Σχολή Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών

Εθνικό Μετσόβιο Πολυτεχνείο

Διπλωματική Εργασία

Χαράλαμπος Μπεκιάρης

«Συμβολική Παλινδρόμηση»

Επιβλέπων:
Φουσκάκης Δημήτριος

Τριμελής Επιτροπή:

1. Δημήτριος Φουσκάκης, Καθηγητής
2. Μιχαήλ Λουλάκης, Καθηγητής
3. Αντώνης Παπαπαντολέων, Αναπληρωτής Καθηγητής

Αθήνα, Μάιος 2024

Ευχαριστίες

Αρχικά θα πρέπει να ευχαριστήσω τον επιβλέποντα καθηγητή, κ. Φουσκάκη Δημήτριο, για την εμπιστοσύνη, την καθοδήγηση και τον χρόνο που αφιέρωσε για να καταλήξει η εργασία στην τελική μορφή της. Επίσης τον ευχαριστώ για την πολύ καλή συνεργασία που είχαμε όλο αυτό το διάστημα.

Τέλος, δεν γίνεται να παραλείψω τους φίλους που απέκτησα κατά την διάρκεια των σπουδών μου, τις δυσκολίες που περάσαμε μαζί, καθώς και την στήριξη της οικογένειάς μου κατά την διάρκεια των σπουδών μου.

Περίληψη

Η Συμβολική Παλινδρόμηση είναι μία μέθοδος που στοχεύει να ανακτήσει από τα δεδομένα έναν κλειστό τύπο για την περιγραφή τους. Σε αντίθεση με την απλή παλινδρόμηση, όπου το μοντέλο είναι γνωστό, εδώ ο αλγόριθμος επιχειρεί να συνδυάσει απλές συναρτήσεις, καθώς και να εκτιμήσει τις κατάλληλες παραμέτρους του μοντέλου, ώστε αυτό να έχει την καλύτερη δυνατή προσαρμογή.

Στο παρελθόν η συμβολική παλινδρόμηση γινόταν από τους ίδιους τους επιστήμονες, όταν προσπαθούσαν να επεξηγήσουν τις παρατηρήσεις τους μέσω ενός κλειστού μαθηματικού τύπου. Παραδείγματα πρώιμης συμβολικής παλινδρόμησης, υπάρχουν κυρίως στην φυσική, με τον νόμο του Kepler για την κίνηση των πλανητών, καθώς και τον νόμο του Planck. Η διαδικασία βάση της οποίας επιχειρούσαν να καταλήξουν στο τελικό μοντέλο, ήταν αρκετά επίπονη και χρονοβόρα, καθώς έπρεπε να γίνουν πολλές δοκιμές και υποθέσεις προκειμένου να εξηγηθεί πλήρως η συμπεριφορά των παρατηρήσεων.

Στην σύγχρονη εποχή, έχουν αναπτυχθεί τεχνικές μηχανικής μάθησης, με ευρεία εφαρμογή σε προβλήματα ταξινόμησης και παλινδρόμησης. Η συνεισφορά όμως των εν λόγω μοντέλων είναι περιορισμένη, ειδικά σε πεδία της επιστήμης, όπου πολλές φορές δεν επαρκεί απλώς να εκπαιδευτεί ένα μοντέλο το οποίο λειτουργεί σαν μαύρο κουτί, αλλά να προσφέρει περισσότερες πληροφορίες και γνώση στον τελικό χρήστη, σχετικά με τα μοτίβα που παρουσιάζονται στα δεδομένα.

Γι΄ αυτό τα τελευταία χρόνια έχει τραβήξει το ενδιαφέρον η συμβολική παλινδρόμηση, η οποία στηρίζεται σε στοχαστικούς αλγορίθμους για την εξερεύνηση του σύνθετου χώρου των συναρτήσεων. Δύο πρόσφατες υλοποιήσεις παρουσιάζονται στο Κεφάλαιο 2, οι οποίες είναι ανοιχτού κώδικα, και η μία στηρίζεται σε γενετικούς αλγορίθμους, ενώ η άλλη σε Μαρκοβιανές αλυσίδες Μόντε Κάρλο. Συγκεκριμένα η υλοποίηση του γενετικού αλγόριθμου, παρουσιάζεται στο Κεφάλαιο 3, μαζί με τις διαθέσιμες εναλλακτικές επιλογές για κάθε λειτουργία του. Το εν λόγω κεφάλαιο καταλήγει σε μια υλοποίηση γενετικού αλγορίθμου, συνδυάζοντας χαρακτηριστικά από την διαθέσιμη βιβλιογραφία, η οποία στο Κεφάλαιο 4 καλείται να επιλύσει ένα άγνωστο πρόβλημα ταξινόμησης.

Τα αποτελέσματα επιβεβαιώνουν την ανωτερότητα που προσφέρει η παραπάνω μέθοδος. Η επίδοση του μοντέλου ταξινόμησης, με χρήση του γενετικού αλγορίθμου συμβολικής παλινδρόμησης, είναι συγκρίσιμη με αυτή άλλων ταξινομητών, διατηρώντας το πλεονέκτημα της επεξηγισιμότητας και απλότητας, καθώς το τελικό μοντέλο αποτελείται από λιγότερες μεταβλητές.

Περιεχόμενα

1 Αναλυτική Εισαγωγή

1.1 Πρόλογος

Το παρόν κεφάλαιο αποτελεί μία εισαγωγή σε βασικές έννοιες, οι οποίες θα χρειαστούν στα επόμενα κεφάλαια της εργασίας. Αρχικά γίνεται αναφορά στην κλασσική παλινδρόμηση, δηλαδή μοντέλα με γνωστό τύπο, καθώς και τεχνικές εύρεσης των παραμέτρων του μοντέλου. Στην συνέχεια γίνεται μία σύντομη εισαγωγή στους γενετικούς αλγορίθμους και τις σημαντικότερες παραλλαγές τους βάση των γενετικών τελεστών, αλλά και τον τρόπο εξέλιξης του πληθυσμού. Η εκτίμηση των παραμέτρων του μοντέλου, και οι γενετικοί αλγόριθμοι, αποτελούν σημαντικά στοιχεία που θα συνθέσουν τον αλγόριθμο συμβολικής παλινδρόμησης του Κεφαλαίου 3.

1.2 Παλινδρόμηση

Η παλινδρόμηση αποτελεί έναν κλάδο της στατιστικής, με κύριο στόχο τη δημιουργία και τον έλεγχο προβλεπτικών μοντέλων. Στην φυσική οι σχέσεις ανάμεσα στα μεγέθη είναι καθορισμένη από τους νόμους που διέπουν το κάθε φαινόμενο, όπως για παράδειγμα, στο σύστημα μάζας - ελατηρίου - αποσβεστήρα, η σχέση μεταξύ της δύναμης, της θέσης και της ταχύτητας δίνεται από τον τύπο:

$$F = -k \cdot x - c \cdot u$$

όπου k η σταθερά του ελατηρίου και c η σταθερά απόσβεσης.

