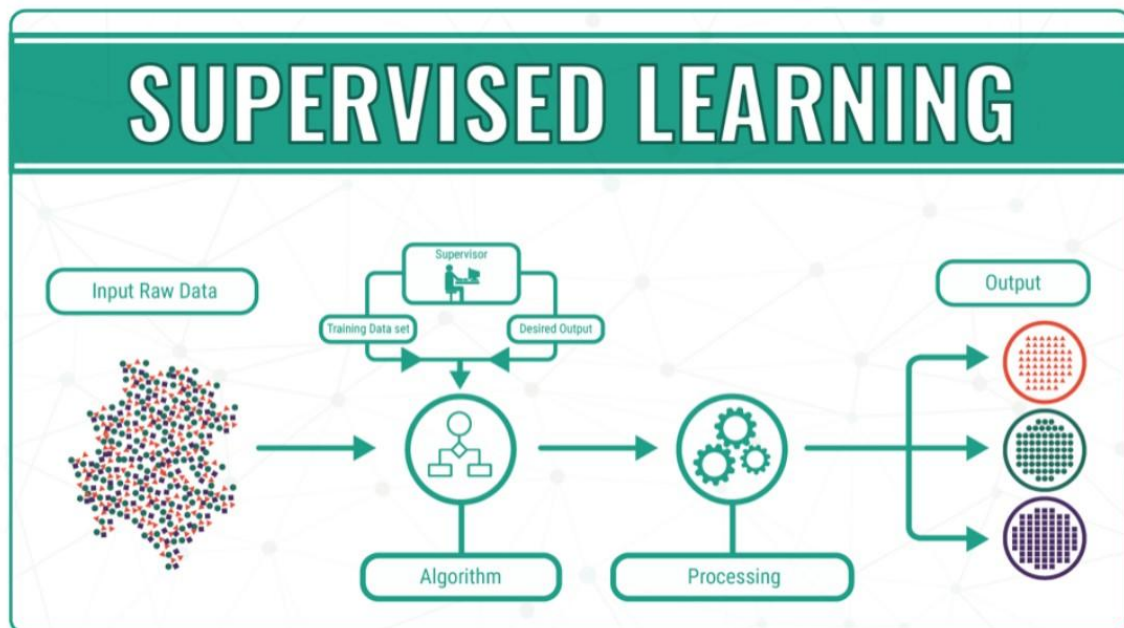
συνάρτηση στόχος χρησιμοποιείται για την πρόβλεψη της τιμής μιας μεταβλητής, που ονομάζεται εξαρτημένη μεταβλητή ή μεταβλητή εξόδου.

Μάθηση με επίβλεψη είναι μία κατηγορία μηχανικής μάθησης, στόχος της οποίας είναι ο χαρακτηρισμός δεδομένων με βάση κάποια δεδομένα εκπαίδευσης. Τα δεδομένα εκπαίδευσης αποτελούνται από ένα σύνολο παραδειγμάτων τα οποία χρησιμοποιούνται για εκπαίδευση μοντέλων. Στην επιβλεπόμενη μάθηση, κάθε παράδειγμα αποτελείται από ένα σύνολο εισόδου (συνήθως ένα διάνυσμα από χαρακτηριστικά) και μια επιθυμητή τιμή εξόδου. Οι Αλγόριθμοι επιβλεπόμενης μάθησης αναλύουν τα δεδομένα εκπαίδευσης και παράγουν ένα μοντέλο το οποίο μπορεί να χρησιμοποιηθεί για να χαρακτηρίσει νέα παραδείγματα. Το βέλτιστο σενάριο επιτρέπει στον αλγόριθμο να καθορίσει σωστά την ετικέτα της κατηγορίας για άγνωστα μέχρι τώρα παραδείγματα. Για να επιτευχθεί αυτό, απαιτείται ο αλγόριθμος μάθησης να γενικεύει από τα δεδομένα εκπαίδευσης σε αθέατες καταστάσεις με ένα "λογικό" τρόπο.

Στη Μάθηση με Επίβλεψη εντάσσονται δύο κατηγορίες προβλημάτων (learning tasks) εξόρυξης γνώσης: Τα προβλήματα ταξινόμησης και τα προβλήματα παρεμβολής. Η ταξινόμηση (classification) αφορά στη δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων, όπως για παράδειγμα η ομάδα αίματος, ενώ η παρεμβολή (regression) αφορά στη δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών.

Οι κυριότερες τεχνικές Μηχανικής Μάθησης με επίβλεψη είναι:

- Μάθηση Εννοιών (Concept Learning)
- Δέντρα Απόφασης (Decision Trees)
- Μάθηση Κανόνων (Rule Learning)
- Μάθηση κατά Περίπτωση (Instance Based Learning)
- Μάθηση κατά Bayes
- Γραμμική Παρεμβολή (Linear Regression)
- Νευρωνικά Δίκτυα (Neural Networks)
- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)



Εικόνα 1.1.1 Η Μάθηση με επίβλεψη βασίζεται σε μια διαδικασία επισημανσης προκειμένου να κατανοήσει τις πληροφορίες. Το σύστημα μαθαίνει από την επισημάνση πολλών δεδομένων. Αυτό είναι ωφέλιμο μόνο όταν γνωρίζουμε ακριβώς τι ψάχνουμε.

Ένα σύστημα Μηχανικής Μάθησης απαιτεί μία γλώσσα αναπαράστασης του κόσμου του προβλήματος, δηλαδή των υποθέσεων. Επιπροσθέτως, απαιτείται ένα σύνολο από τελεστές (operators) οι οποίοι θα επιτρέψουν στο σύστημα να οδηγηθεί σε μία γενίκευση (επαγωγή), έναν ευριστικό κανόνα ή ένα πλάνο που ικανοποιεί τους στόχους του.

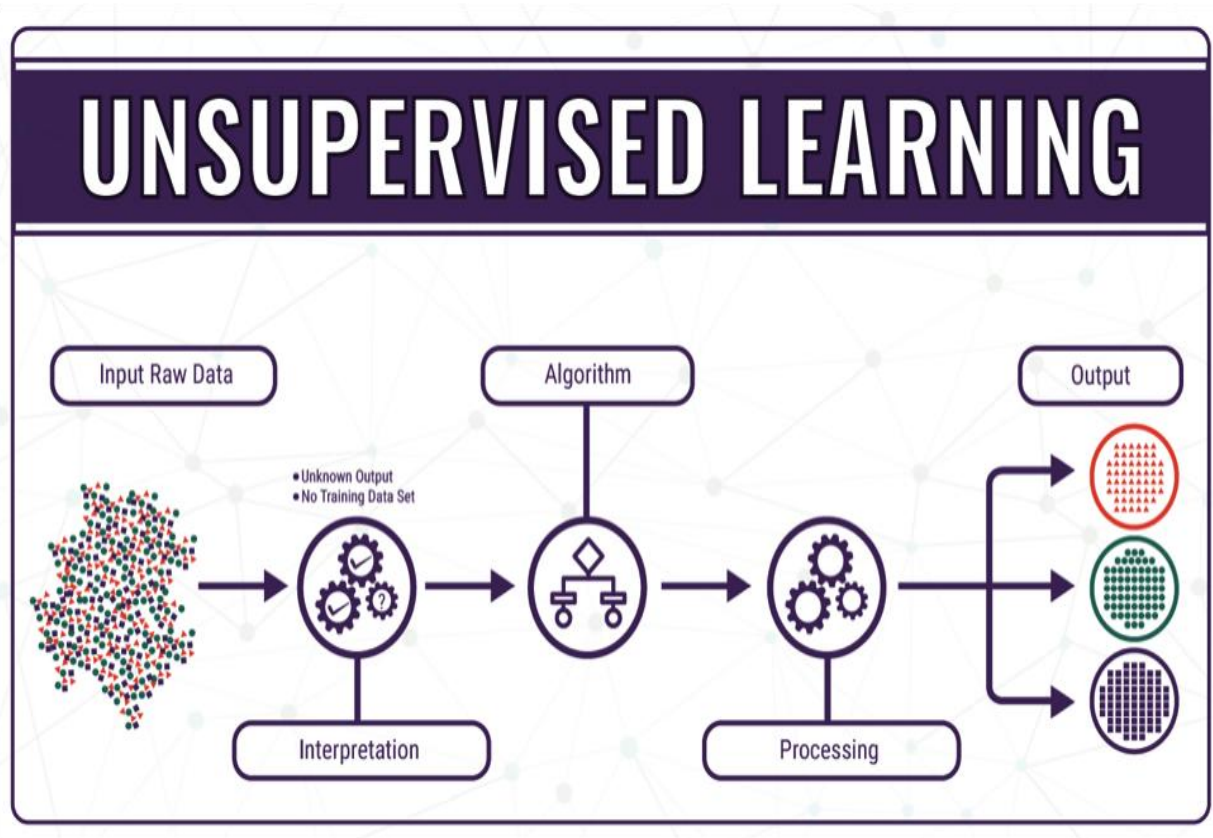
1.2 Μάθηση χωρίς επίβλεψη

Η μάθηση χωρίς επίβλεψη μελετά πώς ένα σύστημα μπορεί να μάθει να αναπαριστά συγκεκριμένα παραδείγματα εισόδου, με τρόπο που να αντανakλά τη στατιστική δομή της συνολικής συλλογής στην οποία αυτά περιέχονται. Σε αντίθεση με την μάθηση με επίβλεψη, δεν υπάρχουν σαφείς έξοδοι (ή

περιβαλλοντικές αξιολογήσεις) συνδεδεμένες με κάθε είσοδο. Οι αλγόριθμοι μάθησης χωρίς επίβλεψη καλούνται να επιλέξουν ποιες πτυχές της δομής της εισόδου πρέπει να συμπεριληφθούν στην έξοδο. Οι προκύπτουσες αναπαραστάσεις γνώσης μπορούν, μεταξύ άλλων, να χρησιμοποιηθούν για λήψη αποφάσεων, πρόβλεψη μελλοντικών εισόδων, ομαδοποίηση παρόμοιων εισόδων ή δημιουργία πρωτοτύπων τα οποία προωθούνται σε άλλες τεχνικές . Οι βασικές προσεγγίσεις μάθησης χωρίς επίβλεψη είναι η ομαδοποίηση (clustering) και ο διαχωρισμός σημάτων με χρήση τεχνικών εξαγωγής γνωρισμάτων και στόχο τη μείωση της διαστατικότητας .

Μάθηση χωρίς επίβλεψη όπου ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Χρησιμοποιείται σε προβλήματα:

- Ανάλυσης Συσχετίσεων
- Ομαδοποίησης



Εικόνα 1.2.1 Η Μάθηση χωρίς επίβλεψη αντλεί συμπεράσματα από σύνολα δεδομένων, χωρίς να ξέρει τι ακριβώς ψάχνουμε. Στη μάθηση χωρίς επίβλεψη το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι.

Σε αυτήν την κατηγορία Μηχανικής Μάθησης το Σύστημα έχει ως στόχο να ανακαλύψει συσχετίσεις και ομάδες από τη Βάση Γνώσης που του παρέχεται, βασιζόμενο αποκλειστικά και μόνο στις ιδιότητές τους. Σαν αποτέλεσμα (έξοδος) προκύπτουν πρότυπα (περιγραφές), κάθε ένα από τα οποία περιγράφει ένα μέρος από τα δεδομένα. Χαρακτηριστικά παραδείγματα προτύπων πληροφόρησης στη Μάθηση χωρίς επίβλεψη είναι (α) οι Κανόνες Συσχέτισης (association rules) και (β) οι Ομάδες (clusters), οι οποίες προκύπτουν από τη διαδικασία της ομαδοποίησης (clustering).

Στη μάθηση χωρίς επίβλεψη το σύστημα έχει στόχο να ανακαλύψει συσχετίσεις και ομάδες από τα δεδομένα, βασιζόμενο μόνο στις ιδιότητές τους. Σαν

αποτέλεσμα προκύπτουν πρότυπα (περιγραφές), κάθε ένα από τα οποία περιγράφει ένα μέρος από τα δεδομένα.

Παραδείγματα προτύπων πληροφόρησης είναι

- οι κανόνες συσχέτισης (association rules).
- οι ομάδες (clusters), οι οποίες προκύπτουν από τη διαδικασία της ομαδοποίησης (clustering).

1.3 Προβλήματα μάθησης συναρτήσεων.

Μεταξύ της επιτηρούμενης και της μη επιτηρούμενης μάθησης είναι η ημι-επιτηρούμενη μάθηση, όπου ο δάσκαλος δίνει ένα ελλειπές εκπαιδευτικό σήμα: ένα σύνολο εκπαίδευσης με κάποια (συχνά πολλά) από τα αποτελέσματα στόχους να λείπουν. Η μεταγωγή είναι μια ειδική περίπτωση της αρχής αυτής, όπου το σύνολο των καταστάσεων του προβλήματος είναι γνωστό κατά το χρόνο εκμάθησης, όμως ένα μέρος των στόχων λείπουν.

Μεταξύ άλλων κατηγοριών μηχανικής μάθησης, υπάρχει ακόμα διαδικασία εκμάθησης (meta learning) που μαθαίνει στην μηχανή (να αναπτύσσει) τις δικές της επαγωγικές μεθόδους, βασιζόμενο στην προηγούμενη εμπειρία.

Η Αναπτυξιακή μάθηση (Developmental robotics), η οποία έχει αναπτυχθεί για την εκμάθηση από ρομπότ, δημιουργεί τη δική της ακολουθία μαθησιακών καταστάσεων, ώστε το ρομποτ συσσωρευτικά αποκτά ποικιλία δεξιοτήτων μέσω της αυτόνομης αυτοεξερεύνησης και της κοινωνικής αλληλεπίδρασης με ανθρώπους εκπαιδευτές και χρησιμοποιώντας μηχανισμούς καθοδήγησης, όπως η ενεργητική μάθηση, η ωρίμανση και η μίμηση. Μια άλλη κατηγοριοποίηση των προβλημάτων μηχανικής μάθησης προκύπτει όταν κάποιος θεωρήσει το επιθυμητό αποτέλεσμα του συστήματος μηχανικής μάθησης.

- Στην ταξινόμηση, τα δεδομένα εισόδου χωρίζονται σε δύο ή περισσότερες κλάσεις, και η μηχανή πρέπει να κατασκευάσει ένα μοντέλο, το οποίο θα αντιστοιχίζει τα δεδομένα σε μία ή περισσότερες (multi-label ταξινόμηση) κλάσεις. Αυτό συνήθως εμπίπτει στην επιτηρούμενη μάθηση. Τα φίλτρα Spam είναι ένα παράδειγμα ταξινόμησης, όπου οι εισοδοί είναι τα emails

ή άλλα μηνύματα και οι κλάσεις είναι "spam" και "όχι spam". Τα προβλήματα ταξινόμησης ή κατηγοριοποίησης (classification), που αφορούν στη δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων/κατηγοριών), ονομαστικών (nominal) ή βαθμωτών (ordinal) τιμών, όπως για παράδειγμα η ομάδα αίματος ή η πιστοληπτική ικανότητα ενός πελάτη τράπεζας.

- Στην παλινδρόμηση, επίσης πρόβλημα επιτηρούμενης μάθησης, τα αποτελέσματα είναι συνεχή και όχι διακριτά. Τα προβλήματα παλινδρόμησης ή παρεμβολής (regression), αφορούν στη δημιουργία μοντέλων πρόβλεψης συνεχών αριθμητικών τιμών, όπως για παράδειγμα η τιμή της θερμοκρασίας σε πρόγνωση καιρού ή μιας μετοχής.
- Στην συσταδοποίηση, ένα σύνολο εισόδων πρόκειται να χωριστεί σε ομάδες. Σε αντίθεση με την ταξινόμηση, οι ομάδες δεν είναι γνωστές εκ των προτέρων, καθιστώντας αυτόν τον διαχωρισμό τυπική εργασία μη επιτηρούμενης μάθησης.
- Στην εκτίμηση πυκνότητας βρίσκει την κατανομή των δεδομένων εισόδου σε κάποιο χώρο.
- Σε προβλήματα μείωσης διαστασιμότητας (dimensionality reduction), τα δεδομένα απλοποιούνται και αντιστοιχίζονται σε ένα χώρο λιγότερων διαστάσεων. Το στατιστικό μοντέλο θεμάτων (Topic modeling) είναι ένα σχετικό πρόβλημα, όπου η μηχανή καλείται να βρει έγγραφα που καλύπτουν παρόμοια θέματα από ένα σύνολο εγγράφων γραμμένων σε φυσική γλώσσα.

1.4 Προβλήματα κατηγοριοποίησης.

Η κατηγοριοποίηση (classification) είναι μία τεχνική της εξόρυξης δεδομένων, κατά την οποία ένα στοιχείο ανατίθεται σε ένα προκαθορισμένο σύνολο κατηγοριών. Ο όρος κατηγοριοποίηση συναντάται στην βιβλιογραφία και ως ταξινόμηση. Γενικότερα, ο στόχος της διαδικασίας αυτής είναι η ανάπτυξη ενός μοντέλου, το οποίο αργότερα θα μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δεδομένων. Τέτοια παραδείγματα είναι ο διαχωρισμός των emails με βάση την επικεφαλίδα τους ή το περιεχόμενό τους, η

πρόβλεψη καρκινικών κυττάρων χαρακτηρίζοντας τα ως καλοήθη ή κακοήθη, η κατηγοριοποίηση πελατών μιας τράπεζας ανάλογα με την πιστωτική τους ικανότητα κ.α.

Η κατηγοριοποίηση μπορεί να περιγραφεί ως μία διαδικασία δύο βημάτων:

- Εκμάθηση(Learning):Στο πρώτο βήμα της διαδικασίας δημιουργείται/προσδιορίζεται το μοντέλο με βάση ένα σύνολο προκατηγοριοποιημένων παραδειγμάτων, που ονομάζεται δεδομένα εκπαίδευσης(training data).Τα δεδομένα εκπαίδευσης αναλύονται από ένα αλγόριθμο κατηγοριοποίησης, προκειμένου να σχηματιστεί το μοντέλο. Λόγω του ότι τα δεδομένα εκπαίδευσης ανήκουν σε μία προκαθορισμένη κατηγορία, η οποία είναι γνωστή, η κατηγοριοποίηση αποτελεί μέθοδο εποπτευομένης μάθησης(supervised learning). Το μοντέλο, που λέγεται και αλλιώς κατηγοριοποιητής(classifier), αναπαρίσταται με τη μορφή κανόνων κατηγοριοποίησης(classification rules), δέντρων απόφασης(decision trees) ή μαθηματικών τύπων.
- Κατηγοριοποίηση(Classification): Μετά την δημιουργία του μοντέλου, το επόμενο βήμα είναι η αξιολόγησή του. Για να επιτευχθεί αυτό, χρησιμοποιούμε τα δοκιμαστικά δεδομένα(test data) για να υπολογίσουν την ακρίβεια του μοντέλου. Το μοντέλο κατηγοριοποιεί τα δοκιμαστικά δεδομένα. Έπειτα, η κατηγορία που σχηματίστηκε με βάση τα δοκιμαστικά δεδομένα συγκρίνεται με την πρόβλεψη που έγινε για τα δεδομένα εκπαίδευσης, τα οποία είναι ανεξάρτητα από αυτά της δοκιμής. Η ακρίβεια του μοντέλου υπολογίζεται από το ποσοστό των δειγμάτων δοκιμής που κατηγοριοποιήθηκαν σωστά σε σχέση με το υπό εκπαίδευση μοντέλο.