Με άλλα λόγια, μπορεί κανείς με βεβαιότητα να υπολογίσει την δύναμη F, αρκεί να διαθέτει τις ακριβείς μετρήσεις για την θέση και την ταχύτητα της μάζας.

Στα στατιστικά μοντέλα οι σχέσεις, όχι μόνο δεν είναι αυστηρά καθορισμένες, αλλά υπεισέρχονται τυχαία σφάλματα τα οποία προσδίδουν στοχαστικότητα στα εν λόγω μοντέλα. Επιπλέον, δεν είναι δεδομένο ότι για το μοντέλο που θέλουμε να δημιουργήσουμε διαθέτουμε όλες τις μεταβλητές που απαιτούνται για την πρόβλεψή του.

Προκειμένου να δημιουργηθεί ένα μοντέλο χρειάζονται κάποιες μετρήσεις - παρατηρήσεις ή δεδομένα. Η μεταβλητή που ενδιαφερόμαστε να προβλέψουμε λέγεται μεταβλητή απόκρισης (Y), ενώ αυτές που χρησιμοποιούνται για τον υπολογισμό της πρόβλεψης χαρακτηρίζονται ως επεξηγηματικές $(\boldsymbol{X})$.

Υποθέτουμε λοιπόν ότι έχουμε n παρατηρήσεις y_i κάποιας μεταβλητής απόκρισης Y, με τις αντίστοιχες παρατηρήσεις $\boldsymbol{x_i}$ της επεξηγηματικής μεταβλητής $\boldsymbol{X}$, όπου $i = 1, \ldots, n$. Οι εκτιμήσεις των παρατηρηθέντων y_i με χρήση κάποιου στατιστικού μοντέλου, $\hat{y}_i = g\left(\boldsymbol{x_i}|\boldsymbol{\theta}\right)$, καλούνται προβλεπόμενες ή προσαρμοσμένες τιμές. Η παράμετρος $\boldsymbol{\theta}$ του μοντέλου, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ με $p = dim\left(\boldsymbol{\theta}\right) \geq 1$, συνήθως αναπαριστά τους συντελεστές του μοντέλου. Για παράδειγμα σε μία ευθεία, το $\boldsymbol{\theta}$ είναι η κλίση και το σημείο τομής της με τον άξονα $y'y$.

Μία από τις σημαντικότερες (προ)υποθέσεις που μπορεί να γίνει για το παραπάνω μοντέλο, αφορά τα υπόλοιπα

$$r_i = y_i - \hat{y}_i \tag{1.2.1}$$

ότι ακολουθούν κανονική κατανομή, δηλαδή $r_i \sim N(\mu, \sigma^2)$.

Για την εκτίμηση των παραμέτρων $\boldsymbol{\theta} = (\mu, \sigma^2)$ του μοντέλου, πρέπει να μεγιστοποιηθεί η λογαριθμισμένη πιθανοφάνεια:

$$\boldsymbol{\theta}^* = argmax\left\{L(\boldsymbol{\theta})\right\} \tag{1.2.2}$$

$$L(\boldsymbol{\theta}) = ln\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(r_i-\mu)^2}{2\sigma^2}}\right) \tag{1.2.3}$$

$$= ln\left(\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot e^{-\frac{\sum_{i=1}^{n}(r_i-\mu)^2}{2\sigma^2}}\right) \tag{1.2.4}$$

$$= -n \cdot ln\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (r_i - \mu)^2 . \tag{1.2.5}$$

Η παράμετρος $\boldsymbol{\theta}^*$ καλείται εκτιμητής μέγιστης πιθανοφάνειας. Στην περίπτωση όπου τα υπόλοιπα ακολουθούν κανονική κατανομή με μέση τιμή 0, με άλλα λόγια $r_i \sim N(0, \sigma^2)$:

$$L(\boldsymbol{\theta}) \propto \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 . \tag{1.2.6}$$

Επομένως το πρόβλημα της εύρεσης παραμέτρου $\boldsymbol{\theta}^*$ που να ελαχιστοποιεί την λογαριθμισμένη πιθανοφάνεια, είναι ισοδύναμο με αυτό των ελάχιστων τετραγώνων. Επιπλέον, σε αυτή την περίπτωση λέμε ότι η μεταβλητή απόκρισης ακολουθεί κατά μέσο όρο την συναρτησιακή σχέση $g(\boldsymbol{x}|\boldsymbol{\theta})$, καθώς ισχύει ότι:

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = E[g(\boldsymbol{x}|\boldsymbol{\theta}) + r] \tag{1.2.7}$$

$$= g(\boldsymbol{x}|\boldsymbol{\theta}) . \tag{1.2.8}$$

Η παραπάνω σχέση καλείται «επιφάνεια παλινδρόμησης της Υ επί της $\boldsymbol{X}$» ή «συστηματικό» ή «μη στοχαστικό μέρος του μοντέλου».

Αξίζει να σημειωθεί ότι στην παραπάνω διαδικασία η μορφή της συνάρτησης $g(\boldsymbol{x}|\boldsymbol{\theta})$ είναι γνωστή. Η επιλογή κατάλληλης συνάρτησης είναι καίριας σημασίας για την απόδοση του μοντέλου, και στηρίζεται σε θεωρητικές προσδοκίες ή εμπειρικές γνώσεις σχετικά με τον τρόπο εξάρτησης της μεταβλητής απόκρισης από τις επεξηγηματικές.

1.3 Ελάχιστα Τετράγωνα

Στο αντικείμενο της παρούσας διπλωματικής εργασίας, που είναι η συμβολική παλινδρόμηση, η μορφή της συνάρτησης δεν θεωρείται γνωστή, και θα επιχειρεί ένας αλγόριθμος να την προσδιορίσει.

Στο παρελθόν έχουν κληθεί να επιλύσουν προβλήματα συμβολικής παλινδρόμησης κυρίως φυσικοί ή εφευρέτες, χωρίς όμως η επιτυχία να είναι δεδομένη. Ένα τέτοιο παράδειγμα, αποτελεί ο Γερμανός αστρονόμος Johannes Kepler, ο οποίος αξιοποιώντας τις παρατηρήσεις του Δανού αστρονόμου Tycho Brahe, έπειτα από πολλές αποτυχημένες προσπάθειες κατέληξε το 1604 ότι το μοντέλο που περιγράφει καλύτερα τα δεδομένα της τροχιάς του Άρη, είναι η έλλειψη. Αυτό οδήγησε αργότερα στην διατύπωση του πρώτου νόμου της κίνησης των πλανητών του ηλιακού μας συστήματος, ότι όλοι πλανήτες κινούνται σε ελλειπτική τροχιά γύρω από τον ήλιο.

Στην περίπτωση όπου η μορφή της συνάρτησης είναι γνωστή, ένα πολύτιμο εργαλείο για την εκτίμηση των συντελεστών ή παραμέτρων του μοντέλου, αποτελεί η μέθοδος των ελάχιστων τετραγώνων. Τα υπόλοιπα 1.2.1 της προηγούμενης υποενότητας μπορούν να γραφτούν σε μορφή διανύσματος:

$$\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{1.3.1}$$

$$= \boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}; \boldsymbol{x}) \tag{1.3.2}$$

$$= [r_1, \dots r_n]^T, \quad n \geq 1. \tag{1.3.3}$$

Η σχέση 1.3.1 γίνεται συνάρτηση της άγνωστης παραμέτρου $\boldsymbol{\theta}$, αφού τα παρατηρηθέντα y_i, $\boldsymbol{x_i}$ είναι γνωστά και βάση αυτών ψάχνουμε το $\boldsymbol{\theta}^*$ που θα ελαχιστοποιήσει το άθροισμα των τετραγωνικών υπολοίπων (Sum of Squared Residuals, SSR):

$$SSR = f(\boldsymbol{\theta}) = \|\boldsymbol{r}\|_2^2 \tag{1.3.4}$$

$$= \boldsymbol{r} \cdot \boldsymbol{r}^T \tag{1.3.5}$$

$$= \sum_{i=1}^{n} r_i^2, \tag{1.3.6}$$

όπου $\|\cdot\|_2$ η νόρμα 2 στον $\mathbb{R}^n$.

Εν γένει το πρόβλημα των ελάχιστων τετραγώνων είναι ένα πρόβλημα βελτιστοποίησης, το οποίο διατυπώνεται ως εξής:

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^p}{minimize} \left\{ \|\boldsymbol{r}(\boldsymbol{\theta})\|_2^2 \right\}. \tag{1.3.7}$$

Η συνάρτηση 1.3.4 καλείται αντικειμενική συνάρτηση (objective function). Οι υποψήφιες λύσεις $\boldsymbol{\theta}^*$ χαρακτηρίζονται ως:

- Ολικά ελάχιστα αν και μόνο αν: $f(\boldsymbol{\theta}^*) \leq f(\boldsymbol{\theta})$ για όλα τα $\theta \in \mathbb{R}^n$.
- Τοπικά ελάχιστα αν και μόνο αν: $\exists r > 0$ τέτοιο ώστε $f(\boldsymbol{\theta}^*) \leq f(\boldsymbol{\theta})$ για όλα τα $\boldsymbol{\theta} \in B(\boldsymbol{\theta}^*, r)$.

Η αναγκαία συνθήκη που πρέπει να ικανοποιούν τα παραπάνω $\boldsymbol{\theta}^*$ για να είναι τοπικά ελάχιστα, είναι:

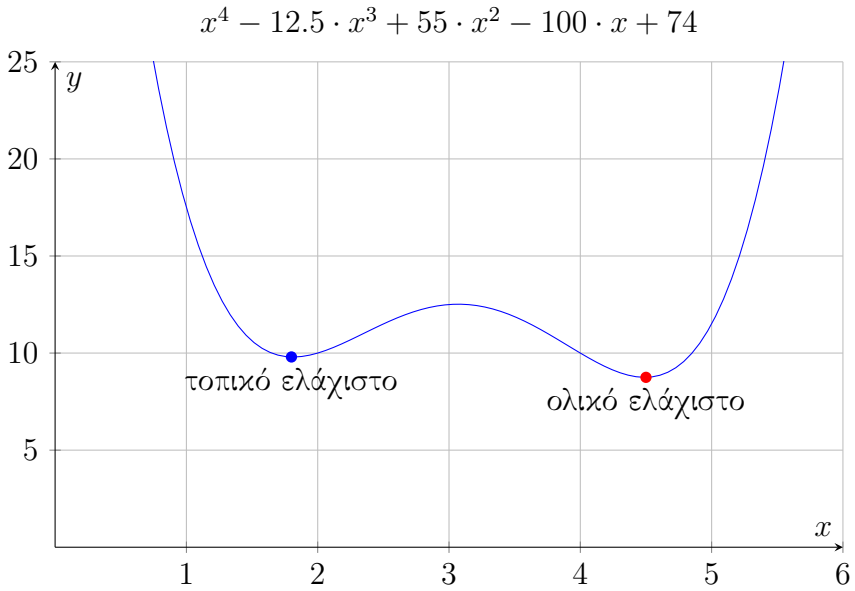
$$\nabla f(\boldsymbol{\theta}^*) = 0 \quad \& \quad \nabla^2 f(\boldsymbol{\theta}^*) \geq 0. \tag{1.3.8}$$

Στην πράξη, όταν αναζητάμε τα εν λόγω σημεία, αναγκαία συνθήκη που πρέπει να ικανοποιείται είναι η εξής:

$$\nabla f(\boldsymbol{\theta}^*) = 0 \quad \& \quad \nabla^2 f(\boldsymbol{\theta}^*) > 0. \tag{1.3.9}$$

1.3.1 Κυρτότητα

Όπως φαίνεται και στο Διάγραμμα 1.1, υπάρχουν διαφορετικά σημεία που πληρούν τις παραπάνω συνθήκες για να είναι ελάχιστα, όμως δεν επαρκούν στην περίπτωση που ενδιαφερόμαστε στην εύρεση του ολικού ελάχιστου.



Διάγραμμα 1.1: Γράφημα συνάρτησης με δύο ακρότατα, τοπικό & ολικό ακρότατο.

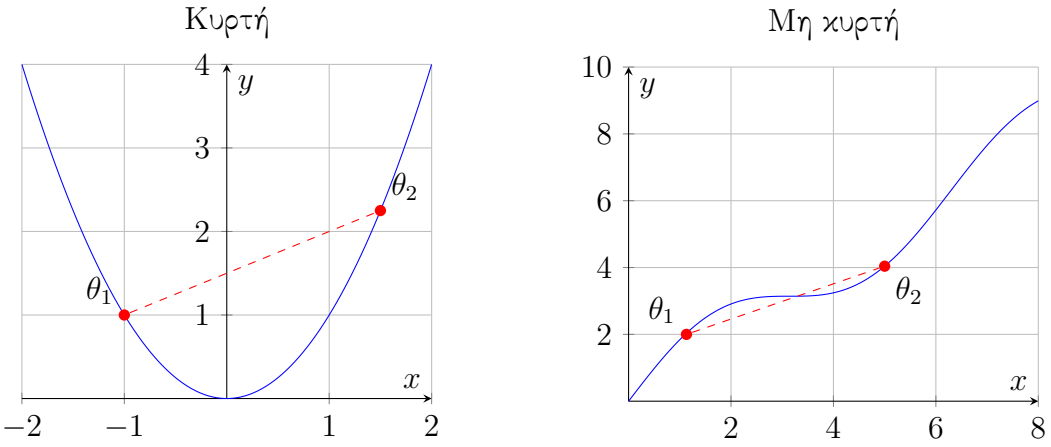
Για τον σκοπό αυτό, αποκτά ιδιαίτερο ενδιαφέρον η μελέτη των προβλημάτων ελάχιστων τετραγώνων, όπου η αντικειμενική συνάρτηση είναι κυρτή. Σε αυτή την περίπτωση, αποδεικνύεται ότι οι λύσεις του προβλήματος είναι ολικά ελάχιστα. Επομένως η κυρτότητα αποτελεί

έναν τρόπο χαρακτηρισμού των εν λόγω προβλημάτων βελτιστοποίησης ως «εύκολα», αφού δεν υπάρχουν τοπικά ελάχιστα στα οποία μπορεί να εγκλωβιστεί ο αλγόριθμος επίλυσης.