Στην περίπτωση που το μοντέλο κριθεί αποδεκτό, τότε μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δειγμάτων δεδομένων, των οποίων η κατηγοριοποίηση είναι άγνωστη.

Κατηγορίες μεθόδων κατηγοριοποίησης.

- Η Bayesian κατηγοριοποίηση αποτελεί μία κατηγορία μεθόδων της κατηγοριοποίησης και βασίζεται στη στατιστική θεωρία κατηγοριοποίησης του Bayes. Αυτό σημαίνει ότι πραγματοποιείται μια πιθανοτική πρόβλεψη, δηλαδή προβλέπει την πιθανότητα ένα δείγμα X να ανήκει σε κάποια κατηγορία. Ο απλούστερος Bayesian κατηγοριοποιητής είναι ο Naïve Bayesian. Αυτός υποθέτει ότι η επίδραση ενός γνωρίσματος σε μία κατηγορία είναι ανεξάρτητη από τις τιμές των υπόλοιπων γνωρισμάτων. Ο λόγος που γίνεται αυτό είναι για να αποφεύγονται οι πολύπλοκοι υπολογισμοί κατά τη συνθήκη ανεξαρτησίας της κατηγορίας. Βασικό τους πλεονέκτημα είναι ότι όλες οι ποσότητες μπορούν να υπολογιστούν με χρήση ερωτημάτων προς τη βάση δεδομένων. Μειονέκτημα τους είναι ότι δεν μπορεί να εντοπίσει πρότυπα που βασίζονται σε αλληλεπίδραση δύο ή περισσότερων χαρακτηριστικών, διότι βασίζεται στην ακριβώς αντίθετη παραδοχή.
- Δέντρα απόφασης. Τα δέντρα απόφασης χρησιμοποιούνται ευρέως για την κατηγοριοποίηση και πρόβλεψη δεδομένων. Ένα δέντρο απόφασης κατασκευάζεται σύμφωνα με ένα σύνολο εκπαίδευσης προκατηγοριοποιημένων δεδομένων. Κάθε εσωτερικός κόμβος προσδιορίζει τον έλεγχο των γνωρισμάτων και κάθε κλαδί που συνδέει τους εσωτερικούς με τους απόγονους αντιστοιχεί σε μία πιθανή τιμή για το γνώρισμα, όπως εμφανίζεται και στην διπλανή εικόνα. Βασικά τους πλεονεκτήματα είναι η ευκολία με την οποία ερμηνεύονται και η πληθώρα αλγορίθμων κατασκευής τους.
- Νευρωνικά δίκτυα Μια άλλη τεχνική κατηγοριοποίησης που χρησιμοποιείται σε εφαρμογές εξόρυξης γνώσης για πρόβλεψη και κατηγοριοποίηση στηρίζεται στα νευρωνικά δίκτυα. Τα βήματα αυτής της διαδικασίας χονδρικά είναι:

1) Η αναγνώριση των χαρακτηριστικών εισόδου και εξόδου

2) Δημιουργία ενός δικτύου με την κατάλληλη τοπολογία

3)Επιλογή του συνόλου εκπαίδευσης(train data)

4)Εφαρμογή του δικτύου με ένα αντιπροσωπευτικό σύνολο δεδομένων, ώστε να μεγιστοποιείται η δυνατότητα του δικτύου να αναγνωρίζει τα πρότυπα

5)Επαλήθευση-αξιολόγηση του δικτύου με την χρήση ενός σύνολο ελέγχου(test data).

- Παραγωγή κανόνων κατηγοριοποίησης Η γνώση που αποκτούμε κατά την διαδικασία της κατηγοριοποίησης μπορεί να αναπαρασταθεί και με τη χρήση κανόνων. Οι κανόνες κατηγοριοποίησης, σε σχέση με τα δέντρα απόφασης, γίνονται ευκολότερα κατανοητοί όταν το δέντρο που παράχθηκε είναι μεγάλο. Έτσι μπορούμε να μετατρέψουμε ένα δέντρο απόφασης σε ένα σύνολο κανόνων κατηγοριοποίησης. Αυτό μπορεί να επιτευχθεί εάν θεωρήσουμε ότι κάθε κανόνας αντιστοιχεί σε ένα μονοπάτι του δέντρου από τη ρίζα μέχρι ένα κόμβο φύλλο. Άρα κάθε φύλλο παράγει ένα κανόνα. Οι συνθήκες που θα μας οδηγήσουν στο φύλλο(υπόθεση) αποτελούν το αριστερό μέρος του κανόνα, ενώ το φύλλο(αποτέλεσμα) αντιστοιχεί στο δεξιό μέρος του κανόνα.

Τα σημαντικότερα προβλήματα κατηγοριοποίησης είναι τα παρακάτω:

- Περιορισμένη πληροφορία.Οι βάσεις δεδομένων δεν είναι πάντα σχεδιασμένες για αναζήτηση γνώσης και συχνά οι ιδιότητες και τα πεδία που θα απλοποιούσαν τη διαδικασία αναζήτησης όχι μόνο λείπουν αλλά δεν είναι δυνατόν να συλλεχθούν από το χρήστη.
- Δείγματα.Η χρήση δείγματος είναι σχεδόν πάντα επιβεβλημένη καθώς οι πραγματικές βάσεις δεδομένων έχουν τεράστιο αριθμό εγγραφών ενώ πολλοί αλγόριθμοι απαιτούν τα δεδομένα εκπαίδευσης να είναι όλα στη μνήμη.Η λήψη ενός δείγματος απαιτεί μεγάλη προσοχή και εφαρμογή στατιστικών τεχνικών,ώστε να αντιπροσωπεύει ικανοποιητικά την αρχική βάση.
- Θόρυβος και πεδία χωρίς τιμή.Οι βάσεις δεδομένων συχνά περιέχουν λάθη.Πεδία των οποίων η τιμή προέρχεται από υποκείμενικές κρίσεις ή

μετρήσεις μπορεί να προκαλέσουν λάθη,όπως εσφαλμένη κατηγοριοποίηση ορισμένων (παρα)δειγμάτων.

- Μέγεθος βάσης δεδομένων.Κάνει χρονοβόρα τη συγκέντρωση στατιστικών στοιχείων για τον έλεγχο της ποιότητας της γνώσης που προκύπτει.Η ύπαρξη μεγάλου αριθμού πεδίων ανά εγγραφή,αν και θεωρείται πλεονέκτημα για την ακρίβεια του παραγόμενου μοντέλου,δημιουργεί τεράστιο αριθμό εναλλακτικών κανόνων κατηγοριοποίησης.
- Αραιά δεδομένα.Ένα σύστημα δεν είναι σε θέση να προσδιορίσει επακριβώς τα όρια μιας κατηγορίας,αν στη βάση δεδομένων υπάρχουν μόνο εγγραφές που σίγουρα ανήκουν ή σίγουρα δεν ανήκουν στην κατηγορία.
- Πρόσφατα δεδομένα.Κατά πόσο μπορεί να θεωρηθεί ότι οι κανόνες που κάποτε παρήχθησαν ανταποκρίνονται στην πλέον ενημερωμένη και πρόσφατη έκδοση της βάσης δεδομένων.

2 Τεχνητά Νευρωνικά Δίκτυα

2.1 Ιστορική αναδρομή

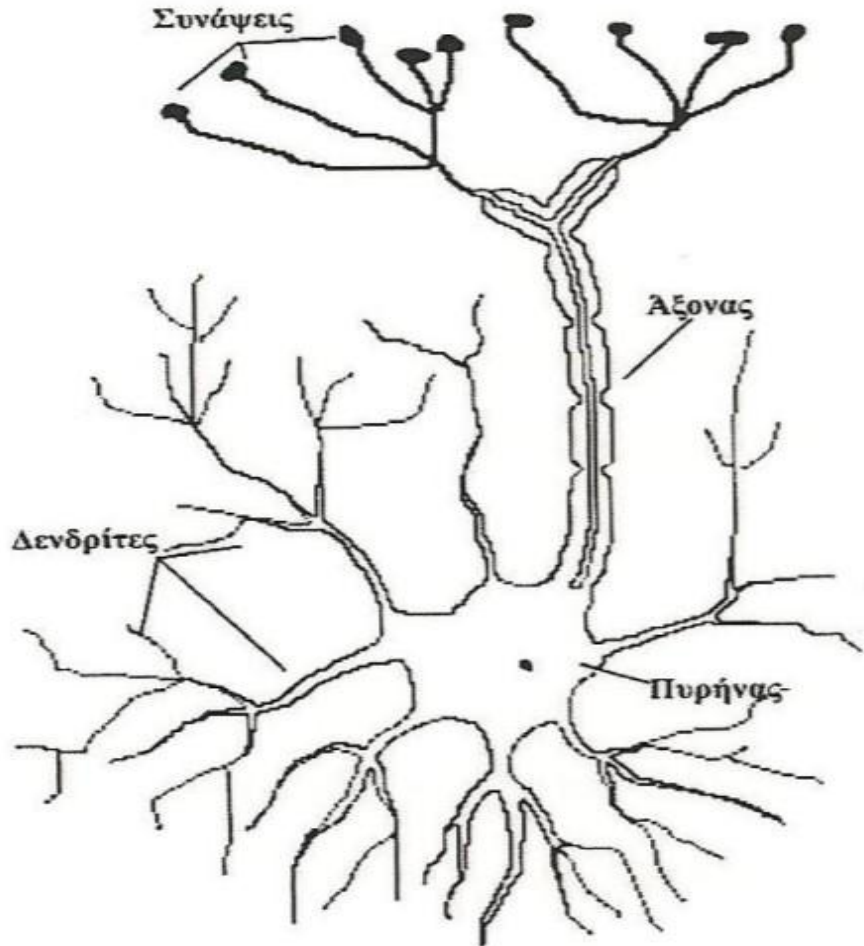
Η έρευνα σχετικά με τα τεχνητά νευρωνικά δίκτυα είναι εμπνευσμένη από τη δομή και τη λειτουργία του εγκεφάλου. Βασικό δομικό στοιχείο του εγκεφάλου είναι οι νευρώνες, δηλαδή τα νευρικά κύτταρα τα οποία δημιουργούν ένα πυκνό δίκτυο επικοινωνίας μεταξύ τους. Κίνητρο για τη μελέτη του νευρώνα και των νευρωνικών δικτύων είναι η ελπίδα ανακάλυψης ενός νέου υπολογιστικού μοντέλου βασισμένου σε μια δικτυακή δομή παρόμοια με αυτή του εγκεφάλου. Αυτή η κανούργια υπολογιστική πλατφόρμα θα είναι πιο κατάλληλη για ανάπτυξη ευφυών αλγορίθμων και γενικότερα διαδικασιών σχετιζόμενων με τη νοημοσύνη, όπως η μάθηση, η μνήμη, η γενίκευση, η ομαδοποίηση προτύπων, κλπ.

Τα συνήθη Τεχνητά Νευρωνικά Δίκτυα χρησιμοποιούν πολύ απλοποιημένα μοντέλα νευρώνων τέτοια ώστε να διατηρούν μόνο τα πολύ αδρά χαρακτηριστικά των λεπτομερών μοντέλων που χρησιμοποιούνται στη νευρολογία. Θα έλεγε κανείς ότι τα συνήθη τεχνητά νευρωνικά μοντέλα έχουν ελάχιστη σχέση με τα βιολογικά νευρωνικά συστήματα. Ωστόσο πιστεύεται ότι οι λεπτομέρειες δεν έχουν ιδιαίτερη σημασία στην κατανόηση της ευφυούς συμπεριφοράς των βιολογικών νευρωνικών συστημάτων. Ακόμη και αυτά τα απλά μοντέλα νευρώνων μπορούν να δημιουργήσουν ιδιαίτερος ενδιαφέροντα δίκτυα αρκεί να πληρούν δύο βασικά χαρακτηριστικά:

- Οι νευρώνες να έχουν ρυθμιζόμενες παραμέτρους ώστε να διευκολύνεται η διαδικασία της μάθησης-ιδιότητα γνωστή ως πλαστικότητα των νευρώνων.
- Το δίκτυο να αποτελείται από μεγάλο πλήθος νευρών ώστε να επιτυγχάνεται παραλληλισμός της επεξεργασίας και κατανομή της πληροφορίας.

Η πρόκληση που αντιμετωπίζει η θεωρία των τεχνητών νευρωνικών δικτύων είναι η εύρεση κατάλληλων αλγορίθμων εκπαίδευσης των δικτύων και ανάκλησης της πληροφορίας που αυτά περιέχουν έτσι ώστε να προσομοιάζονται ευφυείς διαδικασίες όπως αυτές που αναφέρθηκαν προηγουμένως. Για την επίτευξη αυτού του στόχου απαιτείται ο ορισμός του κατάλληλου περιβάλλοντος εκπαίδευσης, πχ. αν το δίκτυο θα εκπαιδεύεται με επίβλεψη δηλαδή με τη χρήση κάποιων

δεδομένων οδηγών-δασκάλων, ή αν το δίκτυο θα αφήνεται μόνο του να αυτο-οργανωθεί κα με ποιό συγκεκριμένο κριτήριο και στόχο



Εικόνα 2.1.1 Ο Νευρώνας

Το νευρωνικό κύτταρο ή νευρώνας είναι το βασικό δομικό στοιχείο του εγκεφάλου τόσο στον άνθρωπο όσο και στα ζώα. Ο νευρώνας είναι ένα μεγάλο σε μέγεθος κύτταρο το οποίο, ανατομικά, αποτελείται από τα εξής τμήματα :

- Οι δενδρίτες είναι οι πύλες εισόδου του νευρώνα.
- Ο άξονας είναι η πύλη εξόδου του νευρώνα.
- Οι συνάψεις είναι τα σημεία ένωσης μεταξύ διακλαδώσεων του άξονα ενός νευρώνα και των δενδριτών από άλλους νευρώνες.

Η λειτουργία του νευρώνα γίνεται ως εξής: φορείς πληροφορίας είναι ηλεκτρικοί παλμοί που ταξιδεύουν στον άξονα κάθε νευρώνα και μέσω των συνάψεων διαδίδονται στους δενδρίτες των παραληπτών νευρώνων. Κάθε νευρώνας Α συλλέγει όλο το ηλεκτρικό φορτίο που δέχεται από κάθε σύναψη στους δενδρίτες του ζυγίζοντας το εισερχόμενο φορτίο με το αντίστοιχο συναπτικό βάρος. Έτσι, όσο πιο ισχυρή είναι η συναπτική ζεύξη τόσο πιο πολύ έντονα συμμετέχει το συγκεκριμένο φορτίο εισόδου στο συνολικό άθροισμα. Αν το άθροισμα του φορτίου ξεπερνάει κάποιο κατώφλι τότε ο άξονας του Α αρχίζει να παράγει ηλεκτρικούς παλμούς με μεγάλη συχνότητα οπότε λέμε ότι ο νευρώνας πυροβολεί (fires). Αν όμως το φορτίο δεν περνάει το συγκεκριμένο αυτό όριο τότε λέμε ότι ο νευρώνας είναι αδρανής.

Πλεονεκτήματα νευρωνικών δικτύων.

- η ακρίβεια πρόβλεψης είναι γενικά υψηλή,
- δουλεύει όταν τα παραδείγματα του συνόλου εκπαίδευσης περιέχουν λάθη,
- η έξοδος μπορεί να είναι διακριτή, συνεχών τιμών, ή ένα διάνυσμα διαφόρων, διακριτών ή συνεχών τιμών χαρακτηριστικών,
- γρήγορη αξιολόγηση της μαθημένης συνάρτησης στόχου.

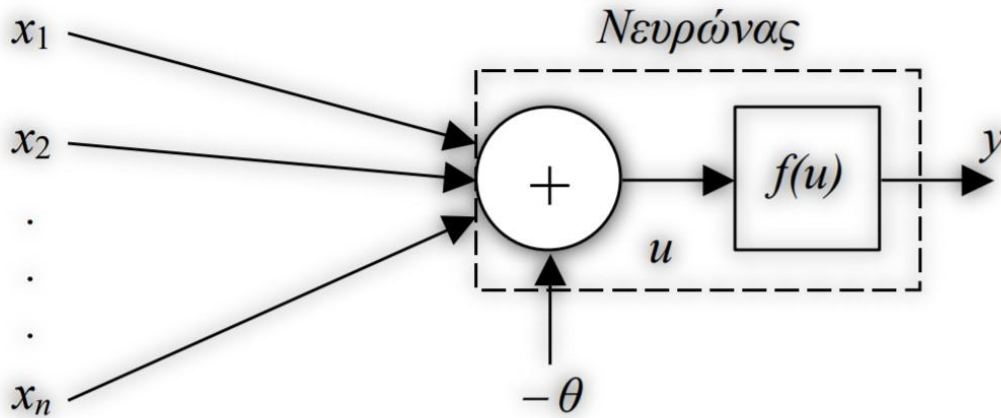
Προβλήματα νευρωνικών δικτύων.

- μεγάλος χρόνος εκπαίδευσης,
- απαιτούν παραμέτρους (που συνήθως καθορίζονται εμπειρικά) (π.χ. η τοπολογία του δικτύου),
- δύσκολη η κατανόηση της μαθημένης συνάρτησης (βάρη),
- δεν είναι εύκολη η ενσωμάτωση της γνώσης του εκάστοτε γνωστικού πεδίου.