Μια συνάρτηση $f(\boldsymbol{\theta})$ με $f : \mathbb{R}^p \to \mathbb{R}$ λέγεται κυρτή αν και μόνο αν (τα παρακάτω είναι ισοδύναμα):

$$\begin{cases} \forall \boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \mathbb{R}^p \text{ και } u \in [0,1] : & f(u \cdot \boldsymbol{\theta_1} + (1-u) \cdot \boldsymbol{\theta_2}) \leq u \cdot f(\boldsymbol{\theta_1}) + (1-u) \cdot f(\boldsymbol{\theta_2}). \\ \forall \boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \mathbb{R}^p : & f(\boldsymbol{\theta_2}) \geq f(\boldsymbol{\theta_1}) + \nabla f(\boldsymbol{\theta_1})^T \cdot (\boldsymbol{\theta_2} - \boldsymbol{\theta_1}). \\ \forall \boldsymbol{\theta} \in \mathbb{R}^p : & \nabla^2 f(\boldsymbol{\theta}) \geq 0. \end{cases}$$

Η γραφική ερμηνεία του παραπάνω ορισμού, φαίνεται στο Διάγραμμα 1.2. Για να είναι κυρτή μια συνάρτηση, θα πρέπει για οποιαδήποτε δύο σημεία επιλέξουμε, η ευθεία που τα ενώνει να βρίσκεται πάνω από το γράφημα της συνάρτησης.



Διάγραμμα 1.2: Γράφημα κυρτής και μη κυρτής συνάρτησης.

1.3.2 Γραμμικά ελάχιστα τετράγωνα

Ένα σημαντικό κομμάτι της ανάλυσης παλινδρόμησης αφορά τα γραμμικά μοντέλα, ή τα μοντέλα που ανάγονται σε γραμμικά. Τότε, το μη στοχαστικό κομμάτι του μοντέλου παλινδρόμησης 1.2.8 είναι της μορφής:

$$\begin{aligned} g(\boldsymbol{x}|\boldsymbol{\theta}) &= \boldsymbol{x}^T \boldsymbol{\theta} \\ &= \theta_0 + \theta_1 x_1 + \ldots + \theta_{p-1} x_{p-1} \end{aligned}$$

όπου $\boldsymbol{\theta} = [\theta_0, \ldots, \theta_{p-1}]^T \in \mathbb{R}^p$ και $\boldsymbol{x} = [1, x_1, \ldots, x_{p-1}]^T \in \mathbb{R}^p$.

Σε αυτή την περίπτωση, τα υπόλοιπα 1.3.1 είναι γραμμικά, και το πρόβλημα των ελάχιστων τετραγώνων 1.3.7 καλείται γραμμικό. Πιο συγκεκριμένα έχουμε ότι:

$$\begin{aligned}\boldsymbol{r}(\boldsymbol{\theta}) &= \boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}; \boldsymbol{x}) \\ &= \boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{\theta}.\end{aligned}$$

Επομένως τα υπόλοιπα είναι της μορφής:

$$\boldsymbol{r}(\boldsymbol{\theta}) = A\boldsymbol{\theta} - b \tag{1.3.10}$$

όπου $A = -\boldsymbol{x}^T$ και $b = -\boldsymbol{y}$. Εν συνεχεία, αντικαθιστώντας τα παραπάνω γραμμικά υπόλοιπα 1.3.10 στην αντικειμενική συνάρτηση των τετραγωνικών υπολοίπων 1.3.4, λαμβάνουμε τα γραμμικά ελάχιστα τετράγωνα:

$$\begin{aligned}f(\boldsymbol{\theta}) &= (A\boldsymbol{\theta} - b)(A\boldsymbol{\theta} - b)^T \\ &= (A\boldsymbol{\theta})^T A\boldsymbol{\theta} - 2(A\boldsymbol{\theta})^T b + bb^T.\end{aligned}$$

Τα σημεία $\boldsymbol{\theta}^*$ που ελαχιστοποιούν το άθροισμα των τετραγωνικών υπολοίπων πρέπει να ικανοποιούν τις αναγκαίες συνθήκες 1.3.9:

$$\nabla f(\boldsymbol{\theta}^*) = 0 \rightarrow \left(A^T A\right)\boldsymbol{\theta}^* = A^T b \tag{1.3.11}$$

$$\nabla^2 f(\boldsymbol{\theta}^*) > 0 \rightarrow 2A^T A > 0. \tag{1.3.12}$$

Παρατηρούμε ότι ο πίνακας $A^T A$ είναι θετικά ορισμένος, αφού για κάθε διάνυσμα $\boldsymbol{u}$:

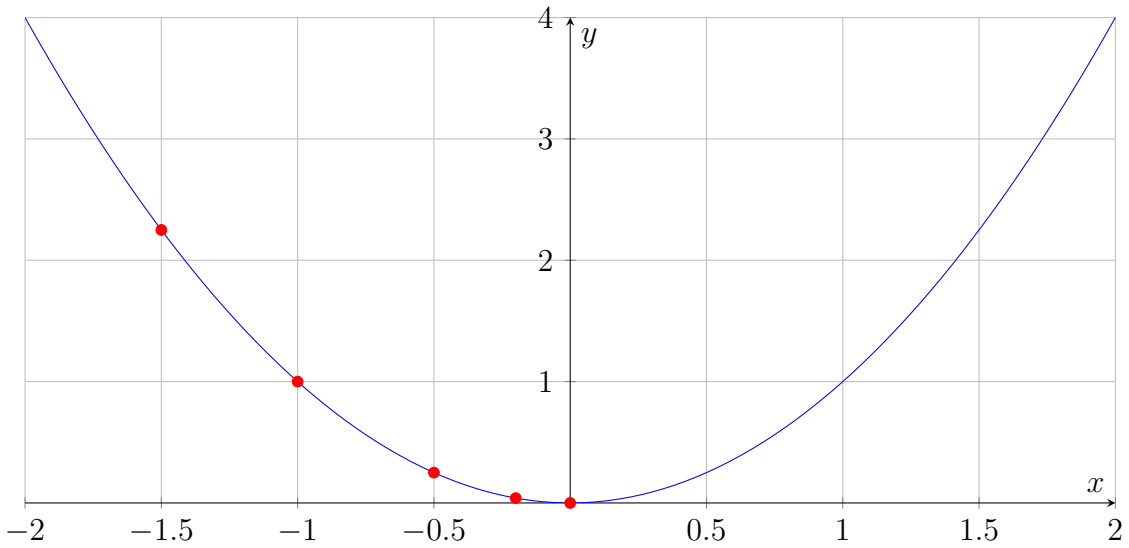
$$\boldsymbol{u}^T A^T A\boldsymbol{u} = (A\boldsymbol{u})^T(A\boldsymbol{u}) = \|A\boldsymbol{u}\|_2^2 \geq 0.$$