2.2 Δίκτυα ενός επιπέδου

Το πιο απλό δίκτυο που μπορεί να σχεδιαστεί και να μελετηθεί είναι ένα δίκτυο που αποτελείται από ένα μόνο νευρώνα. Η λέξη «δίκτυο» σε αυτή την περίπτωση χρησιμοποιείται καταχρηστικά αφού δεν υπάρχουν περισσότεροι του ενός νευρώνες για να συνδεθούν μεταξύ τους. Οι μόνες συνδέσεις που υπάρχουν είναι

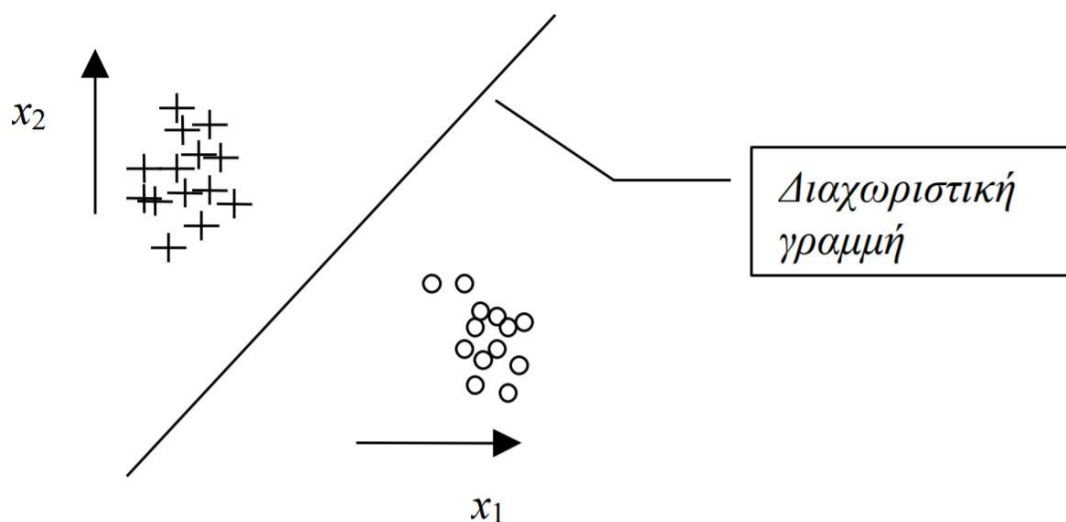
αυτές μεταξύ των εισόδων και του νευρώνα όπως φαίνεται και στο παρακάτω σχήμα.



Εικόνα 2.2.1 Τεχνητό νευρωνικό δίκτυο ενός επιπέδου(Perceptron).

Στο μοντέλο Perceptron ο ένας και μοναδικός νευρώνας υλοποιεί τη συνάρτηση μεταφοράς. Η συνάρτηση μεταφοράς απεικονίζει το διάνυσμα εισόδου x στην έξοδο y . Οι παράμετροι w_0, w_1, w_2 είναι τα συνοπτικά βάρη του νευρώνα ενώ η παράμετρος θ λέγεται κατώφλι ενεργοποίησης. Ο όρος αυτός εξηγείται από το γεγονός ότι η διέγερση u του νευρώνα η οποία προκύπτει από το περιβάλλον (εισόδους) είναι θετική αν το άθροισμα ξεπεράσει το όριο θ . Με το σκεπτικό αυτό λέμε ότι το θ είναι το κατώφλι πάνω από το οποίο ενεργοποιείται ο νευρώνας. Η συνάρτηση ενεργοποίησης $F()$ τροφοδοτείται από τη διέγερση u και δίνει την έξοδο y του νευρώνα. Η συνάρτηση ενεργοποίησης είναι μη γραμμική και ειδικά στο Perceptron παίρνει μια από τις εξής μορφές $(1/0)$ ή $(-1/1)$. Οι παράμετροι που ρυθμίζουν τη συμπεριφορά του νευρώνα είναι το διάνυσμα των συναπτικών βαρών w .

Το ζητούμενο σε ένα δίκτυο Perceptron είναι η αυτόματη εκμάθηση των παραμέτρων του συστήματος ώστε να επιτυγχάνεται ο επιθυμητός στόχος που είναι η εύρεση της διαχωριστικής γραμμής όπως φαίνεται στο παρακάτω σχήμα.



Εικόνα 2.1.2 Πρόβλημα γραμμικά διαχωρίσιμο.

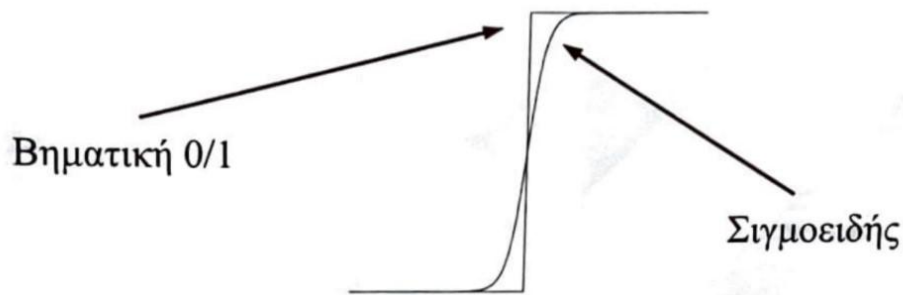
Αν το πρόβλημα δεν είναι γραμμικά διαχωρίσιμο τότε ο αλγόριθμος Perceptron δεν συγκλίνει ποτέ. Αυτό είναι ένα σοβαρό μειονέκτημα το οποίο αποτελεί και βασικό σημείο κριτικής ενάντιον του. Ακόμη πολλά προβλήματα στον πραγματικό κόσμο είναι μη γραμμικά διαχωρίσιμα και επομένως το μοντέλο είναι άχρηστο γι' αυτά.

2.3 Δίκτυα πολλών επιπέδων

Μία από τις χρησιμότερες και πλέον επιτυχημένες εφαρμογές των νευρωνικών δικτύων στην ανάλυση δεδομένων και στην πρόβλεψη αποτελεί το μοντέλο πολλαπλών επιπέδων perceptron (multilayer perceptron model) MLP. Τα μοντέλα πολλαπλών επιπέδων perceptron είναι μη γραμμικά μοντέλα νευρωνικών δικτύων που μπορούν να χρησιμοποιηθούν σχεδόν με όλες τις συναρτήσεις με υψηλό βαθμό ακρίβειας. Τα MLP περιέχουν ένα ή περισσότερα κρυμμένα επίπεδα νευρώνων που χρησιμοποιούν μη γραμμικές συναρτήσεις ενεργοποίησης. Ένα MLP αποτελείται συνήθως από πολλά επίπεδα κόμβων. Το πρώτο (χαμηλότερο), είναι το επίπεδο εισόδου όπου λαμβάνεται η εξωτερική πληροφορία. Το τελευταίο (υψηλότερο), είναι το επίπεδο εξόδου όπου επιτυγχάνεται η λύση του προβλήματος. Ενδιάμεσα υπάρχουν τα κρυμμένα επίπεδα. Οι κόμβοι στα γειτονικά επίπεδα συνήθως είναι πλήρως συνδεδεμένοι μέσω τόξων από χαμηλότερα επίπεδα προς υψηλότερα. Τα δίκτυα perceptron πολλών στρωμάτων MLP όπου οι νευρώνες χρησιμοποιούν τη βηματική συνάρτηση 0/1 ή -1/1, μπορούν να υλοποιήσουν συναρτήσεις που δεν

μπορεί να υλοποιήσει ένα απλό δίκτυο perceptron.Ωστόσο, η χρήση της βηματικής συνάρτησης δεν προτιμάται.Ο λόγος είναι ότι

οι περισσότεροι κανόνες εκπαίδευσης βασίζονται σε μεθόδους βελτιστοποίησης που χρησιμοποιούν παραγώγους,ενώ η βηματική συνάρτηση δεν είναι παραγωγίσιμη.Αυτή η δυσκολία ξεπερνιέται με τη χρήση της σιγμοειδούς συναρτήσεως.



Εικόνα 2.3.1 Βηματική και σιγμοειδής συνάρτηση.

Η χρήση των «μαλακών» συναρτήσεων κατωφλίωσης όπως η σιγμοειδής συνάρτηση δημιουργεί ομαλές επιφάνειες χωρίς απότομες μεταβολές στη τιμή y της εξόδου του δικτύου.Στη περίπτωση της βηματικής συνάρτησης δεν υπάρχουν ενδιάμεσες τιμές ενώ στη περίπτωση της σιγμοειδούς,η μετάβαση γίνεται ομαλά χωρίς απότομες ακμές.Τα δίκτυα perceptron πολλαπλών στρωμάτων που ενσωματώνουν τη σιγμοειδή συνάρτηση αποδεικνύεται ότι έχουν πολλές δυνατότητες αναπαράστασης συναρτήσεων.Το βασικό θεώρημα ουσιαστικά λείπει ότι δίκτυα αυτής της μορφής μπορούν να προσεγγίσουν οποιαδήποτε ομαλή συνάρτηση, όσο κοντά επιθυμούμε.Το βασικό χαρακτηριστικό σε ένα δίκτυο perceptron πολλών στρωμάτων είναι ότι κάθε στρώμα τροφοδοτείται από τους νευρώνες του προηγούμενου στρώματος και μόνον από αυτούς.Για λόγους ευκολίας καλούμε το στρώμα εισόδου,καταχρηστικά ,μηδενικό στρώμα,οπότε ο κανόνας γενικεύεται και για το πρώτο στρώμα,δηλαδή το πρώτο στρώμα τροφοδοτείται και αυτό αποκλειστικά από το μηδενικό στρώμα ,ήτοι από το στρώμα εισόδου.Η εκπαίδευση ενός δικτύου πολλών στρωμάτων είναι η διαδικασία ρύθμισης των συναπτικών βαρών του έτσι ώστε να ικανοποιείται κάποιο κριτήριο καταλληλότητας.Αυτός άλλωστε είναι και ο στόχος της εκπαίδευσης σε οποιοδήποτε νευρωνικό δίκτυο.Αυτό που κάνει την εκπαίδευση

ενός δικτύου MLP πιο ενδιαφέρουσα είναι η ιδιότητα του καθολικού προσεγγιστή που περιγράψαμε παραπάνω, δηλαδή αν έχουμε το κατάλληλο σε μέγεθος δίκτυο μπορούμε να το εκπαιδεύσουμε να μάθει οποιαδήποτε συνάρτηση εμείς επιθυμούμε με οποιαδήποτε ποιότητα προσέγγισης εμείς επιθυμούμε. Σε αντίθεση με το απλό δίκτυο perceptron μπορεί να υλοποιήσει μόνο γραμμικές διαχωριστικές επιφάνειες.

2.4 Ο αλγόριθμος Back-Propagation

Ο αλγόριθμος back-propagation προτάθηκε από τον Paul Werbos τη δεκαετία του 1970 στα πλαίσια της ανάλυσης μοντέλων οικονομικής και πολιτικής πρόβλεψης. Τότε, τα μοντέλα αυτά δεν είχαν σχεδιαστεί για να έχουν σχέση με νευρωνικά δίκτυα. Αργότερα τη δεκαετία του 1980, έγινε αντιληπτό ότι η μέθοδος μπορούσε να μεταφερθεί αυτούσια στην εκπαίδευση νευρωνικών δικτύων πολλών στρωμάτων, και από τότε έγινε η πιο γνωστή και η πιο διαδεδομένη μέθοδος γι' αυτό το σκοπό. Βασικό χαρακτηριστικό της μεθόδου αυτής είναι η ύπαρξη στόχων, όπως και ακριβώς και στο απλό δίκτυο perceptron. Συνεπώς το μοντέλο ανήκει στην κατηγορία των δικτύων που εκπαιδεύονται με επίβλεψη.

Για πολλά χρόνια δεν υπήρχε τρόπος εκπαίδευσης δικτύων με πολλά επίπεδα και οι γνώσεις μας περιορίζονταν. Έτσι, έγινε γρήγορα αντιληπτό ότι ήταν απαραίτητο να βρούμε ένα τρόπο να εκπαιδεύσουμε νευρωνικά δίκτυα που μπορούν να αναπαραστήσουν πιο περίπλοκες διεργασίες. Το κενό αυτό ήρθε να καλύψει η μέθοδος εκπαίδευσης Back-Propagation. Το νευρωνικό δίκτυο στο οποίο εφαρμόζεται είναι αρκετά πιο περίπλοκο από το μοντέλο του απλού Perceptron. Είναι ένα δίκτυο πολλαπλών επιπέδων και κάθε επίπεδο μπορεί να έχει πολλούς νευρώνες. Οι νευρώνες μέσα στο ίδιο επίπεδο δεν συνδέονται μεταξύ τους, αλλά οι νευρώνες που ανήκουν σε διαφορετικά επίπεδα συνδέονται ως συνήθως με τις γνωστές συνάψεις. Η κεντρική ιδέα της δομής και λειτουργίας τέτοιων δικτύων είναι σχετικά απλή. Το δίκτυο ξεκινά την διαδικασία μάθησης από τυχαίες τιμές των βαρών του. Εάν δώσει λάθος απάντηση, τότε τα βάρη διορθώνονται έτσι ώστε το σφάλμα να γίνει μικρότερο. Η ίδια διαδικασία επαναλαμβάνεται πολλές φορές έτσι ώστε σταδιακά το σφάλμα ελαττώνεται μέχρις ότου γίνει πολύ μικρό και ανεκτό. Στο σημείο αυτό λέμε ότι το δίκτυο έχει μάθει τα

παραδείγματα που του διδάξαμε με την ακρίβεια που θέλαμε να μάθει. Η διαδικασία εκπαίδευσης κατά τον αλγόριθμο Back-Propagation λαμβάνει χώρα με την παρουσίαση και εφαρμογή στο νευρωνικό δίκτυο ενός συνόλου παραδειγμάτων εκπαίδευσης. Η παρουσίαση όλων των προτύπων εκπαίδευσης μια φορά το καθένα στο δίκτυο ονομάζεται εποχή. Κατά τη συνολική διαδικασία εκπαίδευσης εκτελούνται επαναλήψεις των εποχών, ώσπου τα βάρη του δικτύου να σταθεροποιηθούν σε συγκεκριμένες τιμές που θα προκαλούν σύγκλιση της μέσης τιμής των σφαλμάτων, για όλα τα πρότυπα εκπαίδευσης, στην ελάχιστη δυνατή τιμή της.

Ο αλγόριθμος back-propagation είναι μια οικογένεια μεθόδων που χρησιμοποιούνται για την αποτελεσματική εκπαίδευση των τεχνητών νευρωνικών δικτύων (συντομογραφία ΤΝΔ), ακολουθώντας έναν αλγόριθμο βελτιστοποίησης βασισμένο στην κλίση που εκμεταλλεύεται τον κανόνα της αλυσίδας. Το κύριο χαρακτηριστικό του back-propagation είναι η επαναληπτική, αναδρομική και αποδοτική μέθοδος για τον υπολογισμό των ανανεώσεων των βαρών για τη βελτίωση του δικτύου έως ότου είναι σε θέση να εκτελέσει το έργο για το οποίο εκπαιδεύεται. Είναι στενά συνδεδεμένη με τον αλγόριθμο Gauss-Newton.

Ο back-propagation απαιτεί ότι οι παράγωγοι των συναρτήσεων ενεργοποίησης κατά το σχεδιασμό του δικτύου είναι γνωστοί. Η αυτόματη διαφοροποίηση είναι μια τεχνική που παρέχει αυτόματα και αναλυτικά τις παραγώγους των συναρτήσεων στον αλγόριθμο εκπαίδευσης. Στο πλαίσιο της μάθησης, ο αλγόριθμος back-propagation χρησιμοποιείται συνήθως από τον αλγόριθμο βελτιστοποίησης μείωσης κλίσης ώστε να ρυθμίσει το βάρος των νευρώνων υπολογίζοντας την κλίση της συνάρτησης απώλειας. Ο back-propagation υπολογίζει τις κλίσεις, ενώ η (στοχαστική) μείωση της κλίσης χρησιμοποιεί τις κλίσεις για την εκπαίδευση του μοντέλου (μέσω βελτιστοποίησης).

Ο αλγόριθμος back-propagation λειτουργεί ως εξής, παρακάτω σας παραθέτω τα βασικά του σημεία.

- Κατάβαση δυναμικού

Τα συνοπτικά βάρη w είναι οι παράμετροι που πρέπει να διορθωθούν ώστε να ελαχιστοποιηθεί το σφάλμα j , καθώς τόσο οι εισόδους x όσο και οι στόχοι d είναι δεδομένοι και σταθεροί. Σύμφωνα με τη μέθοδο κατάβασης

δυναμικού, η μεταβολή της παραμέτρου w ως προς το χρόνο t γίνεται χρησιμοποιώντας την παράγωγο του j ως προς w , η οποία ονομάζεται και κλίση του j ως προς w .

- Εξίσωση και ενημέρωση βαρών και υπολογισμός με τον κανόνα δέλτα.

Εξίσωση ενημέρωσης των βαρών:

$$w_{ij}(l, k+1) = w_{ij}(l, k) + \beta \delta_i^{(k)}(l) a_j^{(k)}(l-1)$$

$$j = 0, 1, \dots, N(l), \quad l = 1, \dots, L$$

Τύποι υπολογισμού δ :

Στρώμα $L =$ τελευταίο στρώμα:

$$\delta_i^{(k)}(L) = f'(u_i^{(k)}(L))(d_i^{(k)} - y_i^{(k)})$$

Στρώμα $l = 1, \dots, L-1$:

$$\delta_i^{(k)}(l) = f'(u_i^{(k)}(l)) \sum_{\mu=1}^{N(l+1)} w_{\mu i}(l+1) \delta_{\mu}(l+1)$$

Εικόνα 2.4.1 Εξίσωση και ενημέρωση βαρών και υπολογισμός με τον κανόνα δέλτα.

- Παράμετρος του βήματος εκπαίδευσης β .

Όπου το βήμα εκπαίδευσης β είναι περίπου ίσο με το ελάχιστο βήμα χρόνου. Αυτό σημαίνει ότι για να προσεγγίσουμε την πραγματική κατάβαση δυναμικού το β πρέπει να είναι πολύ μικρό. Αν όντως επιλέξουμε β πολύ μικρό τότε ο αλγόριθμος θα συγκλίνει στο κοντινότερο τοπικό ελάχιστο, όπως ακριβώς θα συνέβαινε με την πραγματική κατάβαση δυναμικού, αλλά η σύγκλιση μας θα είναι πολύ αργή. Μπορούμε να επιτύχουμε ταχύτερη σύγκλιση αν επιλέξουμε μεγαλύτερο βήμα εκπαίδευσης, Ωστόσο, από κάποιο σημείο και μετά, ο κανόνας του Back-Propagation θα αποτελεί φτωχή προσέγγιση της κατάβασης δυναμικού, με αποτέλεσμα να έχουμε κακή σύγκλιση ή, στη χειρότερη περίπτωση ακόμη και απόκλιση του αλγορίθμου (τιμές των βαρών θα οδηγούνται στο $+\infty$). Η σωστή επιλογή του β δεν είναι εύκολο να προσδιοριστεί μαθηματικά εκ των προτέρων και συνήθως

χρησιμοποιείται η μέθοδος δοκιμής και σφάλματος.Τυπικά χρησιμοποιούνται τιμές μικρότερες έως πολύ μικρότερες της μονάδας.Σε γενικές γραμμές, αν το δίκτυο έχει πολλές εισόδους ή οι εισόδοι έχουν μεγάλες τιμές τότε καλό είναι το βήμα εκπαίδευσης να έχει μικρή τιμή.Αντίθετα σε δίκτυα με λίγες εισόδους ή εισόδους με μικρές τιμές, το βήμα εκπαίδευσης μπορεί να είναι μεγαλύτερο.Αν και η κατάβαση δυναμικού χρησιμοποιεί το ίδιο βήμα εκπαίδευσης για όλα τα βάρη δικτύου,στη πράξη έχει παρατηρηθεί ότι μπορεί να επιτευχθεί ταχύτερη σύγκλιση αν χρησιμοποιούνται διαφορετικά βήματα εκπαίδευσης για κάθε στρώμα δικτύου ή ακόμα και για κάθε συναπτικό βάρος ξεχωριστά.Συνήθως τα τελευταία στρώματα έχουν μεγαλύτερη παράγωγο από τα πρώτα στρώματα.Έτσι,αν χρησιμοποιούμε το ίδιο βήμα εκπαίδευσης παντού,η διόρθωση των βαρών είναι μεγαλύτερη στα τελευταία στρώματα απ'ότι στα πρώτα.Για να εξισορροπηθεί η ταχύτητα της εκπαίδευσης σε όλα τα στρώματα μπορούμε να χρησιμοποιήσουμε μεγαλύτερο βήμα εκπαίδευσης στα πρώτα στρώματα και μικρότερο στα μεγαλύτερα.

- Κριτήριο τερματισμού

Για να αποφασίσουμε αν ο αλγόριθμος back-propagation πρέπει να τερματιστεί ή όχι,μπορούμε γενικά να χρησιμοποιήσουμε δύο μεγέθη το κόστος ή τα συνοπτικά βάρη.Ίσως το πιο συχνά χρησιμοποιούμενο μέγεθος είναι το κόστος με βάση το οποίο μπορεί κανείς να ορίσει δύο λογικά κριτήρια τερματισμού:

A)Το κόστος για κάποια εποχή εκπαίδευσης είναι μικρότερο απο κάποιο κατώφλι που ορίζει ο χρήστης.

B)Το σφάλμα σε δύο διαδοχικές εποχές $n-1, n$,δεν μειώνεται σημαντικά:

$$J(n-1)-J(n)<\varepsilon$$

Όπου ε είναι κάποιο προκαθορισμένο κατώφλι που ορίζει ο χρήστης

Εναλλακτικά ,κανείς μπορεί να χρησιμοποιήσει τα συναπτικά βάρη για να ορίσει κριτήρια τερματισμού.

- Αρχικοποίηση των βαρών

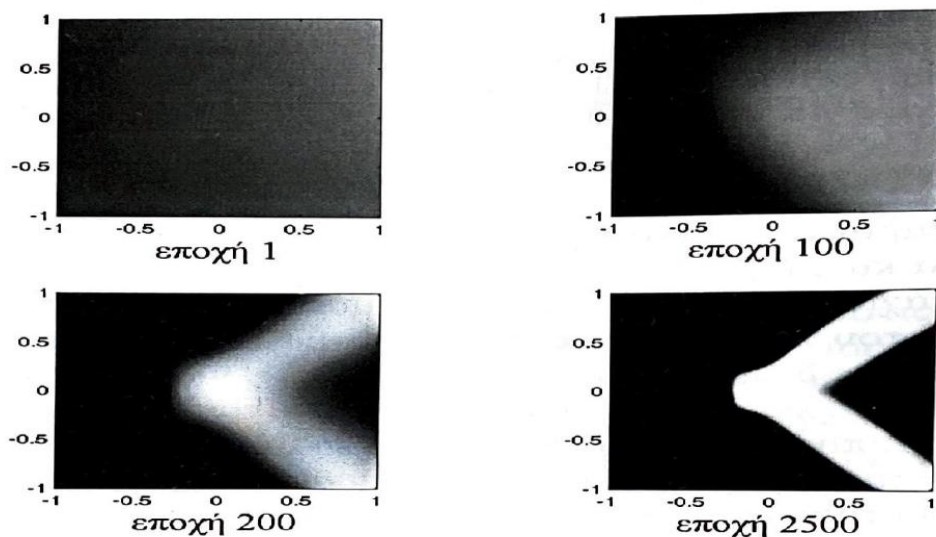
Τα βάρη κατά την εκκίνηση είναι καλό να αρχικοποιούνται σε μικρές τιμές.Ο λόγος είναι ότι τα μεγάλα βάρη οδηγούν στο λεγόμενο «κορεσμό»των νευρώνων.Ο κορεσμός είναι ένα φαινόμενο κατά το οποίο η μεταβολή των

συναπτικών βαρών είναι πολύ μικρή όχι γιατί το δίκτυο έχει φτάσει σε καλό σημείο αλλά διότι η έξοδος των νευρώνων είναι πολύ κοντά στο όριο της συνάρτησης ενεργοποίησης.

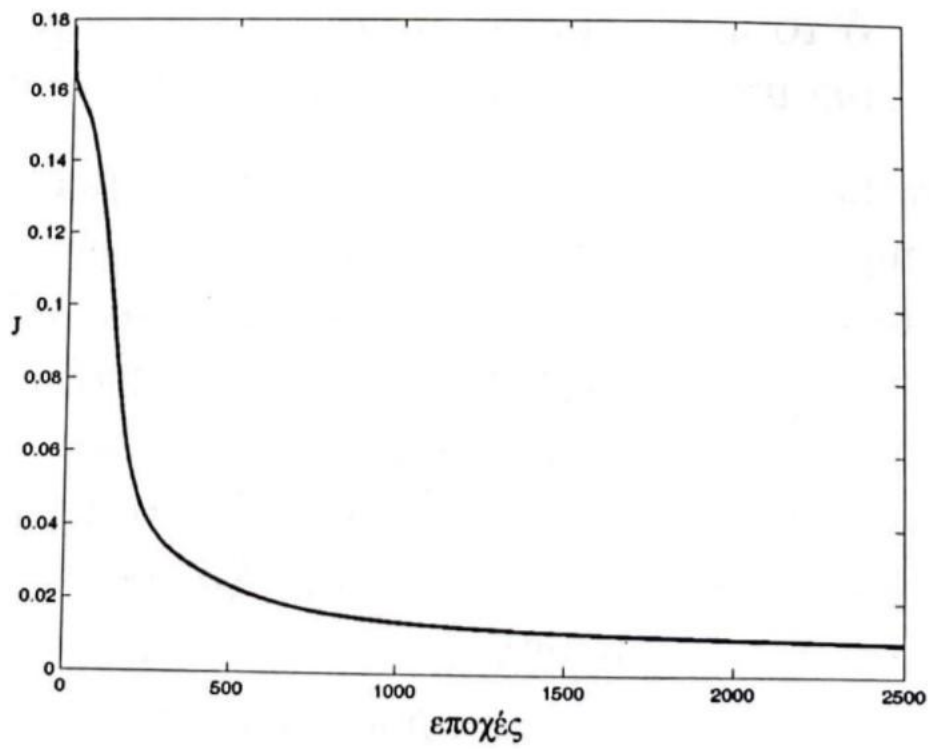
- Τοπικά ελάχιστα

Ένα βασικό χαρακτηριστικό της μεθόδου της κατάβασης δυναμικού είναι η πιθανότητα να καταλήξουμε σε τοπικά ελάχιστα της συνάρτησης κόστους. Τα τοπικά ελάχιστα είναι εν γένει ανεπιθύμητα καθώς αποτελούν υποδεέστερες λύσεις του προβλήματος σε σχέση με το ολικό ελάχιστο. Ωστόσο αυτή η πράξη έχει δείξει ότι αν ο αλγόριθμος δεν σταματήσει από τη αρχή σε κάποιο τοπικό ελάχιστο και αν εξελιχθεί κανονικά για ένα ικανό χρονικό διάστημα, το τοπικό ελάχιστο που θα προκύψει στο τέλος σπανίως δημιουργεί ουσιαστικό πρόβλημα καθώς τις περισσότερες φορές αντιστοιχεί σε πολύ ικανοποιητική τιμή κόστους.

Το δίκτυο MLP εξοπλισμένο με τον κανόνα εκπαίδευσης Back-Propagation μπορεί να προσεγγίσει οποιαδήποτε συνεχή συνάρτηση. Κατά συνέπεια, μπορεί να διαχωρίσει οποιαδήποτε κλάση K σε N διαστάσεις.



Εικόνα 2.4.2 Παράδειγμα εφαρμογής δικτύου MLP με Back-Propagation (γραμμικά μη διαχωρίσιμο)



Εικόνα 2.4.3 Εξέλιξη του σφάλματος με τη πάροδο των εποχών στο παράδειγμα
Παράδειγμα εφαρμογής δικτύου MLP με Back-Propagation(γραμμικά μη
διαχωρίσιμο).

3 ΓΕΝΕΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ

3.1 Ιστορική αναδρομή

Σε αρκετές περιπτώσεις το μέγεθος ενός προβλήματος καθιστά απαγορευτική τη χρήση κλασικών μεθόδων αναζήτησης για την επίλυση του. Στις περιπτώσεις αυτές βρίσκουν εφαρμογή πιθανοκρατικοί αλγόριθμοι οι οποίοι αν και δεν εγγυώνται ότι θα βρουν τη βέλτιστη λύση, είναι ικανοί να επιστρέψουν μια αρκετά καλή λύση σε εύλογο χρονικό διάστημα. Μια κατηγορία τέτοιων αλγορίθμων επίλυσης προβλημάτων είναι οι γενετικοί αλγόριθμοι, των οποίων ο βασικός μηχανισμός είναι εμπνευσμένος από τη Δαρβινική θεωρία της εξέλιξης της φύσης. Οι γενετικοί αλγόριθμοι εκτελούν μία αναζήτηση στο χώρο των υποψήφιων λύσεων με στόχο την εύρεση αποδεκτών λύσεων, σύμφωνα με κάποιο κριτήριο λύσεων.

Οι Γενετικοί αλγόριθμοι ανήκουν στο κλάδο της επιστήμης υπολογιστών και αποτελούν μια μέθοδο αναζήτησης βέλτιστων λύσεων σε συστήματα που μπορούν να περιγραφούν ως μαθηματικό πρόβλημα. Είναι χρήσιμοι σε προβλήματα που περιέχουν πολλές παραμέτρους/διαστάσεις και δεν υπάρχει αναλυτική μέθοδος που να μπορεί να βρει το βέλτιστο συνδυασμό τιμών για τις μεταβλητές ώστε το υπό εξέταση σύστημα να αντιδρά με όσο το δυνατόν με το επιθυμητό τρόπο.

Ο τρόπος λειτουργίας των Γενετικών Αλγορίθμων είναι εμπνευσμένος από τη βιολογία. Χρησιμοποιεί την ιδέα της εξέλιξης μέσω γενετικής μετάλλαξης, φυσικής επιλογής και διασταύρωσης. Οι Γενετικοί Αλγόριθμοι είναι αρκετά απλοί στην υλοποίησή τους. Οι τιμές για τις παραμέτρους του συστήματος πρέπει να κωδικοποιούνται με τρόπο ώστε να αναπαρασταθούν από μια μεταβλητή που περιέχει σειρά χαρακτήρων ή δυαδικών ψηφίων (0/1). Αυτή η μεταβλητή μιμείται το γενετικό κώδικα που υπάρχει στους ζωντανούς οργανισμούς. Αρχικά, ο Γενετικός Αλγόριθμος παράγει πολλαπλά αντίγραφα της μεταβλητής/γεννητικού κώδικα, συνήθως με τυχαίες τιμές, δημιουργώντας ένα πληθυσμό λύσεων. Κάθε λύση (τιμές για τις παραμέτρους του συστήματος) δοκιμάζεται για το πόσο κοντά φέρνει την αντίδραση του συστήματος στην επιθυμητή, μέσω μιας συνάρτησης που δίνει το μέτρο ικανότητας της λύσης και η οποία ονομάζεται συνάρτηση ικανότητας (Σ.Ι).

Οι λύσεις που βρίσκονται πιο κοντά στην επιθυμητή, σε σχέση με τις άλλες, σύμφωνα με το μέτρο που μας δίνει η Σ.Ι, αναπαράγονται στην επόμενη γενιά λύσεων και λαμβάνουν μια τυχαία μετάλλαξη. Επαναλαμβάνοντας αυτή τη διαδικασία για αρκετές γενιές, οι τυχαίες μεταλλάξεις σε συνδυασμό με την επιβίωση και αναπαραγωγή των γονιδίων/λύσεων που πλησιάζουν καλύτερα το επιθυμητό αποτέλεσμα θα παράγουν ένα γονίδιο/λύση που θα περιέχει τις τιμές για τις παραμέτρους που ικανοποιούν όσο καλύτερα γίνεται την Σ.Ι.

Υπάρχουν διάφορες εκδοχές της παραπάνω διαδικασίας για τους Γ.Α από τις οποίες κάποιες περιλαμβάνουν και τη διασταύρωση (ζευγάρωμα) γονιδίων/λύσεων ώστε ο αλγόριθμος να φτάσει στο αποτέλεσμα πιο γρήγορα. Καθώς υπάρχει το στοχαστικό (τυχαίο) συστατικό της μετάλλαξης και ζευγαρώματος, κάθε εκτέλεση του Γ.Α μπορεί να συγκλίνει σε διαφορετική λύση και σε διαφορετικό χρόνο. Η απόδοση του Γ.Α εξαρτάται επί το πλείστον από την συνάρτηση ικανότητας και συγκεκριμένα από το κατά πόσο το μέτρο της περιγράφει την βέλτιστη λύση. Οι γενετικοί αλγόριθμοι είναι ένα πεπερασμένο σύνολο οδηγιών για την εκπλήρωση ενός έργου, το οποίο δεδομένης μιας αρχικής κατάστασης θα οδηγήσει σε μια αναγνωρίσιμη τελική κατάσταση, και το οποίο προσπαθεί να μιμηθεί την διαδικασία της βιολογικής εξέλιξης. Οι γενετικοί αλγόριθμοι προσπαθούν να βρουν τη λύση ενός προβλήματος με το να προσομοιώνουν την εξέλιξη ενός πληθυσμού «λύσεων» του προβλήματος.

Είναι μια τεχνική προγραμματισμού που εισήγαγε στα τέλη της δεκαετίας του 1960 ο Τζον Χόλαντ, ερευνητής του Ινστιτούτου της Σάντα Φε (ΗΠΑ).

Οι γενετικοί αλγόριθμοι είναι μια από τις βάσεις των Προγραμμάτων Τεχνητής Ζωής. Συγκεκριμένα, επιχειρεί να αναπαράγει στους υπολογιστές τους μηχανισμούς της βιολογικής εξέλιξης με τον ίδιο τρόπο που η τεχνητή νοημοσύνη επιχειρεί να αναπαραστήσει και να μιμηθεί τις διαδικασίες της γνώσης.

Τα προγράμματα εξελίσσονται μέχρι να φτάσουν, μέσω μεταλλάξεων, διασταυρώσεων και φυσικής επιλογής, σε μια αποτελεσματική φόρμουλα η οποία θα εκτελεί με τον καλύτερο δυνατό τρόπο μια συγκεκριμένη εργασία.

Οι παράμετροι ενός γενετικού αλγορίθμου είναι οι εξής:

- Η κωδικοποίηση, δηλαδή ο τρόπος με τον οποίο κάθε χρωμόσωμα του πληθυσμού θα εμπεριέχει πληροφορία για τη λύση που περιγράφει.

- Η διασταύρωση,δηλαδή το πως εμπλέκονται μεταξύ τους τα χρωμοσώματα του πληθυσμού για να παράγουν απογόνους.
- Η μετάλλαξη, η οποία έχει σκοπό να εμποδίσει το αλγόριθμο από το να παγιδευτεί σε τοπικά ακρότατα του προβλήματος σε κάθε επανάληψη του.
- Η πιθανότητα διασταύρωσης και μετάλλαξης,πόσο συχνά θα εμφανίζονται δηλαδή αυτές οι δύο λειτουργίες.
- Το μέγεθος του πληθυσμού πόσα χρωμοσώματα αποτελούν τον πληθυσμό.Αν υπάρχουν λίγα χρωμοσώματα τότε ο αλγόριθμος έχει λίγες πιθανότητες να εφαρμόσει διασταύρωση και έτσι μόνο ένα μικρό κομμάτι του πεδίου αναζήτησης εξερευνείται.Αντίθετα αν υπάρχουν πιο πολλά χρωμοσώματα ο αλγόριθμος γίνεται πιο αποδοτικός.

Μερικά από τα πλεονεκτήματα καθώς και εφαρμογές των γενετικών αλγορίθμων παρουσιάζονται παρακάτω:

Πλεονεκτήματα:

- Γρήγοροι και αξιόπιστοι.
- Συνεργάζονται εύκολα με πολλά συστήματα.
- Επεκτείνονται και προσαρμόζονται εύκολα.
- Εφαρμόζονται σε περισσότερα πεδία από άλλους αλγόριθμους.
- Είναι ανθεκτικοί σε σφάλματα.
- Παραλληλίζονται εύκολα.

Εφαρμογές:

- Γραφικά υπολογιστών.
- Σύνθεση μουσικής.
- Κατασκευή τεχνητών νευρωνικών δικτύων.
- Ρομποτική.
- Αεροναυπηγική.

- Επεξεργασία φυσικής γλώσσας.
- Επεξεργασία σήματος.

3.2 Κωδικοποιήσεις προβλημάτων

Ένα από τα πιο σημαντικά ζητήματα κατά τη συγγραφή ενός γενετικού αλγορίθμου είναι η αναπαράσταση του προβλήματος με την μορφή χρωμοσωμάτων. Κατά την αναπαράσταση του προβλήματος με την μορφή χρωμοσωμάτων θα πρέπει να τηρούνται δύο βασικές ιδιότητες.

- Κατά την αναπαράσταση να μην χανεται πληροφορία των δεδομένων.
- Να είναι ξεκάθαρη η διαφορά μεταξύ στα «καλα» και «κακα» χρωμοσώματα.

Πάρακατω παρουσιάζονται μερικοί τρόποι αναπαράστασης των δεδομένων με τη μορφή χρωμοσωμάτων και τα βασικά τους πλεονεκτήματα.

- Δυαδική κωδικοποίηση.

Είναι η πρώτη μορφή κωδικοποίησης που εφαρμόστηκε στους γενετικούς αλγόριθμους και είναι η μόνη που έχουν αποδειχθεί και κριτήρια σύγκλισης των γενετικών αλγορίθμων. Τα χρωμοσώματα είναι ένας ακέραιος αριθμός και τα γονίδια τους είναι δυαδικά στοιχεία(0,1).

Chromosome A	101100101100101011100101
Chromosome B	111111100000110000011111

Εικόνα 3.2.1 Δυαδική κωδικοποίηση χρωμοσώματος.

Το μεγάλο μειονέκτημα της δυαδικής αναζήτησης είναι ότι για μια μικρή αλλαγή σε ένα χρωμόσωμα μπορεί να προκαλέσει μεγάλες αλλαγές στην επίδοση του χρωμοσώματος.