Επίσης για να έχει λύση το σύστημα 1.3.11, ο πίνακας $A^T A$ πρέπει να είναι αντιστρέψιμος, δηλαδή να είναι πλήρης τάξης ως προς τις στήλες. Τότε $\|A\boldsymbol{u}\|_2^2 > 0$ για κάθε $\boldsymbol{u} \neq \boldsymbol{0}$, άρα η συνθήκη 1.3.12 ικανοποιείται πάντα. Συνεπώς, η αντικειμενική συνάρτηση είναι κυρτή, και η λύση που προκύπτει λύνοντας το σύστημα 1.3.11 είναι ολικό ελάχιστο του προβλήματος των ελάχιστων τετραγώνων. Το σύστημα 1.3.11 αποτελείται από γραμμικές εξισώσεις οι οποίες καλούνται κανονικές εξισώσεις, και η λύση $\boldsymbol{\theta}^* = \left[\theta_0^*, \ldots, \theta_{p-1}^*\right]^T$ καλείται εκτιμητής ελάχιστων τετραγώνων.

Το εν λόγω γραμμικό σύστημα επιλύεται χωρίς να αντιστραφεί ο πίνακας $A^T A$, καθώς δεν είναι υπολογιστικά αποδοτικό και σε περίπτωση που ο πίνακας έχει κακό δείκτης κατάστασης, η λύση θα είναι ανακριβής. Συνήθως η επίλυση του συστήματος γίνεται με ορθογώνιους μετασχηματισμούς.

1.3.3 Μη γραμμικά ελάχιστα τετράγωνα

Η επίλυση των ελάχιστων τετραγώνων στην περίπτωση που η συνάρτηση των υπολοίπων δεν είναι γραμμική, αποτελεί ένα αρκετά πιο περίπλοκο πρόβλημα, σε σχέση με αυτό της προηγούμενης υποενότητας. Στα μη γραμμικά ελάχιστα τετράγωνα, η συνάρτηση των υπολοίπων 1.3.1 προσεγγίζεται τοπικά με την βοήθεια του αναπτύγματος Taylor. Κατόπιν, βάση ενός αλγορίθμου υπολογίζεται το τοπικό ελάχιστο μέσα από μία επαναληπτική διαδικασία, ξεκινώντας από ένα αρχικό σημείο (βλ. Διάγραμμα 1.3).



Διάγραμμα 1.3: Γράφημα της συνάρτησης $f(x) = x^2$ (μπλε γραμμή). Επαναληπτικά βήματα προσέγγισης του τοπικού ελάχιστου (κόκκινες κουκίδες).

Προσέγγιση πρώτης τάξης

Η προσέγγιση πρώτης τάξης, στην πραγματικότητα γραμμικοποιεί τοπικά την συνάρτηση των υπολοίπων και στην συνέχεια το πρόβλημα λύνεται τοπικά ως πρόβλημα γραμμικών τετραγώνων. Πιο συγκεκριμένα, για κάθε μια παρατήρηση $\boldsymbol{x_i}$ έχουμε το αντίστοιχο υπόλοιπο:

$$r_i(\boldsymbol{\theta}) = y_i - g(x_i; \boldsymbol{\theta}).$$

Αναπτύσσοντας κατά Taylor, γύρω από το αρχικό σημείο $\boldsymbol{\theta_0}$, έχουμε την ζητούμενη προσέγγιση πρώτης τάξης:

$$r_i(\boldsymbol{\theta}) \approx r_i(\boldsymbol{\theta_0}) + \nabla r_i(\boldsymbol{\theta_0})^T (\boldsymbol{\theta} - \boldsymbol{\theta_0}) \quad \forall i = 1, \ldots, n. \tag{1.3.13}$$

Ισοδύναμα, τα παραπάνω υπόλοιπα γράφονται για όλα τα i, σε μορφή ενός διανύσματος:

$$\boldsymbol{r}(\boldsymbol{\theta}) \approx \boldsymbol{r}(\boldsymbol{\theta_0}) + J(\boldsymbol{\theta_0})(\boldsymbol{\theta} - \boldsymbol{\theta_0}), \tag{1.3.14}$$

όπου ο πίνακας $J \in \mathbb{R}^{p\times n}$ είναι η Ιακωβιανή. Αλλάζοντας την μεταβλητή $\boldsymbol{\theta} = \boldsymbol{\theta_0} + \boldsymbol{d}$, λαμβάνουμε την τελική μορφή:

$$\boldsymbol{r}\left(\boldsymbol{\theta_0} + \boldsymbol{d}\right) \approx \boldsymbol{r}\left(\boldsymbol{\theta_0}\right) + J\left(\boldsymbol{\theta_0}\right)\boldsymbol{d}. \tag{1.3.15}$$

Αντικαθιστώντας τώρα την προσέγγιση πρώτης τάξης 1.3.15 στην αντικειμενική συνάρτηση του αρχικού προβλήματος των ελάχιστων τετραγώνων 1.3.4, το εν λόγω πρόβλημα επαναδιατυπώνεται ως:

$$\underset{\boldsymbol{d}\in\mathbb{R}^p}{minimize}\left\{\left\|\boldsymbol{r}\left(\boldsymbol{\theta_0} + \boldsymbol{d}\right)\right\|_2^2\right\}. \tag{1.3.16}$$

Επιπλέον, εφόσον η αντικειμενική συνάρτηση είναι γραμμική τότε το παραπάνω πρόβλημα επιλύεται αντίστοιχα με αυτό των γραμμικών ελάχιστων τετραγώνων της προηγούμενης ενότητας. Πιο συγκεκριμένα, για $A = J\left(\boldsymbol{\theta_0}\right)$ και $b = -\boldsymbol{r}\left(\boldsymbol{\theta_0}\right)$, επιλύοντας το γραμμικό σύστημα:

$$\left(J\left(\boldsymbol{\theta_0}\right)^T J\left(\boldsymbol{\theta_0}\right)\right)\boldsymbol{d} = -J\left(\boldsymbol{\theta_0}\right)^T \boldsymbol{r}\left(\boldsymbol{\theta_0}\right) \tag{1.3.17}$$

προκύπτει η τιμή του $\boldsymbol{d}$ που ελαχιστοποιεί το άθροισμα των τετραγωνικών υπολοίπων γύρω από το $\boldsymbol{\theta_0}$. Ενημερώνοντας την τιμή του $\boldsymbol{\theta_0} \leftarrow \boldsymbol{\theta_0} + \boldsymbol{d}$, επαναλαμβάνουμε την παραπάνω διαδικασία επαναληπτικά μέχρι να συγκλίνει σε κάποια τιμή $\boldsymbol{\theta^*}$.