- Κωδικοποίηση Gray

Η κωδικοποίηση Gray χρησιμοποιείται σε πολλούς τομείς της πληροφορικής πχ. στην ασφάλεια πληροφοριών, αλλά επίσης χρησιμοποιείται και στους γενετικούς αλγόριθμους. Βασική ιδιότητα της κωδικοποίησης Gray είναι ότι δύο γειτονικοί ακέραιοι να απέχουν μόλις ένα bit. Υπάρχουν γενικά πολλές κωδικοποιήσεις Gray και αυτό που τις διαφοροποιεί είναι ο αριθμός που θα ξεκινήσουμε κάθε φορά, αρκεί η βασική ιδιότητα της κωδικοποίησης να διατηρείται μεταξύ δύο συνεχόμενων ακεραίων. Το μεγάλο πλεονέκτημα της κωδικοποίησης Gray είναι ότι δεν προκαλεί μεγάλες στο αποτέλεσμα για μικρές αλλαγές στο χρωμόσωμα.

ΑΡΙΘΜΟΣ	ΔΥΑΔΙΚΟΣ	GRAY
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0010
4	0100	0110
5	0101	0111
6	0110	0101
7	0111	0100
8	1000	1100
9	1001	1101
10	1010	1111
11	1011	1110
12	1100	1010
13	1101	1011
14	1110	1001
15	1111	1000

Εικόνα 3.2.2 Κωδικοποίηση Gray.

- Αναπαράσταση ακεραίων.

Οι ακέραιοι αριθμοί η αναπαράσταση γίνεται σχετικά εύκολα σχετικά με τη χρήση χρωμοσωμάτων ενός γενετικού αλγόριθμου, ωστόσο και εδώ υπάρχουν ορισμένα προβλήματα. Έστω ότι θέλουμε να κωδικοποιήσουμε αριθμούς στο διάστημα $\{0,10\}$. Για να το κάνουμε αυτό χρειαζόμαστε 4bits. Ωστόσο κατά τη διάρκεια εφαρμογής των γενετικών τελεστών μπορεί να προκύψει ο αριθμός 1011 που στο δεκαδικό σύστημα είναι ο αριθμός

11.Αυτός ο αριθμός είναι εκτός ορίων και έτσι δεν μπορεί να αναπαρασταθεί στο σύστημα μας.

- Δεκαδική αναπαράσταση.

Η δεκαδική αναπαράσταση δεν είναι το ίδιο απλή με την αναπαράσταση ακεραίων,διότι δεν υπάρχει άμεση αντιστοίχιση μεταξύ δεκαδικών και ακεραίων.Έχουμε δύο τρόπους να αναπαραστήσουμε δεκαδικούς αριθμούς με τη χρήση χρωμοσωμάτων που είναι οι εξής:

1)Δυαδική κωδικοποίηση

2)Άμεση κωδικοποίηση

- Αλφαβητική αναπαράσταση.

Αν έχουμε να αναπαραστήσουμε αλφαριθμητικές μεταβλητές με τη χρήση χρωμοσωμάτων,τότε αν αυτές οι μεταβλητές είναι πεπερασμένες σε αριθμό τότε αναπαρίστανται με συνδυασμούς ακεραίων,πχ έστω ότι έχουμε το πρόβλημα να βρέθει μια λύση στο σύνολο $\{a,b,c,d\}$,μπορούμε να κωδικοποιήσουμε το γράμμα a με τον αριθμό 1(ή δυαδικό σύστημα 01) και ούτω καθεξής.Στην περίπτωση που το σύνολο αναπαράστασης του προβλήματος δεν είναι πεπερασμένο τότε θα πρέπει να χρησιμοποιούμε αλφαριθμητικά ως χρωμοσώματα.Στην περίπτωση αυτή θα πρέπει να δημιουργήσουμε εξαρχής καινούργιους γενετικούς τελεστές και η απόδειξη σύγκλισης που υπάρχει για την περίπτωση των ακέραιων χρωμοσωμάτων δεν θα ισχύει.

3.3 Γενετικοί τελεστές

Οι Γενετικοί Αλγόριθμοι διατηρούν έναν πληθυσμό πιθανών λύσεων, του προβλήματος που μας ενδιαφέρει, πάνω στον οποίο δουλεύουν, σε αντίθεση με άλλες μεθόδους αναζήτησης που επεξεργάζονται ένα μόνο σημείο του διαστήματος αναζήτησης. Έτσι ένας Γενετικός Αλγόριθμος πραγματοποιεί αναζήτηση σε πολλές κατευθύνσεις και υποστηρίζει καταγραφή και ανταλλαγή πληροφοριών μεταξύ αυτών των κατευθύνσεων. Ο πληθυσμός υφίσταται μια προσομοιωμένη γενετική

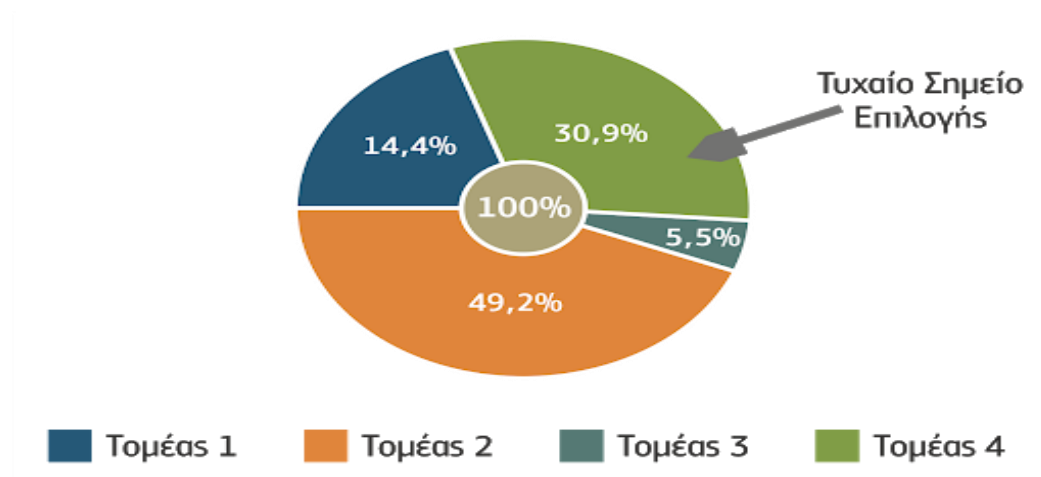
εξέλιξη χρησιμοποιώντας διάφορους γενετικούς τελεστές όπως η επιλογή, η διασταύρωση και η μετάλλαξη.

Επιλογή.

Επιλογή: Όπως στη φύση που επιλέγονται τα καλύτερα μέλη ενός πληθυσμού για αναπαραγωγή έτσι και στους γενετικούς αλγορίθμους επιλέγονται μόνο τα καλύτερα μέλη βάση κάποιων κριτηρίων για να διασταυρωθούν. Έτσι κατασκευάζεται μία λίστα από γονείς που θα ανταλλάξουν γενετικό υλικό. Η λίστα αυτή ονομάζεται *mating pool*. Ο απλούστερος τρόπος συμπλήρωσης αυτής της λίστας θα ήταν να επιλεγθούν τα καλύτερα μέλη του πληθυσμού για διασταύρωση. Ωστόσο κάτι τέτοιο θα οδηγούσε σε πρόωρη σύγκλιση του αλγορίθμου και θα ήταν αντίθετο με τους κανόνες της φύσης, όπου και τα άτομα με «κακά» χαρακτηριστικά ως προς το περιβάλλον επιλέγονται για διασταύρωση οδηγώντας πολλές φορές σε «καλούς» απογόνους.

Μέθοδος επιλογής ρουλέτας.

Μια μέθοδος επιλογής χρωμοσωμάτων είναι αυτή της ρουλέτας όπου κάθε μέλος έχει ανάλογο χώρο στη ρουλέτα με την καταλληλότητα του.



Εικόνα 3.3.1 Μέθοδος επιλογής ρουλέτας.

Για να υλοποιηθεί η επιλογή ρουλέτας υπάρχουν οι εξής τρόποι:

- Υλοποίηση με πίνακα ακεραίων.

Σε αυτόν τον πίνακα κάθε χρωμόσωμα έχει τόσες θέσεις όσο πιο μεγάλη είναι η καταλληλότητα του.

- Υλοποίηση με χρήση δεκαδικών αριθμών.

Σε αυτή την περίπτωση δεν χρειάζεται κάποιος πίνακας ακεραίων για την αποθήκευση των χρωμοσωμάτων. Οι καταλληλότητες των χρωμοσωμάτων ταξινομούνται από την χειρότερη προς την καλύτερη.

Αν και η μέθοδος επιλογής ρουλέτας εμφανίζεται στις περισσότερες επιστημονικές εργασίες, δείχνει να υστερεί έναντι άλλων μεθόδων επιλογής. Τα προβλήματα που παρουσιάζονται είναι τα εξής:

- Τα χρωμοσώματα με καταλληλότητα μεγαλύτερη από τη μέση τείνουν να έχουν περισσότερα από ένα αντίγραφα τους στην mating pool, αντίθετα τα χρωμοσώματα με μικρή καταλληλότητα πολλές φορές δεν εμφανίζονται καθόλου.
- Πρόωρη σύγκλιση.
- Στασιμότητα.

Βέβαια προβλήματα σαν αυτά μπορούν να επιλυθούν σχετικά εύκολα με τεχνικές κλιμάκωσης της καταλληλότητας.

Μέθοδος επιλογής Tournament.

Η μέθοδος επιλογής Tournament χρησιμοποιείται αρκετά και θεωρείται ο απλούστερος τρόπος επιλογής. Η μέθοδος αυτή προτιμά να επιλέγει τον καλύτερο από μια μικρή ομάδα υποψηφίων, η μέθοδος αυτή είναι και αυτή που προτιμάται από τη φύση, καθώς η φύση δεν χρησιμοποιεί εξαντλητικούς αλγόριθμους για την επιλογή των καλύτερων για αναπαραγωγή. Τα μεγάλα πλεονεκτήματα που έχει αυτή η μέθοδος είναι ότι δεν υπάρχει πρόωρη σύγκλιση αλλά ούτε και στασιμότητα όπως συμβαίνει με τη μέθοδο της επιλογής ρουλέτας. Ακόμη δεν απαιτεί ταξινόμηση των μελών του πληθυσμού με βάση τη καταλληλότητα, κάτι που έχει σαν συνέπεια να την καθιστά πιο βέλτιστη μέθοδο από αυτήν της ρουλέτας. Η μέθοδος αυτή είναι η γρηγορότερη μέθοδος επιλογής που έχει αναπτυχθεί επειδή η μέθοδος αυτή διαλέγει τους καλύτερους από κάθε ομάδα.

Μέθοδος ελυστικής επιλογής.

Η μέθοδος αυτή είναι η τροποποίηση των προηγούμενων μεθόδων, δεν είναι μια μέθοδος επιλογής που είναι εμπνευσμένη από τη φύση. Η ελυστική μέθοδος επιλογής δεν ακολουθεί τα πρότυπα της φύσης επειδή κατά κάποιο τρόπο επιβάλλει στον μηχανισμό του γενετικού αλγορίθμου να επιλέξει κάποιο στοιχείο του πληθυσμού που ίσως δεν θα επιλεγόταν. Γενικά η μέθοδος ελυστικής επιλογής χρησιμοποιείται σαν προσθήκη στις προηγούμενες μεθόδους επιλογής, επειδή μας εξασφαλίζει ότι θα επιβιώσουν οι καλύτεροι, σε ένα ή περισσότερα αντίγραφα που εμφανίζονται σε κάθε γενιά. Το βασικότερο πλεονέκτημα της μεθόδου της ελυστικής επιλογής είναι ότι η συνθήκη σύγκλισης είναι πάντα εγγυημένη. Το βασικότερο μειονέκτημα της μεθόδου αυτής είναι ότι υπάρχει κίνδυνος να παγιδευτούμε σε τοπικά ελάχιστα ή μέγιστα της συνάρτησης που εξετάζουμε.

Διασταύρωση.

Η μέθοδος της επιλογής δεν έχει κάποιο πρακτικό αποτέλεσμα από μόνη της στα χρωμοσώματα, καθώς δεν προκαλεί καμία αλλαγή στα γονίδια τους. Είναι όμως το πρώτο βήμα που πρέπει να κάνουμε πριν προχωρήσουμε στη διαδικασία της διασταύρωσης. Τα μέλη του πληθυσμού που βρίσκονται στην mating pool κατά την μέθοδο της επιλογής, με την διαδικασία της διασταύρωσης επιλέγονται και ανταλλάσσουν γενετικό υλικό. Βασικά σημεία της διασταύρωσης είναι τα εξής:

Πιθανότητα διασταύρωσης.

Η διασταύρωση δημιουργεί μια σειρά από καινούργια χρωμοσώματα με βάση τα παλιά. Η πιθανότητα διασταύρωσης μας καθορίζει το ποσοστό των παλιών χρωμοσωμάτων που πρέπει να αντικατασταθούν από τα καινούργια. Συνήθως ο αριθμός που επιλέγεται για τη πιθανότητα διασταύρωσης είναι μεταξύ 0-1. Έχει παρατηρηθεί ότι όσο μικρότερος είναι ο αριθμός πιθανότητας διασταύρωσης υπάρχει κίνδυνος να παρατηρηθεί το φαινόμενο του εγκλωβισμού του γενετικού πληθυσμού σε κάποιο τοπικό ελάχιστο. Αντίθετα αν ο αριθμός πιθανότητας διασταύρωσης είναι μεγάλος τότε είναι δυνατόν ο πληθυσμός να επικεντρωθεί στην εύρεση του ολικού ελάχιστου της αντικειμενικής συνάρτησης. Υπάρχουν διάφορα είδη διασταύρωσης που παρουσιάζονται παρακάτω.

1) Διασταύρωση ενός σημείου.

Είναι η απλούστερη διασταύρωση και η πιο συνηθισμένη, χρησιμοποιείται συνήθως σε ακέραια χρωμοσώματα αλλά και σε χρωμοσώματα διαφορετικής μορφής.

2) Διασταύρωση δύο σημείων.

Εδώ επιλέγονται δύο τυχαία σημεία αντί για ένα και ανταλλάσσονται τα αντίστοιχα τμήματα.

3) Ομοιόμορφη διασταύρωση.

Σε αυτήν τη περίπτωση οι απόγονοι λαμβάνουν τα στοιχεία τους είτε από τον πρώτο είτε από τον δεύτερο γονέα με πιθανότητα 50% για κάθε στοιχείο.

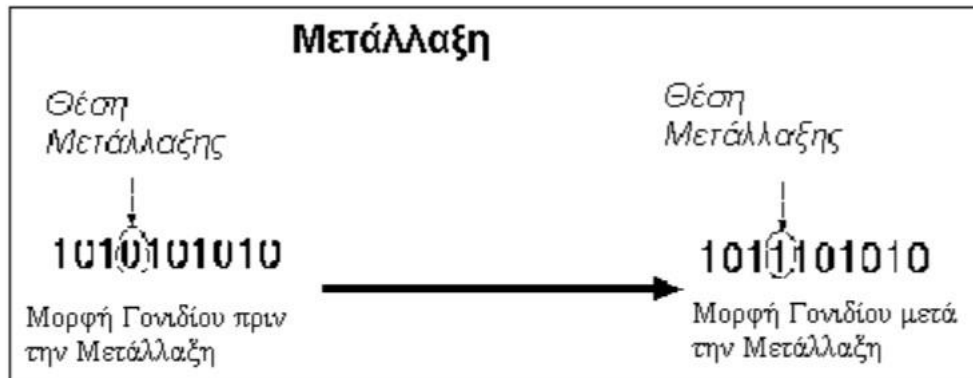
4) Δεκαδική διασταύρωση.

Όταν χρησιμοποιείται δεκαδική αναπαράσταση για τους γενετικούς αλγόριθμους, χρησιμοποιείται ένας διαφορετικός τρόπος διασταύρωσης, ο οποίος σε πολλές περιπτώσεις δίνει καλύτερα αποτελέσματα από τους προηγούμενους. Σε αυτή τη περίπτωση ο αλγόριθμος επιλέγει δύο γονείς με κάποιον αλγόριθμο επιλογής και στην συνέχεια παράγει δύο απογόνους με τη χρήση απλών γραμμικών σχέσεων.

Μετάλλαξη.

Μετάλλαξη είναι το γενικό όνομα που δόθηκε στους τελεστές μεταβλητών που χρησιμοποιούν μόνο ένα γονέα και δημιουργούν ένα παιδί εφαρμόζοντας κάποιο είδος τυχαίας επιλογής για την αναπαράσταση (γονότυπου). Η μορφή που παίρνει εξαρτάται από την επιλογή της κωδικοποίησης που χρησιμοποιείται, καθώς αυτό σημαίνει τις παραμέτρους που συνδέονται, η οποία συχνά αναφέρεται ως ρυθμός μετάλλαξης. Στις λεπτομέρειες παρακάτω συγκεντρωνόμαστε στην επιλογή των τελεστών αντί των παραμέτρων. Ωστόσο, το τελευταίο μπορεί να έχει σημαντική διαφορά στη συμπεριφορά του γενετικού αλγορίθμου. Η μετάλλαξη βασίζεται στη χρήση ενός συντελεστή μεταλλάξεως ο οποίος είναι ένας αριθμός μεταξύ

$\{0,1\}$. Όσο μεγαλύτερος είναι αυτός ο αριθμός τόσο μεγαλύτερη είναι η πιθανότητα μετάλλαξης σε ένα στοιχείο ενός χρωμοσώματος.



Εικόνα 3.3.2 Παράδειγμα μετάλλαξης.

Η διαδικασία αλλάζει με τυχαίο τρόπο το στοιχείο του χρωμοσώματος. Η αλλαγή γίνεται είτε με τυχαίο τρόπο είτε με τη χρήση κάποιου έξυπνου αλγόριθμου. Στην δεκαδική μετάλλαξη χρησιμοποιείται ένας διαφορετικός αλγόριθμος για μετάλλαξη, ο οποίος λαμβάνει υπόψιν του την ιδιαιτερότητα κάθε προβλήματος και έχει βρεθεί να είναι περισσότερο αποδοτικός από τον τυπικό αλγόριθμο μετάλλαξης.

3.4 Γενετικός Προγραμματισμός

Ο John Koza εισήγαγε την έννοια γενετικός προγραμματισμός (genetic programming) το 1992 ως μετεξέλιξη των γενετικών αλγορίθμων: ο πληθυσμός αντί να αποτελείται από δυαδικές αναπαραστάσεις, συνίσταται από συμβολικές μαθηματικές εκφράσεις, δηλαδή προγράμματα υπολογιστών. Κύριος στόχος του γενετικού προγραμματισμού είναι να παρέχει γενικές μεθόδους επίλυσης προβλημάτων με τις οποίες κτίζεται ένα πρόγραμμα βάσει της ανάλυσης απαιτήσεων. Ξεκινώντας από μια σειρά τυχαία κατασκευασμένων προγραμμάτων, η τεχνική αυτή προοδευτικά εξελίσσει ένα πληθυσμό προγραμμάτων επί σειρά γενεών, ακολουθώντας τις βιολογικές έννοιες της διασταύρωσης, της μετάλλαξης και της φυσικής επιλογής. Όπως μια ομάδα προγραμματιστών επιφέρουν προοδευτικές βελτιώσεις στον κώδικα ενός προγράμματος σε αλληπάλλληλες εκδοχές (versions), έτσι και με το γενετικό προγραμματισμό δημιουργούνται νέα

προγράμματα από ήδη υπάρχοντα, αρκεί να είναι συντακτικά σωστά και εκτελέσιμα [Lan98]. Τα προγράμματα που αποτελούν τον πληθυσμό κάθε γενιάς είναι σύνθεση συναρτήσεων (αριθμητικών πράξεων, λογικών και ανισωτικών τελεστών, ειδικών συναρτήσεων του προβλήματος) με τερματικούς όρους (terminals, λ.χ. σταθερές, μεταβλητές). Έτσι περιττεύει η αναγκαιότητα μετασχηματισμού δυαδικών αναπαραστάσεων σ' ένα μοντέλο, παρέχοντας μεγαλύτερη ευελιξία. Η επιλογή κατάλληλων συναρτήσεων και τερματικών όρων προσδιορίζει το χώρο λύσεων του προβλήματος και τελικά κρίνει την επιτυχία της διαδικασίας.

Η αναπαράσταση των προγραμμάτων γίνεται ως εξής:

Τα προγράμματα αναπαρίστανται με τη μορφή S-εκφράσεων σε γλώσσα LISP (LISt Processing language) και μάλιστα με προθεματικό συμβολισμό (prefix notation). Οι εκφράσεις αυτές μπορούν να αναπαρασταθούν και ως δένδρα δομικής ανάλυσης (parse trees) των οποίων οι κόμβοι περιέχουν τις συναρτήσεις ενώ τα φύλλα τηρούν τους τερματικούς όρους. Έτσι επιτυγχάνεται καλύτερη εποπτεία των δομών που χρησιμοποιούνται και διευκολύνεται η διαισθητική κατανόηση των λειτουργιών που θα εφαρμοστούν στα μέλη κάθε γενιάς. Από το συνδυασμό ενός αρχικού συνόλου συναρτήσεων και ενός αντίστοιχου συνόλου τερματικών όρων (μεταβλητών και σταθερών), μπορεί να σχηματιστεί ένας πολύ μεγάλος αριθμός μαθηματικών εκφράσεων μεταβαλλόμενου μήκους και πολυπλοκότητας. Η βασική διαφορά από την δυαδική αναπαράσταση που χρησιμοποιούν οι γενετικοί αλγόριθμοι είναι ότι με το γενετικό προγραμματισμό επιτυγχάνεται ιεραρχική δομή, γι' αυτό άλλωστε είναι δυνατή η δενδρική αναπαράσταση. Παρόλο που η φύση του προβλήματος επηρεάζει την επιλογή των συναρτήσεων και των όρων που θα σχηματίσουν τις μαθηματικές εκφράσεις, από κοινού τα δύο σύνολα πρέπει να εμφανίζουν δύο βασικές ιδιότητες:

- Κλειστότητα (closure). Το σύνολο των συναρτήσεων είναι κλειστό όταν κάθε συνάρτηση μπορεί να δεχθεί ως όρισμα όποια τιμή ή τύπο δεδομένων επιστρέφεται από κάποια άλλη συνάρτηση ή λαμβάνεται από κάποιο τερματικό όρο. Έτσι, λ.χ. προλαμβάνεται διαίρεση με 0 ή μπορεί να οριστεί ότι το αποτέλεσμα μιας τέτοιας διαίρεσης είναι 1. Αυτό εγγυάται ότι τα προγράμματα που θα προκύψουν θα είναι συντακτικά ορθά και συνήθως

επιτυγχάνεται επιλέγοντας συναρτήσεις και όρους από το ίδιο πεδίο τιμών (λ.χ. ακεραίους).

- **Επάρκεια (sufficiency).** Το σύνολο των συναρτήσεων και των τερματικών όρων όχι μόνο μπορεί να εκφράσει τη λύση του προβλήματος, αλλά είναι ικανό να οδηγήσει σ' αυτήν τη λύση μετά από πεπερασμένο αριθμό γενεών.

Όπως και στους γενετικούς αλγορίθμους, με τη συνάρτηση προσαρμογής αποφασίζεται πόσο καλή είναι η λύση που κάποιο πρόγραμμα δίνει στο προς επίλυση πρόβλημα. Το αποτέλεσμα της συνάρτησης προσαρμογής συνδέεται στενά με το μέγεθος των μελών. Ο γενετικός προγραμματισμός είναι δυνατόν να οδηγήσει σε ανεξέλεγκτα μεγάλες εκφράσεις (προγράμματα) με πολύ μικρή πρακτική αξία. Για την καταπολέμηση του φαινομένου, η έννοια της φειδούς (parsimony) προβλέπει εξαναγκασμό των μελών να μην υπερβαίνουν ένα ορισμένο μέγεθος, συνήθως ένα προκαθορισμένο μέγιστο μήκος προγράμματος ή ύψος δένδρου.

Τελεστές και συναρτήσεις.

Οι τελεστές που χρησιμοποιούνται για το γενετικό προγραμματισμό είναι αντίστοιχοι όσων εφαρμόζονται στους γενετικούς αλγορίθμους.

- **Επιλογή.** Συνήθως, χρησιμοποιείται η απόλυτη τιμή της προσαρμογής κάθε λύσης σε σχέση με τις υπόλοιπες της ίδιας γενιάς. Όσο υψηλότερη είναι η πιθανότητα, τόσο πιο σίγουρη είναι η αντιγραφή του μέλους αυτού στη νέα γενιά. Ο κίνδυνος πρόωρης σύγκλισης είναι πιθανός όταν ο αρχικός πληθυσμός περιλαμβάνει προγράμματα που δίνουν σχετικά καλές λύσεις. Όπως και στους γενετικούς αλγορίθμους, διάφορες τεχνικές κλιμάκωσης προσαρμογής αποπειρώνται να αποκλείσουν το ενδεχόμενο κάποιο μέλος να κυριαρχήσει επί των υπολοίπων ήδη από τις πρώτες γενιές. Επίσης, εφαρμόζονται οι μέθοδοι της επιλογής κατά σειρά κατάταξης και της ανταγωνιστικής επιλογής.
- **Διασταύρωση.** Αρχικά, βάσει της τιμής προσαρμογής επιλέγονται οι δύο γονείς και στη συνέχεια, ανεξάρτητα για κάθε γονέα, το σημείο διασταύρωσης (crossover point), δηλαδή το σημείο όπου θα αποκοπεί ένας

κλάδος (branch, subtree) απ' τον καθένα, χωρίς να είναι απαραίτητο να διαλεχτεί στην ίδια θέση για κάθε γονέα. Η διασταύρωση συνίσταται στην ανταλλαγή (swap) του επιλεγμένου κλάδου κάθε γονέα με τον αντίστοιχο κλάδο του άλλου, οπότε από δύο γονείς παράγονται ακριβώς δύο απόγονοι. Σε αντίθεση με τους γενετικούς αλγορίθμους, όπου διασταυρώσεις ίδιων γονέων οδηγούν σε απαράλλαχτους απογόνους, με την μεθοδολογία του γενετικού προγραμματισμού ακόμη και δύο πανομοιότυποι γονείς θα δώσουν δύο διαφορετικούς απογόνους εάν το σημείο διασταύρωσης δεν επιλεγεί στην ίδια θέση. Επομένως, δύο ταυτόσημες λύσεις είναι δυνατόν να δώσουν στην επόμενη γενιά δύο εντελώς διαφορετικές, γεγονός που αποτελεί συγκριτικό πλεονέκτημα ως προς τους γενετικούς αλγορίθμους. Επιπλέον, αντί να υπάρχουν σταθερού μήκους δομές σε κάθε γενιά, επιτρέπονται και αναπαράγονται μεταβλητά μεγέθη προγραμμάτων, γεγονός που προσδίδει περισσότερο δυναμισμό στην τεχνική αυτή. Βέβαια, προκειμένου να αποφευχθούν υπερβολικά μεγάλες και συνεπώς υπολογιστικά περίπλοκες εκφράσεις, συνήθως τίθεται ένα όριο στο μέγιστο βάθος του δένδρου, ώστε ο αριθμός των επιπέδων να μην ξεπερνά κάποιον σταθερό αριθμό.

- Μετάλλαξη. Ένα μικρό ποσοστό από τα μέλη κάθε γενιάς θα υποστεί μετάλλαξη, με δύο διαφορετικούς τρόπους. Στην πρώτη περίπτωση, μια συνάρτηση από το σύνολο των συναρτήσεων που έχουν προκαθοριστεί μπορεί να αντικαταστήσει μιαν άλλη συνάρτηση ή ένας τερματικός όρος κάποιον άλλο τερματικό όρο. Εναλλακτικά, ένας ολόκληρος κλάδος μπορεί να υποκατασταθεί απ' έναν άλλο κλάδο ορισμένου βάθους που δημιουργείται κατά τυχαίο τρόπο.

Τερματισμός εκτέλεσης.

Η εκτέλεση μιας διαδικασίας γενετικού προγραμματισμού μπορεί να τερματιστεί είτε όταν συμπληρωθεί ένας αριθμός γενεών (συνήθως 50) είτε όταν επιτευχθεί ένα προκαθορισμένο επίπεδο της συνάρτησης προσαρμογής. Βέβαια, χρειάζεται αρκετός πειραματισμός ώστε να οριστεί ο κατάλληλος αριθμός γενεών ή ένα ικανοποιητικό επίπεδο τιμών προσαρμογής, δεδομένου ότι δεν είναι γνωστό εκ των προτέρων εάν και πού ακριβώς η συνάρτηση εμφανίζει μέγιστο

Εκτέλεση διαδικασίας γενετικού προγραμματισμού.

Τα βασικότερα στάδια εκτέλεσης είναι τα εξής:

- Αρχικά επιλέγονται τα σύνολα των συναρτήσεων και των τερματικών όρων που θα συμμετάσχουν στο πρόγραμμα. Δεδομένου ότι δεν είναι γνωστό ευθύς εξαρχής ποιες είναι οι κατάλληλες συναρτήσεις και όροι, ο χρήστης θα πρέπει να κάνει καλή επιλογή των μελών κάθε συνόλου, κρίνοντας από τη φύση του προβλήματος που πρέπει να επιλυθεί.
- Από τις επιλεγμένες συναρτήσεις και όρους δημιουργείται ένας αρχικός πληθυσμός με μαθηματικές εκφράσεις (προγράμματα που απεικονίζονται και ως δένδρα). Αυτές οι εκφράσεις επιτρέπεται να είναι μεταβλητού μεγέθους, πρέπει όμως να είναι συντακτικά σωστές.
- Η τιμή προσαρμογής κάθε προγράμματος υπολογίζεται εκτελώντας το για έναν ορισμένο αριθμό χαρακτηριστικών περιπτώσεων (ενδεικτικές τιμές εισόδου, των οποίων το σωστό αποτέλεσμα είναι γνωστό). Για κάθε πρόγραμμα, η τιμή αυτή ισούται προς τον αριθμό των περιπτώσεων που εκτελέστηκε σωστά ή για όσες το αποτέλεσμα προσεγγίζει τη σωστή τιμή με κάποια ανοχή. Είναι εύλογο ότι τα τυχαία προγράμματα που συγκροτούν τον αρχικό πληθυσμό μάλλον δεν θα επιτύχουν να βρουν τη λύση· ωστόσο, αν ο πληθυσμός είναι αρκετά μεγάλος, υπάρχουν ελπίδες κάποια απ' αυτά κατά τύχη να προκύψουν ευνοϊκά.
- Ακολούθως, με τους τελεστές της επιλογής, της διασταύρωσης και της μετάλλαξης δημιουργείται η καινούργια γενιά. Προφανώς, ως αποτέλεσμα της διασταύρωσης το μέγεθος των εκφράσεων (ισοδύναμα, των δένδρων) μπορεί να αυξομειώνεται.
- Τα δύο προηγούμενα βήματα (3 και 4) επαναλαμβάνονται για έναν ορισμένο αριθμό γενεών.

Αξιολόγηση των γενετικών τεχνικών.

Η λειτουργία των γενετικών αλγορίθμων θεμελιώνεται πάνω στο λεγόμενο Θεώρημα Σχημάτων που διατύπωσε ο Holland. Το θεώρημα προβλέπει ότι μικρά, αλλά αποτελεσματικά δομικά στοιχεία (σχήματα) λαμβάνουν εκθετικά αυξημένη πιθανότητα αναπαραγωγής από γενιά σε γενιά, δίνοντας πολύ καλούς απογόνους. Αυτό το γεγονός εξηγεί το δυναμισμό των γενετικών τεχνικών και δικαιολογεί την αποτελεσματικότητά τους. Βεβαίως, στο επίπεδο της υλοποίησης έχουν γίνει

αρκετές παραδοχές, παραλλαγές ή και απλουστεύσεις, καθώς και ανάμιξη στοιχείων από άλλα πεδία (λ.χ. ευρετικές τεχνικές). Επιπλέον, η κωδικοποίηση με δένδρικές δομές συντελεί στην διερεύνηση ενός ανοικτού χώρου λύσεων, χωρίς να υφίσταται περιορισμούς από τους πολλούς αλλά πεπερασμένους συνδυασμούς της τυπικής κωδικοποίησης με σταθερό μέγεθος συμβολοσειρών. Πράγματι, η κωδικοποίηση της αναπαράστασης μαζί με τους γενετικούς τελεστές κρίνουν την αποτελεσματικότητα της λύσης. Οι τελεστές της διασταύρωσης και της μετάλλαξης που χρησιμοποιούνται από τις γενετικές μεθόδους είναι δυνατό να δώσουν λύσεις ριζικά διαφορετικές από τις υπάρχουσες. Η λειτουργία της διασταύρωσης παρέχει τον κύριο μηχανισμό αναζήτησης λύσεων για το πρόβλημα, γι' αυτό και συνήθως της αποδίδεται υψηλή πιθανότητα (της τάξης του 70%) να συμβεί. Η σημασία της είναι ιδιαίτερα μεγάλη κατά τα αρχικά στάδια της εκτέλεσης (τις πρώτες γενιές), όταν αναζητείται η περιοχή του χώρου λύσεων του προβλήματος όπου πιθανόν εντοπίζεται το ολικό μέγιστο. Απεναντίας, η μετάλλαξη λαμβάνει σημαντικά μικρότερη πιθανότητα εμφάνισης (της τάξης του 1% ή λιγότερο) και ο ρόλος της έγκειται κυρίως στη διατήρηση της ποικιλίας στον πληθυσμό κάθε γενιάς. Γι' αυτό και η σημασία της γίνεται ολοένα και μεγαλύτερη κατά τα τελευταία στάδια της εκτέλεσης, όταν γίνονται πλέον μικροβελτιώσεις στη διαφαινόμενη λύση. Οι γενετικές μέθοδοι έχουν ιδιαίτερη συμβολή στη μελέτη προβλημάτων βελτιστοποίησης (optimization). Αν η βέλτιστη λύση ενός προβλήματος αναζητείται με τυπικές ντετερμινιστικές μεθόδους, ελλοχεύει ο κίνδυνος να προσδιοριστεί τελικά κάποιο τοπικό ακρότατο (μέγιστο ή ελάχιστο, ανάλογα με τη συνάρτηση προσαρμογής), αφού κάθε φορά ερευνάται ένα διαφορετικό σημείο του χώρου λύσεων. Αντίθετα, με τις γενετικές τεχνικές πολλαπλά σημεία του χώρου λύσεων αναζητούνται ταυτόχρονα, οπότε η πιθανότητα εγκλωβισμού σε τοπικά ακρότατα περιορίζεται δραστικά. Επειδή, τέλος, τα μέλη του πληθυσμού επιλέγονται με τυχαίο τρόπο, η αναζήτηση κατευθύνεται προς περιοχές του χώρου λύσεων που είναι πιθανότερο να οδηγήσουν σε βελτιωμένα αποτελέσματα. Βασικό ζήτημα έρευνας στον τομέα των γενετικών μεθόδων αποτελεί η ευρωστία τους (robustness), δηλαδή πώς η αποτελεσματικότητα της προτεινόμενης λύσης συμβαδίζει με την ικανότητα επίλυσης σε διαφορετικά περιβάλλοντα. Η έννοια αυτή συμπυκνώνει το ερώτημα κατά πόσο είναι δυνατή η δημιουργία αλγορίθμων που να μην απαιτούν επανασχεδιασμό ή επαναπρογραμματισμό, αλλά να είναι δυνατόν να

προσαρμόζονται στις απαιτήσεις του εκάστοτε προβλήματος. Διάφορες προηγμένες τεχνικές έχουν προταθεί κατά καιρούς. Ο εκλεκτικισμός προνοεί ώστε οι εκλεκτότεροι εκπρόσωποι κάθε γενιάς να διατηρούνται πάντοτε άθικτοι και στην επόμενη, προστατεύοντάς τους από τυχόν παραποίηση λόγω διασταύρωσης ή μετάλλαξης. Άλλη ενδιαφέρουσα τεχνική είναι η λεγόμενη επιλογή κατά Boltzmann, όπου αρχικά τα κριτήρια είναι μεν χαλαρά για να εξασφαλίζεται μεγαλύτερη ποικιλία στον πληθυσμό, βαθμιαία όμως γίνονται αυστηρότερα καθώς ο αλγόριθμος πλησιάζει προς την σύγκλιση.

4. Ανάπτυξη εφαρμογής κατηγοριοποίησης δεδομένων σε κινητά τηλέφωνα.