Η παραπάνω διαδικασία είναι γνωστή και ως ο αλγόριθμος των Gauss-Newton. Φυσικά το όνομα του αλγορίθμου προέρχεται από τους γνωστούς μαθηματικούς Carl Friedrich Gauss και Isaac Newton και πρωτοεμφανίσθηκε το 1809. Στο Διάγραμμα 1.4, παρουσιάζεται το διάγραμμα ροής του αλγορίθμου. Ενδεικτικά, ο αλγόριθμος τερματίζει όταν η μεγαλύτερη κατ΄ απόλυτο τιμή μεταβολή σε κάποια συνιστώσα $\boldsymbol{\theta_0}$ είναι μικρότερη από κάποιο προκαθορισμένο όριο ϵ. Εναλλακτικά κριτήρια τερματισμού, αποτελούν ο μέγιστος αριθμός επαναλήψεων, η ελάχιστη ποσοστιαία μεταβολή της αντικειμενικής συνάρτησης, ή συνδυασμός των παραπάνω.

Προσέγγιση δεύτερης τάξης

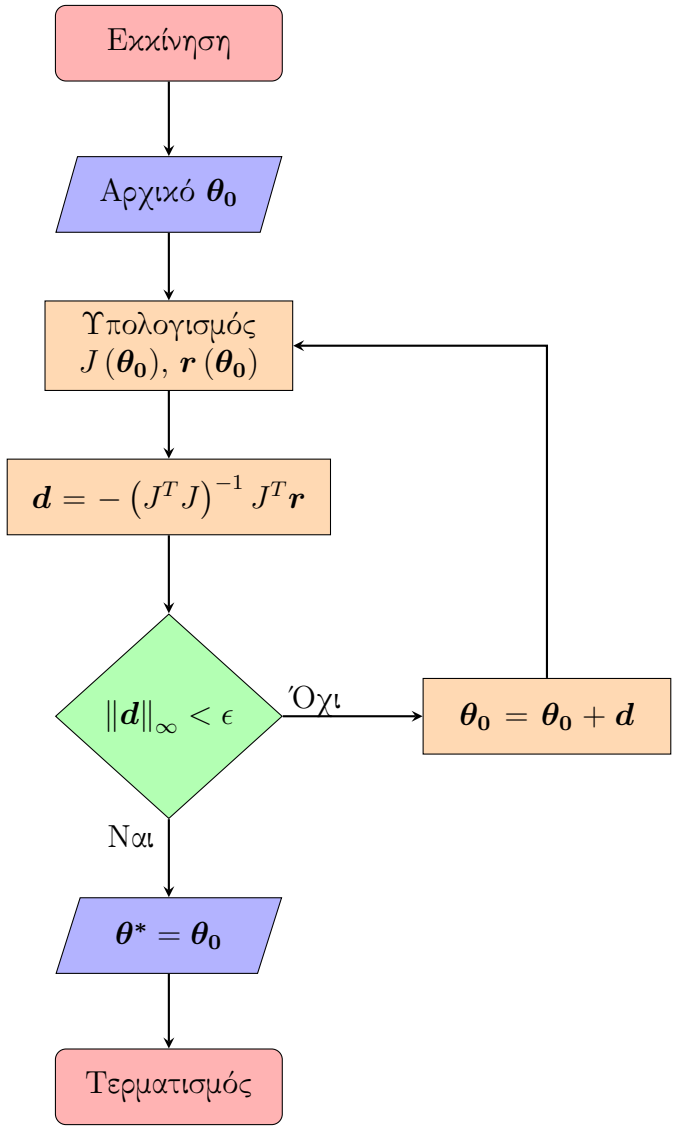
Σε αντίθεση με την πρώτης τάξης προσέγγιση, εδώ η αντικειμενική συνάρτηση 1.3.4 προσεγγίζεται τοπικά από μία τετραγωνική συνάρτηση. Η εν λόγω προσέγγιση, προκύπτει από το ανάπτυγμα κατά Taylor γύρω από το σημείο $\boldsymbol{\theta_0}$.

$$f\left(\boldsymbol{\theta}\right) \approx f\left(\boldsymbol{\theta_0}\right) + \nabla f\left(\boldsymbol{\theta_0}\right)^T\left(\boldsymbol{\theta} - \boldsymbol{\theta_0}\right) + \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta_0}\right)^T \nabla^2 f\left(\boldsymbol{\theta_0}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta_0}\right).$$

Εφαρμόζοντας την ίδια αλλαγή μεταβλητής, $\boldsymbol{\theta} = \boldsymbol{\theta_0} + \boldsymbol{d}$, λαμβάνουμε την τελική μορφή της προσέγγισης της αντικειμενικής συνάρτησης:

$$f\left(\boldsymbol{\theta_0} + \boldsymbol{d}\right) \approx f\left(\boldsymbol{\theta_0}\right) + \nabla f\left(\boldsymbol{\theta_0}\right)^T \boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^T \nabla^2 f\left(\boldsymbol{\theta_0}\right)\boldsymbol{d}, \tag{1.3.18}$$

όπου:



Διάγραμμα 1.4: Διάγραμμα ροής για τον αλγόριθμο των Gauss-Newton.

$$\nabla f(\boldsymbol{\theta_0}) = 2J(\boldsymbol{\theta_0})^T \boldsymbol{r}(\boldsymbol{\theta_0}) \tag{1.3.19}$$

$$\nabla^2 f(\boldsymbol{\theta_0}) = 2J(\boldsymbol{\theta_0})^T J(\boldsymbol{\theta_0}) + 2\sum_{i=1}^{n} r_i(\boldsymbol{\theta_0}) \nabla^2 r_i(\boldsymbol{\theta_0}). \tag{1.3.20}$$

Άρα το πρόβλημα 1.3.7, γράφεται ως:

$$\underset{\boldsymbol{d}\in\mathbb{R}^p}{minimize} \{f(\boldsymbol{\theta_0} + \boldsymbol{d})\}. \tag{1.3.21}$$

Οι ζητούμενες λύσεις πρέπει να ικανοποιούν τις συνθήκες 1.3.9. Πιο συγκεκριμένα, από

την πρώτη συνθήκη, και με χρήση της προσέγγισης 1.3.18, έχουμε ότι:

$$\nabla f\left(\boldsymbol{\theta_0}+\boldsymbol{d}\right)=0\rightarrow\nabla f\left(\boldsymbol{\theta_0}\right)^T+\nabla^2 f\left(\boldsymbol{\theta_0}\right)\boldsymbol{d}=0.$$

Στην συνέχεια αντικαθιστώντας τα 1.3.19 & 1.3.20 στην παραπάνω σχέση, προκύπτει το τελικό γραμμικό σύστημα που επιλύει τοπικά γύρω από το $\boldsymbol{\theta_0}$ το πρόβλημα ελάχιστων τετραγώνων 1.3.21:

$$\left(J\left(\boldsymbol{\theta_0}\right)^T J\left(\boldsymbol{\theta_0}\right)+\sum_{i=1}^{n} r_i\left(\boldsymbol{\theta_0}\right)\nabla^2 r_i\left(\boldsymbol{\theta_0}\right)\right)\boldsymbol{d}=-J\left(\boldsymbol{\theta_0}\right)^T\boldsymbol{r}\left(\boldsymbol{\theta_0}\right). \tag{1.3.22}$$