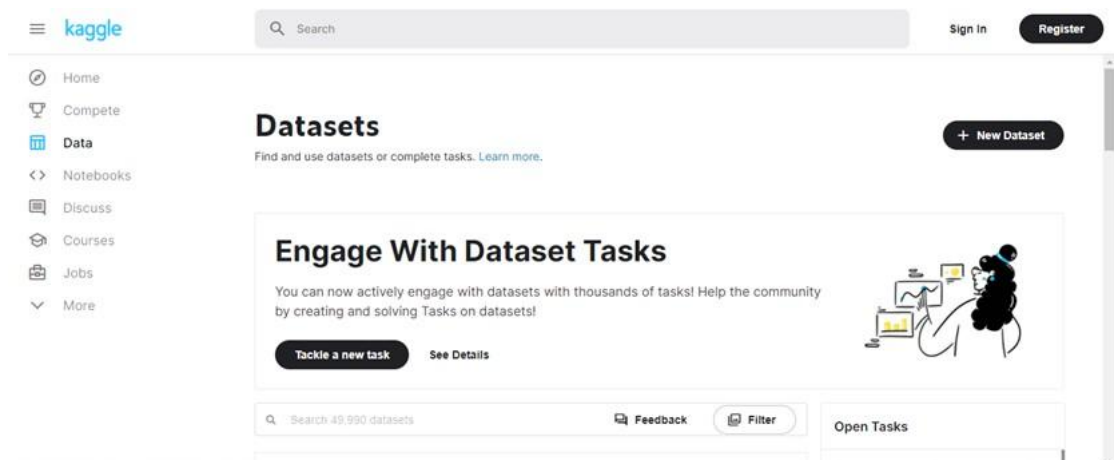
Στόχος λοιπόν είναι να δημιουργηθεί ένα μοντέλο που με κάποια στοιχεία σε ένα κινητό να μπορεί να αναγνωρίζει με σχετικά μεγάλη ακρίβεια το φύλο του χρήστη. Η υλοποίηση του μοντέλου θα γίνει σε γλώσσα προγραμματισμού Python.

4.1 Τα δεδομένα του προβλήματος.

Το σύνολο δεδομένων που επέλεξα έχει βρεθεί σε διαδικτυακή σελίδα με διάφορα σύνολα για ανάλυση classification ή clustering και συγκεκριμένα την σελίδα:

<https://www.kaggle.com/datasets>

Στην σελίδα αυτή υπάρχουν σειρά από σύνολα που μπορεί κάποιος να χρησιμοποιήσει προκειμένου να ερευνήσει για διάφορες περιπτώσεις όπως και το δικό μας σύνολο που αφορά την χρήση κινητών συσκευών.



Εικόνα 4.1.1 Σελίδα δεδομένων προβλήματος.

Πιο συγκεκριμένα το σύνολο μας είναι ένα σύνολο σε μορφή “csv” όπου έχει τις παρακάτω στήλες:

- FB,
- Instagram,
- News,
- TikTok,

- GamesA,
- GamesB,
- Banks,
- Sex,
- Age.

Κάθε στήλη χαρακτηρίζει τον αριθμό που κάποιος χρήστη άνοιξε καινούριο session σε μία εφαρμογή σε μια μέρα

To FB αφορά την εφαρμογή Facebook

To Instagram την εφαρμογή Instagram

To News εφαρμογές ειδήσεων

To TikTok την εφαρμογή TikTok

To GamesA εφαρμογή παιχνιδιών δράσης , πολέμου , αθλητισμού

To GamesB εφαρμογή παιχνιδιών απλής διασκέδασης

To Banks συστήματα ebanking

To Sex αφορά το φύλο (1 άντρας, 2 γυναίκα)

To Age την ηλικία του χρήστη (από 24 και πάνω).

Ένα μέρος των δεδομένων παρασυσιάζεται παρακάτω:

FB	Instagram	News	TikTok	GamesA	GamesB	Banks	Sex	Age
1	5	1	4	5	2	3	1	35
2	9	2	2	7	2	3	1	29
9	4	7	2	1	1	3	1	36
5	13	3	4	5	4	3	1	24
4	9	1	3	5	4	4	2	26
10	6	6	1	0	5	2	2	60
9	5	5	1	3	9	2	2	50
8	6	6	2	3	10	5	2	36
7	1	9	1	1	2	5	1	55

Εικόνα 4.1.2 Δεδομένα προβλήματος.

Το σύνολο των δεδομένων αποτελείται από 600 τέτοιες εγγραφές από διαφορετικά άτομα.

4.2 Εφαρμογή τεχνητών νευρωνικών δικτύων στα δεδομένα.

Για την επίλυση του προβλήματος χρειαστηκε να εφαρμοστούν μοντέλα όπως τα νευρωνικά δίκτυα με τον πλέον κατάλληλο τρόπο ορίζοντας τις βέλτιστες παραμέτρους.

Π.χ αν ορίσουμε ένα νευρωνικό δίκτυο τριών (3) επιπέδων τότε έχουμε να δούμε παραμέτρους όπως τους νευρώνες σε κάθε επίπεδο, τις παραμέτρους learning rate και ορμή (momentum) κ.α.

Πάρακατω παρουσιάζεται η συνάρτηση του νευρωνικού δικτύου που χρησιμοποιήθηκε στο πρόβλημα.

Η συνάρτηση του νευρωνικού δικτύου

```
def MLP(X,y, lr,mom,lev1,lev2,lev3)
    Xtrain=X[1:500]
    Ytrain=y[1:500]
    clf =
MLPClassifier(hidden_layer_sizes=(lev1,lev2,lev3),momentum=mom,
learning_rate_init=lr,random_state=1, max_iter=300)

    clf = clf.fit(Xtrain, Ytrain)
    Xtest=X[500:600]
    Ytest=y[500:600]
    return (clf.score(Xtest,Ytest))
```

Όπως βλέπουμε χρησιμοποιεί τις 500 εγγραφές για σύνολο εκπαίδευση και 100 για σύνολο test και δίνει σαν αποτέλεσμα το score του νευρωνικού δηλαδή το ποσοστό των σωστών προβλέψεων.

4.3 Χρήση γενετικού αλγόριθμου.

Για να βρούμε τις καλύτερες παραμέτρους στο πρόβλημα μας πρέπει να κάνουμε πολλές δοκιμές που ουσιαστικά είναι άπειροι συνδυασμοί. Ένας τρόπος για να ξεπεράσουμε ένα τέτοιο πρόβλημα είναι η εφαρμογή γενετικών αλγορίθμων. Πάρακατω παρουσιάζεται ο γενετικός αλγόριθμος που χρησιμοποιήθηκε στο πρόβλημα μας.

Γενετικός Αλγόριθμος

Στην συνέχεια εφαρμόζουμε το γενετικό αλγόριθμο

```
for i in range (0,N):
    print("Gen: ",i)
    ev,crom=zip(*sorted(zip(ev,crom),reverse=True))
    ev=list(ev)
    crom=list(crom)
    if(random.random()<pcross):
        i1=random.randrange(0,int(MC/2))
        i2=random.randrange(0,int(MC/2))

        lr=(crom[i1][0]+crom[i2][0])/2
        mom=(crom[i1][1]+crom[i2][1])/2
        lev1=int((crom[i1][2]+crom[i2][2])/2)
        lev2=int((crom[i1][3]+crom[i2][3])/2)
        lev3=int((crom[i1][4]+crom[i2][4])/2)
        c=[lr, mom ,lev1, lev2, lev3 ]

    for j in range (0,5):
        crom[MC-1][j]=c[j]
        score=MLP(X,y, lr,mom,lev1,lev2,lev3)
        ev[MC-1]=score
```

```

if(random.random()< pmul):
    i1=random.randrange(0,int(MC/2))
    i2=random.randrange(0,5)
    k=[]
    k.append(random.random())
    k.append(random.random())
    k.append(random.randrange(2,100))
    k.append(random.randrange(2,100))
    k.append(random.randrange(2,100))

    crom[i1][i2]=k[i2]
    i2=random.randrange(0,5)
    crom[i1][i2]=k[i2]
    score=MLP(X,y, crom[i1][0],crom[i1][1],crom[i1][2],crom[i1][3],crom[i1][4])
    ev[MC-1]=score

```

Όπως βλέπουμε όταν έχουμε διασταύρωση γίνεται για κάθε χρωμόσωμα του παιδιού στα αντίστοιχα στοιχεία του ο μέσος όρος των δύο γονέων.

Στη μετάλλαξη έχουμε 2 στοιχεία ενός χρωμοσωματος να αλλάζουν τυχαία.

Για κάθε στοιχείο εφαρμόζεται το νευρωνικό και παίρνουμε την νεά τιμή του score δηλαδή την πιθανότητα να έχουμε σωστή πρόβλεψη.

4.4 Ενδεικτικά αποτελέσματα.

Πάρακατω παρουσιάζονται τα αποτελέσματα της εκτέλεσης της εφαρμογής.

```
start
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
```

Εικόνα 4.4.1 Έναρξη εκτέλεσης αλγόριθμου.

```
Gen : 0
Gen : 1
Gen : 2
Gen : 3
Gen : 4
Gen : 5
Gen : 6
Gen : 7
Gen : 8
Gen : 9
Gen : 10
Gen : 11
Gen : 12
Gen : 13
Gen : 14
Gen : 15
Gen : 16
Gen : 17
Gen : 18
Gen : 19
Gen : 20
```

Εικόνα 4.4.2 Συνέχεια εκτέλεσης αλγόριθμου.

```
[0.07046332589209292, 0.42672307197664616, 70, 35, 58], [0.33945682574642866, 0.7965507155607557, 62, 36, 62], [0.12635228351418692, 0.6237472435259129, 43, 73, 27], [0.12635228351418692, 0.4074392373265052, 44, 73, 27], [0.12292915944034372, 0.960106306968617, 70, 55, 57], [0.12292915944034372, 0.960106306968617, 39, 55, 17], [0.12292915944034372, 0.41624571420074047, 57, 66, 85], [0.12292915944034372, 0.41624571420074047, 57, 66, 57], [0.12292915944034372, 0.41624571420074047, 57, 66, 57], [0.10262745083581204, 0.5200814118808852, 60, 55, 47], [0.10262745083581204, 0.5200814118808852, 60, 55, 47], [0.08061418019435876, 0.37480522313657383, 68, 40, 77], [0.07890261815743715, 0.6327235864352653, 76, 37, 68], [0.06999285553316226, 0.9320824407981234, 80, 59, 58], [0.05821210562937945, 0.7299627689811949, 45, 58, 39], [0.05821210562937945, 0.7299627689811949, 45, 40, 39], [0.048255653288876627, 0.531663750526801, 62, 27, 54], [0.048255653288876627, 0.531663750526801, 62, 27, 54], [0.048255653288876627, 0.4378615601482808, 20, 27, 54], [0.0382992009483738, 0.33336473207240713, 80, 15, 70], [0.0382992009483738, 0.33336473207240713, 80, 15, 70], [0.5249284640843447, 0.43976911949594555, 41, 42, 44], [0.5249284640843447, 0.43976911949594555, 41, 42, 44], [0.5249284640843447, 0.43976911949594555, 41, 42, 44], [0.12635228351418692, 0.9537011798586947, 43, 52, 27], [0.12635228351418692, 0.7525551953961732, 43, 67, 65], [0.12635228351418692, 0.4074392373265052, 36, 97, 27], [0.12635228351418692, 0.4074392373265052, 14, 72, 27], [0.12464072147726532, 0.6849734470297176, 50, 59, 42], [0.12464072147726532, 0.4118424757636229, 46, 81, 56], [0.12292915944034372, 0.960106306968617, 91, 55, 6], [0.12292915944034372, 0.5812226823671314, 57, 55, 57], [0.12292915944034372, 0.41624571420074047, 98, 66, 34], [0.12292915944034372, 0.41624571420074047, 57, 79, 57], [0.12292915944034372, 0.2595825759310343, 57, 20, 57], [0.11950603536650051, 0.9320824407981234, 72, 59, 66], [0.11950603536650051, 0.9320824407981234, 36, 75, 66], [0.11950603536650051, 0.3622750413496848, 54, 59, 87], [0.11950603536650051, 0.208744184875568, 72, 59, 87], [0.11950603536650051, 0.208744184875568, 72, 59, 87], [0.11950603536650051,
```

Εμφανίζει ταξινομημένες τις καλύτερες τιμές για το νευρωνικό δίκτυο. Η πρώτη τιμή είναι και η καλύτερη.

```
[0.85, 0.84, 0.84, 0.84, 0.84, 0.84, 0.84, 0.84, 0.84, 0.84, 0.84, 0.84, 0.84,
0.84, 0.84, 0.84, 0.84, 0.84, 0.84, 0.84, 0.84, 0.83, 0.83, 0.83, 0.83,
0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 0.83,
0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 0.83, 0.49, 0.82]
```

Εμφανίζονται ταξινομημένα τα ποσοστά επιτυχίας των καλύτερων τιμών του νευρωνικού δικτύου.

```
Best Solution: [0.07046332589209292, 0.42672307197664616, 70, 35, 58]  
Evaluate Solution: 0.85
```

Εικόνα 4.4.5 Εμφάνιση της καλύτερης λύσης και το ποσοστό επιτυχίας.

Εμφανίζει όπως βλέπουμε το καλύτερο νευρωνικό δίκτυο με learning rate :0.70, momentum : 0.42 και κόμβους στα τρία επίπεδα 70,35,58.Που δίνει πιθανότητα εκτίμησης 0.85 στο test σύνολο.

	FB	Instagram	News	TikTok	GamesA	GamesB	Banks	Ytest	Ypred
500	5	6	3	6	9	5	3	2	2
501	10	5	5	0	2	4	5	2	2
502	4	1	5	2	0	2	3	1	1
503	2	12	1	6	3	6	3	2	1
504	6	4	9	0	1	2	3	1	1
..	..	...	...	...	...	...	...	...	...
595	8	2	10	0	2	1	2	1	1
596	6	2	9	1	0	2	6	1	1
597	2	9	4	5	5	6	4	1	1
598	3	11	4	5	10	4	3	1	1
599	6	6	3	2	0	5	2	2	2

Εικόνα 4.4.6 Εμφάνιση του πίνακα δεδομένων καθώς και των προβλέψεων του προγράμματος.

Ο παραπάνω πίνακας είναι ο πίνακας δεδομένων που χρησιμοποιήθηκε με την διαφορά ότι η στήλη Ytest είναι το φύλο του κάθε χρήστη και η στήλη Ypred είναι η πρόβλεψη που έκανε ο αλγόριθμος.Εμφανίζονται δίπλα δίπλα για να δούμε το ποσοστό επιτυχίας και στη πράξη.

4.5 Παραδείγματα από τον κώδικα.

Στόχος λοιπόν είναι να δημιουργηθεί ένα μοντέλο που με αυτά τα στοιχεία(τα δεδομένα που παρουσιάστηκαν προηγουμένως) σε ένα κινητό να μπορεί να αναγνωρίζει με σχετικά μεγάλη ακρίβεια το φύλο του χρήστη.

Για να γίνει αυτό πρέπει να εφαρμοστούν μοντέλα όπως τα νευρωνικά δίκτυα με τον πλέον κατάλληλο τρόπο ορίζοντας τις βέλτιστες παραμέτρους.

Π.χ αν ορίσουμε ένα νευρωνικό δίκτυο τριών (3) επιπέδων τότε έχουμε να δούμε παραμέτρους όπως τους νευρώνες σε κάθε επίπεδο, τις παραμέτρους learning rate και ορμή (momentum) κ.α.

Για να βρούμε τις καλύτερες παραμέτρους πρέπει να κάνουμε πολλές δοκιμές που ουσιαστικά είναι άπειροι συνδυασμοί. Ένας τρόπος για να ξεπεράσουμε ένα τέτοιο πρόβλημα είναι η εφαρμογή γεννητικών αλγορίθμων.

Θεωρώντας σαν χρωμοσώματα τις παραμέτρους που "πειράζουμε" δηλαδή τον αριθμό νευρώνων σε κάθε επίπεδο, το learning rate και το Momentum τότε μπορούμε με χρήση ενός τυχαίου αριθμού και εφαρμογή ενός γεννητικού αλγορίθμου N γενεών να έχουμε το νευρωνικό που με συνεχείς μεταλλάξεις και διασταυρώσεις του πληθυσμού θα πετύχει την καλύτερη επίδοση.

Με βάση τα παραπάνω δημιουργήθηκε κώδικας σε python που είναι ο παρακάτω:

```
import random
import pandas
from sklearn import tree
from sklearn.neural_network import MLPClassifier
import matplotlib.pyplot as plt
import matplotlib.image as pltimg

def MLP(X,y, lr,mom,lev1,lev2,lev3):

Xtrain=X[1:500]
Ytrain=y[1:500]
```

```
clf = MLPClassifier(hidden_layer_sizes=(lev1,lev2,lev3),momentum=mom,  
learning_rate_init=lr,random_state=1, max_iter=300)
```

```
clf = clf.fit(Xtrain, Ytrain)  
Xtest=X[500:600]  
Ytest=y[500:600]  
return (clf.score(Xtest,Ytest))
```

```
df = pandas.read_csv("data.csv")
```

```
features = ['FB','Instagram','News','TikTok','GamesA','GamesB','Banks']
```

```
X= df[features]  
y = df['Sex']
```

```
crom=[]  
ev=[]  
MC=50  
N=100  
pcross=0.9  
pmul=0.9  
print("start")  
for i in range(0,MC):  
    lr=random.random()  
    lev1=random.randrange(2,100)  
    lev2=random.randrange(2,100)  
    lev3=random.randrange(2,100)  
    mom=random.random()  
  
    c=[lr, mom ,lev1, lev2, lev3 ]  
  
    crom.append(c)
```

```

score=MLP(X,y, lr,mom,lev1,lev2,lev3)

    ev.append(score)
    print (i)

for i in range (0,N):
    print("Gen: ",i)

    ev,crom=zip(*sorted(zip(ev,crom),reverse=True))

    ev=list(ev)
    crom=list(crom)

    if(random.random()<pcross):

        i1=random.randrange(0,int(MC/2))
        i2=random.randrange(0,int(MC/2))

        lr=(crom[i1][0]+crom[i2][0])/2
        mom=(crom[i1][1]+crom[i2][1])/2
        lev1=int((crom[i1][2]+crom[i2][2])/2)
        lev2=int((crom[i1][3]+crom[i2][3])/2)
        lev3=int((crom[i1][4]+crom[i2][4])/2)

        c=[lr, mom ,lev1, lev2, lev3 ]

    for j in range (0,5):

```

```

        crom[MC-1][j]=c[j]

    score=MLP(X,y, lr,mom,lev1,lev2,lev3)
    ev[MC-1]=score

    if(random.random()< pmul):

        i1=random.randrange(0,int(MC/2))

        i2=random.randrange(0,5)

        k=[]
        k.append(random.random())
        k.append(random.random())
        k.append(random.randrange(2,100))
        k.append(random.randrange(2,100))
        k.append(random.randrange(2,100))

        c=[0,0,0,0,0]
        for j in range (0,5):
            c[j]=crom[i1][j]

            c[i2]=k[i2]
            i2=random.randrange(0,5)

            c[i2]=k[i2]

        score=MLP(X,y, c[0],c[1],c[2],c[3],c[4])

```

```

    crom[MC-2]=c
    ev[MC-2]=score

    print(crom)
    print(ev)

    print("Best Solution:", crom[0])
    print("Evaluate Solution:", ev[0])

    lr=crom[0][0]
    mom=crom[0][1]
    lev1=crom[0][2]
    lev2=crom[0][3]
    lev3=crom[0][4]

    clf = MLPClassifier(hidden_layer_sizes=(lev1,lev2,lev3),momentum=mom,
    learning_rate_init=lr,random_state=1, max_iter=300)
    Xtrain=X[0:500]
    Ytrain=y[0:500]

    clf = clf.fit(Xtrain, Ytrain)

    Xtest=X[500:600]
    Ytest=y[500:600]

    print(clf.score(Xtest,Ytest))

    Ypred=clf.predict(Xtest)

    R=pandas.DataFrame(Xtest)
    R.insert(7,"Ytest",Ytest,True)

```

```
R.insert(8,"Ypred",Ypred,True)
print(R)
R.to_csv("result.csv")
```



Πάρακατω παρουσιάζονται τα βασικότερα σημεία του αλγόριθμου:

Η συνάρτηση του νευρωνικού δικτύου

```
def MLP(X,y, lr,mom,lev1,lev2,lev3):
    Xtrain=X[1:500]
    Ytrain=y[1:500]
    clf = MLPClassifier(hidden_layer_sizes=(lev1,lev2,lev3),momentum=mom,
learning_rate_init=lr,random_state=1, max_iter=300)
    clf = clf.fit(Xtrain, Ytrain)

    Xtest=X[500:600]
    Ytest=y[500:600]
    return (clf.score(Xtest,Ytest))
```

Όπως βλέπουμε χρησιμοποιεί τις 500 εγγραφές για σύνολο εκπαίδευση και 100 για σύνολο test και δίνει σαν αποτέλεσμα το score του νευρωνικού δηλαδή το ποσοστό των σωστών προβλέψεων.

Διάβασμα αρχείου και καταχώρηση σε λίστες

```
df = pandas.read_csv("data.csv")
features = ['FB','Instagram','News','TikTok','GamesA','GamesB','Banks']

X= df[features]
y = df['Sex']
```

Όπως βλέπουμε διαβάζουμε το αρχείο με χρήση της βιβλιοθήκης pandas και παίρνουμε τα σύνολα που θέλουμε.

Ουσιαστικά το X και y έχει τις τιμές από το αρχείο μας για τα αντίστοιχα πεδία.

Δημιουργία αρχικού συνόλου χρωμοσωμάτων

Με τον παρακάτω κώδικα δημιουργούμε τα χρωμοσώματα δηλαδή ορίζουμε MC τιμές για το learning rate, momentum , Nodes στο επίπεδο 1, Nodes στο επίπεδο 2, Nodes στο επίπεδο 3.

Οι λίστες crom και ev είναι οι λίστες με τις τιμές των χρωμοσωμάτων και η αξιολόγηση της εφαρμογής του νευρωνικού για τις αντίστοιχες τιμές των χρωμοσωμάτων.

Επίσης το N είναι οι γενιές που θα εξετάσουμε , pcross η πιθανότητα διασταύρωσης , pmul η πιθανότητα μετάλλαξης.

```
crom=[]
ev=[]
MC=50
N=100

pcross=0.9
pmul=0.9

print("start")

for i in range(0,MC):
    lr=random.random()
    lev1=random.randrange(2,100)
    lev2=random.randrange(2,100)
    lev3=random.randrange(2,100)
    mom=random.random()
    c=[lr, mom ,lev1, lev2, lev3 ]
    crom.append(c)
    score=MLP(X,y, lr,mom,lev1,lev2,lev3)
    ev.append(score)
    print (i)
```

Γενετικός Αλγόριθμος

Στην συνέχεια εφαρμόζουμε το γενετικό αλγόριθμο

```
for i in range (0,N):
    print("Gen: ",i)
    ev,crom=zip(*sorted(zip(ev,crom),reverse=True))
    ev=list(ev)
    crom=list(crom)
    if(random.random()<pcross):
        i1=random.randrange(0,int(MC/2))
        i2=random.randrange(0,int(MC/2))

        lr=(crom[i1][0]+crom[i2][0])/2
        mom=(crom[i1][1]+crom[i2][1])/2
        lev1=int((crom[i1][2]+crom[i2][2])/2)
        lev2=int((crom[i1][3]+crom[i2][3])/2)
        lev3=int((crom[i1][4]+crom[i2][4])/2)
        c=[lr, mom ,lev1, lev2, lev3 ]

    for j in range (0,5):
        crom[MC-1][j]=c[j]
    score=MLP(X,y, lr,mom,lev1,lev2,lev3)
    ev[MC-1]=score

    if(random.random()< pmul):
        i1=random.randrange(0,int(MC/2))
        i2=random.randrange(0,5)
        k=[]
        k.append(random.random())
        k.append(random.random())
        k.append(random.randrange(2,100))
        k.append(random.randrange(2,100))
        k.append(random.randrange(2,100))

    crom[i1][i2]=k[i2]
    i2=random.randrange(0,5)
    crom[i1][i2]=k[i2]
    score=MLP(X,y, crom[i1][0],crom[i1][1],crom[i1][2],crom[i1][3],crom[i1][4])
    ev[MC-1]=score
```

Όπως βλέπουμε όταν έχουμε διασταύρωση γίνεται για κάθε χρωμόσωμα του παιδιού στα αντίστοιχα στοιχεία του ο μέσος όρος των δύο γονέων.

Στη μετάλλαξη έχουμε 2 στοιχεία ενός χρωμοσωματος να αλλάζουν τυχαία.

Για κάθε στοιχείο εφαρμόζεται το νευρωνικό και παίρνουμε την νεά τιμή του score δηλαδή την πιθανότητα να έχουμε σωστή πρόβλεψη.

5. Συμπεράσματα-Μελλοντικές κατευθύνσεις.

Όπως είδαμε με την εκτέλεση του παραπάνω προγράμματος πετύχαμε έναν ποσοστό επιτυχίας σωστής πρόβλεψης 85%. Φυσικά αυξάνοντας τον πληθυσμό και τις γενεές πιθανά να έχουμε και καλύτερα αποτελέσματα αλλά χρειάζονται πολύ ισχυροί υπολογιστικοί πόροι και μεγάλος χρόνος αναμονής.

Μια τέτοια εφαρμογή θα μπορούσε να βοηθήσει τις εταιρίες για πιο ουσιαστικό και στοχοποιημένο μάρκετινγκ αφού αναγνωρίζοντας το φύλο μπορεί να προβάλει τις κατάλληλες διαφημίσεις που αφορούν το φύλο , όπως γυναικεία ρούχα μόνο στις γυναίκες, αντρικά παπούτσια μόνο σε άντρες κλπ.Επίσης θα μπορούσαμε με μια μικρή τροποποίηση του συγκεκριμένου αλγόριθμου να αναγνωρίσουμε και την ηλικία του κάθε χρήστη(ή κάποιο άλλο χαρακτηριστικό του) χωρίζοντάς την σε ηλικιακές ομάδες. πχ.(18-27,28-37 κτλ.)Και να παρουσιάζουμε σύμφωνα με την ηλικιακή ομάδα του κάθε χρήστη τα ανάλογα ενδιαφέροντα.Η εφαρμογή υλοποιήθηκε σε γλώσσα προγραμματισμού Python διότι είναι μία αντικειμενοστραφής γλώσσα προγραμματισμού,μαςδίνει πολλές βιβλιοθήκες,είναι μια επεκτάσιμη γλώσσα προγραμματισμού και διατίθεται δωρεάν.Δεν είναι τυχαίο ότι τα τελευταία χρόνια στην Ελλάδα, έχει τεράστια απήχηση και αποδοχή, όχι μόνο στον χώρο της επιστήμης των υπολογιστών (Computer Science), αλλά και σε άλλους εμπορικούς και επιστημονικούς τομείς. Χαρακτηριστική είναι η ταχύτατη διείσδυση της γλώσσας τόσο σε πρακτικές εφαρμογές (σε επιχειρήσεις και οργανισμούς), όσο και στον τομέα της διδασκαλίας και εκμάθησης προγραμματισμού σε διάφορα επίπεδα εκπαίδευσης (από τη μέση μέχρι την ανώτατη εκπαίδευση). Δεν είναι τυχαίο το ότι χρησιμοποιείται σε πάρα πολλούς και ασυνήθιστα, θα λέγαμε, διαφορετικούς τομείς δραστηριοτήτων και αυτό είναι κάτι που την κάνει μοναδική στο είδος της. Η ιδιαίτερη ευκολία εκμάθησης είναι ένα από τα χαρακτηριστικά της. Για αυτό τον λόγο πλέον αποτελεί, για πολλούς και πολλές, την πρώτη γλώσσα προγραμματισμού που θα μάθουν.

Η πλατφόρμα που χρησιμοποιήθηκε για την υλοποίηση της εφαρμογής είναι Spyder(Scientific Python Development EnviRonment) .Προορίζεται για χρήση από προγραμματισμό που αφορά την επιστήμη, και αυτό αντικατοπτρίζεται στο σύνολο χαρακτηριστικών, το packaging και τη συνολική συμπεριφορά του IDE. Το Spyder έχει χρήσιμες δυνατότητες για τη γενική Python, αλλά αν δεν δουλεύετε κυρίως με

πακέτα IPython και επιστημονικής πληροφορικής, μάλλον καλύτερα να χρησιμοποιήσετε ένα διαφορετικό IDE.

Ο μεγαλύτερος λόγος να μην χρησιμοποιήσετε το Spyder ως περιβάλλον ανάπτυξης Python γενικής χρήσης δεν είναι το σύνολο χαρακτηριστικών, αλλά η διαδικασία εγκατάστασης. Το Spyder δεν παραδίδεται ως αυτόνομο εκτελέσιμο με τον τρόπο ενός προϊόντος όπως το Visual Studio ή το PyCharm. Αντ' αυτού, είναι εγκατεστημένο ως πακέτο Python. Ο ευκολότερος τρόπος για εγκατάσταση του Spyder είναι μέσω του Anaconda της Continuum Analytics.

Βιβλιογραφία.

[1]ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ.Κωνσταντίνος Διαμαντάρας.Εκδόσεις κλειδάριθμος.[Έντυπη μορφή,Βιβλίο].

[2]Μηχανική Μάθηση.Τεχνητή Νοημοσύνη-Β Έκδοση.Ι.Βλαχάβας, Π.Κεφαλάς,Ν.Βασιλειάδης,Φ.Κόκκορας,Η.Σακελλαρίου.[ONLINE]Available at: <http://aibook.csd.auth.gr/include/slides/Chap18.pdf>

[Accessed 20 January 2021]

[3]Μηχανική Μάθηση.Γεωργούλη Αικατερίνη. [ONLINE]Available at: https://repository.kallipos.gr/bitstream/11419/3382/1/02_chapter_04.pdf

[Accessed 20 January 2021]

[4] Technative.io. [ONLINE]Available at: <https://www.technative.io/why-unsupervised-machine-learning-is-the-future-of-cybersecurity/>

[Accessed 20 January 2021]

[5]E-class ΑΤΕΙ Πελοποννήσου.Μηχανική Μάθηση. [ONLINE]Available at: <http://www.eclass.teipel.gr/modules/document/file.php/ITCOM664/Chap18%20MACHINE%20LEARNING%20%28AT%20v8%CE%B1-intro%2BEvaluation%2BCandElimALgor%29.pdf>

[Accessed 20 January 2021]

[6]ΙΕΚ ΔΕΛΤΑ.Τι είναι μηχανική μάθηση? [ONLINE]Available at: <https://www.iekdelta360.gr/ti-einai-i-michaniki-mathisi-machine-learning>

[Accessed 20 January 2021]

[7]Computer Science Center CSC.Machine Learning-Μηχανική Μάθηση-Τι είναι? [ONLINE]Available at: <https://www.csc.com.gr/machine-learning-%CE%BC%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%AE-%CE%BC%CE%AC%CE%B8%CE%B7%CF%83%CE%B7-%CF%84%CE%B9-%CE%B5%CE%AF%CE%BD%CE%B1%CE%B9/>

[Accessed 20 January 2021]

[8]Τεχνητή Νοημοσύνη.Μια εισαγωγική προσέγγιση.Κατερίνα Γεωργούλη.

[ONLINE]Available at: <http://ai-ebook.cs.teiath.gr/04a-main.html>

[Accessed 20 January 2021]

[9]ΒΙΚΙΠΑΙΔΕΙΑ.Νευρωνικό Δίκτυο. [ONLINE]Available at:

https://el.wikipedia.org/wiki/%CE%9D%CE%B5%CF%85%CF%81%CF%89%CE%BD%CE%B9%CE%BA%CF%8C_%CE%B4%CE%AF%CE%BA%CF%84%CF%85%CE%BF

[Accessed 20 January 2021]

[10]Αλγόριθμος οπισθοδιάδοσης του σφάλματος-Back Propagation.Αναστάσιος Ντούνης.

[ONLINE]Available at: http://auto.teipir.gr/sites/default/files/dialexi_11_bl_mlp.pdf

[Accessed 20 January 2021]

[11]Πανεπιστήμιο Πάτρας,Μεταπτυχιακό δίπλωμα ειδίκευσης μαθηματικά των υπολογιστών και των αποφάσεων.Σύγκριση μεθόδων εκπαίδευσης νευρωνικών δικτύων.Σταθοπούλου Δήμητρα. [ONLINE]Available at:

<https://nemertes.lis.upatras.gr/jspui/bitstream/10889/4419/1/%ce%94%ce%99%ce%a0%ce%9b%ce%a9%ce%9c%ce%91%ce%a4%ce%99%ce%9a%ce%97%20%ce%a3%ce%a4%ce%91%ce%98%ce%9f%ce%a0%ce%9f%ce%a5%ce%9b%ce%9f%ce%a5%20%ce%94%ce%97%ce%9c%ce%97%ce%a4%ce%a1%ce%91%ce%a3.pdf>

[Accessed 20 January 2021]

[12]ΒΙΚΙΠΑΙΔΕΙΑ.Γενετικοί Αλγόριθμοι. [ONLINE]Available at:

https://el.wikipedia.org/wiki/%CE%93%CE%B5%CE%BD%CE%B5%CF%84%CE%B9%CE%BA%CE%BF%CE%AF_%CE%91%CE%BB%CE%B3%CF%8C%CF%81%CE%B9%CE%B8%CE%BC%CE%BF%CE%B9

[Accessed 20 January 2021]

[13]Διπλωματική εργασία βελτιστοποίηση χρονοπρογραμματισμού παραγωγής με χρήση γενετικών αλγορίθμων.Αλέξανδρος Μ.Κελεμένης. [ONLINE]Available at:

<http://artemis.cslab.ece.ntua.gr:8080/jspui/bitstream/123456789/13904/1/DT2003-0130.pdf>

[Accessed 20 January 2021]

[14]Πτυχιακή Εργασία,Σύγκριση νέων εξελικτικών αλγορίθμων.Σιλιαχλή Σουλτάνα.

[ONLINE]Available at: <https://core.ac.uk/download/pdf/132817654.pdf>

[Accessed 20 January 2021]

[15]Διπλωματική Εργασία,Μοντελοποίηση χρονοσειρών με χρήση γεντικού

προγραμματισμού.Θεοφιλάτος Κων/νος. [ONLINE]Available at:

<https://nemertes.lis.upatras.gr/jspui/bitstream/10889/1654/1/thesis.pdf>

[Accessed 20 January 2021]

[16]Εθνικό Μετσόβιο Πολυτεχνίο,Εργασία Μαθήματος:Τεχνητά νευρωνικά δίκτυα και μηχανική μάθηση.Κωνσταντίνος Χ.Πατρούμπας. [ONLINE]Available at:

http://www.dbnet.ece.ntua.gr/~kpatro/Scripta/Patroumpas_GeneticQueryOptimization.pdf

[Accessed 20 January 2021]

[17]ΒΙΚΙΠΑΙΔΕΙΑ.Κατηγοριοποίηση. [ONLINE]Available at:

<https://el.wikipedia.org/wiki/%CE%9A%CE%B1%CF%84%CE%B7%CE%B3%CE%BF%CF%81%CE%B9%CE%BF%CF%80%CE%BF%CE%AF%CE%B7%CF%83%CE%B7>

[Accessed 20 January 2021]

[18]Μηχανική Μάθηση.Η Μηχανική μάθηση ως πρόβλημα αναζήτησης.Σαλώμη

Γεωργίου. [ONLINE]Available at: [https://docplayer.gr/43000668-Mihaniki-mathisi-i-](https://docplayer.gr/43000668-Mihaniki-mathisi-i-uihaniki-uathisi-os-provliua-anazitisis.html)

[uihaniki-uathisi-os-provliua-anazitisis.html](https://docplayer.gr/43000668-Mihaniki-mathisi-i-uihaniki-uathisi-os-provliua-anazitisis.html)

[Accessed 20 January 2021]

[19]Πανεπιστήμιο Πατρών Πολυτεχνική Σχολή.Εισαγωγή στους γενετικούς

αλγόριθμους.Ευστράτιος Φ.Γεωργόπουλος,Σπυρίδων Δ.Λυκοθανάσης.

[ONLINE]Available at: http://edu.eap.gr/pli/pli31/docs/GAs_introduction.pdf

[Accessed 20 January 2021]

[20]Γενετικοί Αλγόριθμοι.Κεφάλαιο 1 Εισαγωγή.Ιωάννης Γ.Τσούλος.

[ONLINE]Available at: http://users.cs.uoi.gr/~itsoulos/genetic/ga_chapter1.pdf

[Accessed 20 January 2021]

[21]Γενετικοί Αλγόριθμοι.Κεφάλαιο 2 Κωδικοποίηση.Ιωάννης Γ.Τσούλος.

[ONLINE]Available at: http://users.cs.uoi.gr/~itsoulos/genetic/ga_chapter2.pdf

[Accessed 20 January 2021]

[22]Γενετικοί Αλγόριθμοι.Κεφάλαιο 3 Γενετικοί Τελεστές.Ιωάννης Γ.Τσούλος.

[ONLINE]Available at: http://users.cs.uoi.gr/~itsoulos/genetic/ga_chapter3.pdf

[Accessed 20 January 2021]

[23]Γενετικοί Αλγόριθμοι.Κεφάλαιο 5 Γενετικός Προγραμματισμός.Ιωάννης Γ.Τσούλος.

[ONLINE]Available at: http://users.cs.uoi.gr/~itsoulos/genetic/ga_chapter5.pdf

[Accessed 20 January 2021]

[24]Ιστοσελίδα Kaggle.Λήψη δεδομένων για το πρόβλημα. [ONLINE]Available at:

<https://www.kaggle.com/datasets>

[Accessed 20 January 2021]

[25] Δημοφιλή IDE για γλώσσα προγραμματισμού Python – Hello World.Κρητικός

Γεώργιος. [ONLINE]Available at: <https://blogs.sch.gr/kritikog19/2021/01/01/dimofili-ide-gia-glossa-programmatismoy-python-hello-world/>

[Accessed 20 January 2021]