Το πρόβλημα των ελάχιστων τετραγώνων με αυτή την διαδικασία, λύνεται επαναληπτικά, όπως και στην προσέγγιση πρώτης τάξης. Ο αλγόριθμος επίλυσης ξεκινάει από μία αρχική εκτίμηση $\boldsymbol{\theta_0}$, και στην συνέχεια λύνοντας το γραμμικό σύστημα 1.3.22 ενημερώνεται η τιμή του $\boldsymbol{\theta}$. Η διαδικασία επαναλαμβάνεται μέχρι να συγκλίνει στην λύση του αρχικού προβλήματος 1.3.7. Ο παραπάνω αλγόριθμος καλείται αλγόριθμος του Newton, και στο Διάγραμμα 1.5 παρουσιάζεται το αντίστοιχο διάγραμμα ροής.

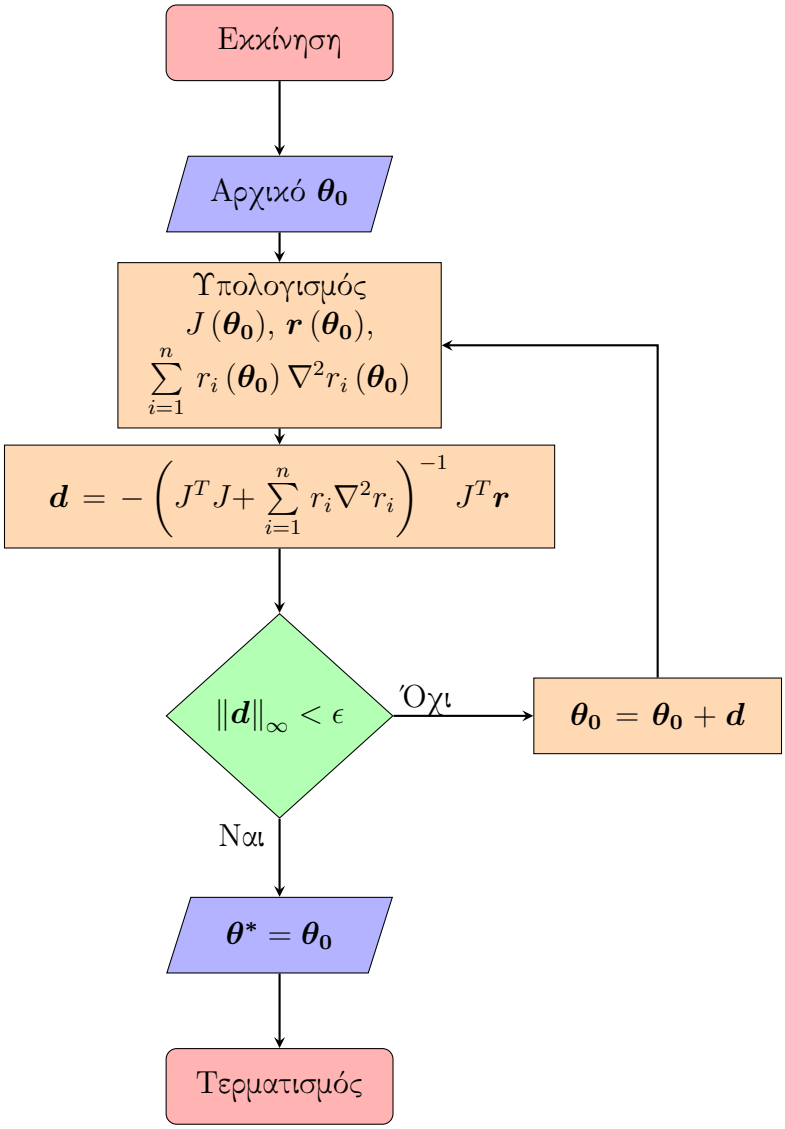
Το υπολογιστικό κόστος της προσέγγισης δεύτερης τάξης είναι μεγαλύτερο, καθώς απαιτούνται παραπάνω πράξεις για την προσέγγιση του Εσσιανού της αντικειμενικής συνάρτησης. Επιπλέον απαιτεί την ύπαρξη 2ης παραγώγου στα υπόλοιπα, και κατ΄ επέκταση στο μοντέλο το οποίο θέλουμε να εκτιμήσουμε τις παραμέτρους τους. Για τους παραπάνω λόγους, η εν λόγω μέθοδος μπορεί να εφαρμοστεί σε περιορισμένο αριθμό προβλημάτων.

Συνοψίζοντας, για τις δύο προσεγγίσεις πρώτης και δεύτερη τάξης, οι παρακάτω διαφορές είναι εμφανής (Esposito & Floudas, 2009):

- Μέθοδος Gauss-Newton (1ης τάξης):
  - Απαιτείται μόνο ο υπολογισμός της πρώτης παραγωγού.
  - Η προσέγγιση του Εσσιανού είναι θετικά ορισμένη και λόγω συμμετρίας το υπολογιστικό κόστος της αντιστροφής είναι μικρότερο.
  - Η προσέγγιση είναι ακριβής μόνο αν τα υπόλοιπα τείνουν στο μηδέν κοντά στην λύση.
- Μέθοδος Newton (2ης τάξης):
  - Η σύγκλιση είναι τετραγωνική.
  - Απαιτείται ο υπολογισμός και η αντιστροφή του Εσσιανού αυξάνοντας το υπολογιστικό κόστος.
  - Αν ο Εσσιανός δεν είναι θετικά ορισμένος τότε δεν θα συγκλίνει η μέθοδος.

Το κοινό των δύο περιπτώσεων είναι ότι σε περιοχές μακρυά από το ελάχιστο (δηλαδή με κακή αρχική συνθήκη) δεν εγγυάται ότι θα συγκλίνει η εκάστοτε μέθοδος σε ελάχιστο. Ο

αλγόριθμος του Gauss-Newton, προτιμάται σε προβλήματα όπου δεν υπάρχει μεγάλος θόρυβος στα δεδομένα και αναμένεται να συγκλίνει σε παρόμοιο αριθμό βημάτων από την μέθοδο Newton. Στην αντίθετη περίπτωση η μέθοδος Newton, έχει ταχύτερη σύγκλιση, όμως ο χρόνος που θα χρειαστεί, λόγω αυξημένου υπολογιστικού κόστους, θα είναι παρόμοιος με την μέθοδο Gauss-Newton.



Διάγραμμα 1.5: Διάγραμμα ροής για τον αλγόριθμο του Newton.

1.4 Γενετικοί Αλγόριθμοι

Η συμβολική παλινδρόμηση αποτελεί ένα πρόβλημα βελτιστοποίησης, όπου ο χώρος μεγαλώνει εκθετικά, αυξάνοντας το πλήθος των τελεστών που ενδέχεται να περιγράφουν τα δεδομένα, καθώς οι συνδυασμοί που μπορούν να γίνουν είναι αμέτρητοι. Οι γενετικοί αλγόριθμοι, αποτελούν μία μέθοδο στοχαστικής βελτιστοποίησης, η οποία μπορεί να λύσει αρκετά περίπλοκα προβλήματα, χωρίς να απαιτούνται ιδιαίτερες προϋποθέσεις. Η ιδέα πίσω

από την υλοποίηση του γενετικού αλγορίθμου, προκύπτει από την θεωρία εξέλιξης του Δαρβίνου.

Οι γενετικοί αλγόριθμοι ανήκουν σε μία ευρύτερη κατηγορία αλγορίθμων, τους επονομαζόμενους εξελικτικούς, καθώς στηρίζονται σε έναν πληθυσμό ο οποίος εξελίσσεται από γενιά σε γενιά. Ο πληθυσμός αποτελείται από υποψήφιες λύσεις του προβλήματος βελτιστοποίησης, οι οποίες καλούνται χρωμοσώματα, ενώ οι πληροφορίες που κωδικοποιεί το κάθε χρωμόσωμα ονομάζονται γονίδια. Παρά τα στοχαστικά χαρακτηριστικά της εν λόγω μεθόδου βελτιστοποίησης, το γεγονός ότι τα «ικανότερα» χρωμοσώματα επιβιώνουν στον πληθυσμό με την πάροδο των γενεών, εγγυάται ότι σε κάθε επανάληψη η ποιότητα της υποψήφιας λύσης διατηρείται ή βελτιώνεται. Ο χαρακτηρισμός ως «ικανότερα», προκύπτει από μια προκαθορισμένη αντικειμενική συνάρτηση, η οποία μπορεί να διαφέρει από πρόβλημα σε πρόβλημα.

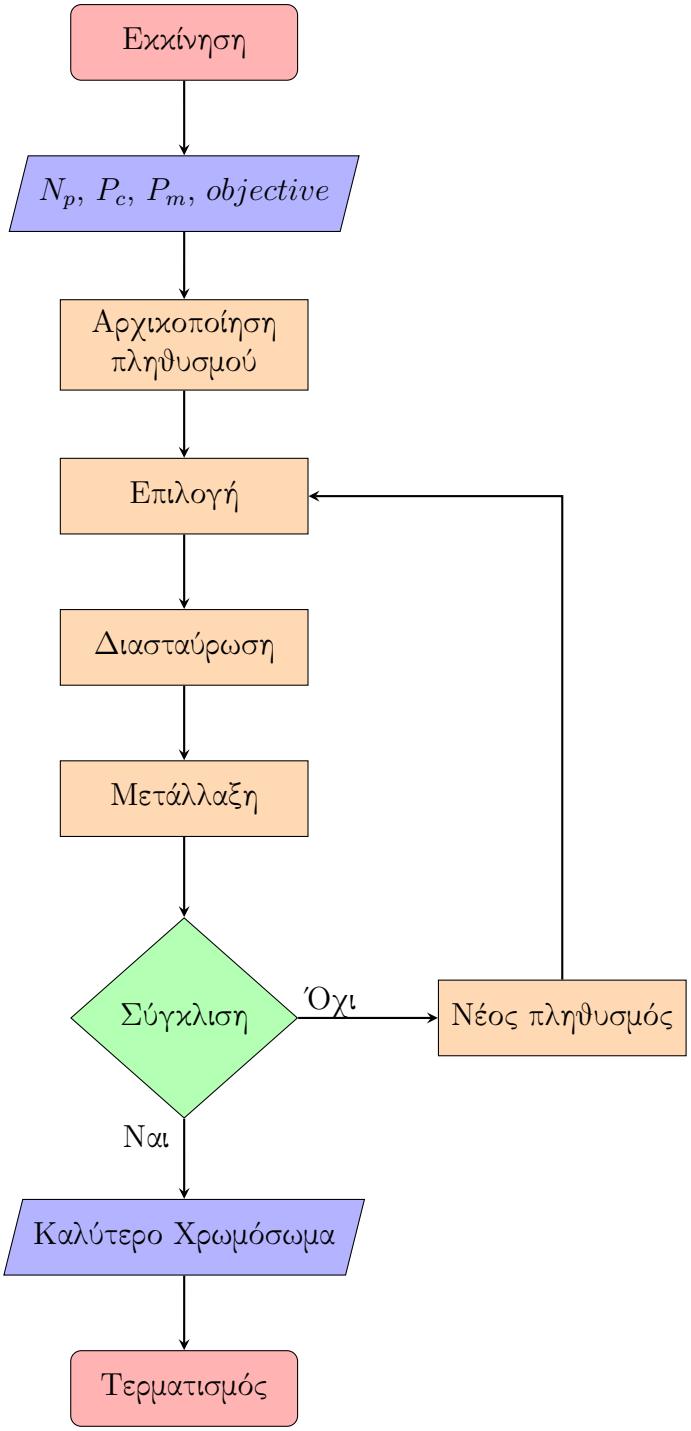
Ο πληθυσμός του γενετικού αλγορίθμου συνήθως αρχικοποιείται με τυχαίο τρόπο, ώστε να υπάρχει ποικιλομορφία μεταξύ των χρωμοσωμάτων, και να εξερευνηθεί μεγαλύτερο κομμάτι του χώρου των πιθανών λύσεων. Σημαντικό χαρακτηριστικό στην εξερεύνηση του χώρου και την βελτίωση των πιθανών λύσεων, αποτελούν οι γενετικοί τελεστές που δρουν πάνω στον πληθυσμό σε κάθε γενιά (βλ. Διάγραμμα 1.6):

- Επιλογή (Selection).
- Διασταύρωση (Crossover).
- Μετάλλαξη (Mutation).

Οι δημοφιλέστεροι τρόποι κωδικοποίησης των γονιδίων του χρωμοσώματος, είναι σε δυαδική ή πραγματική μορφή. Στην πρώτη περίπτωση, τα γονίδια απεικονίζονται σε μια συμβολοσειρά από 0 και 1 (bits). Η δυαδική κωδικοποίηση έχει το μειονέκτημα του αυξημένου υπολογιστικού κόστους, όταν πρέπει να κωδικοποιηθούν άλλοι τύποι δεδομένων, όπως πραγματικοί αριθμοί, σε δυαδική μορφή. Επιπλέον, οι τροποποιήσεις των γονιδίων γίνονται αλλάζοντας τα bits, με συνέπεια η αναζήτηση του χώρου των λύσεων να περιορίζεται σε γειτονικές περιοχές, εγκλωβίζοντας τον αλγόριθμο σε τοπικά ελάχιστα. Κλείνοντας, ένας ακόμη περιοριστικός παράγοντας αποτελεί η συμβολοσειρά των 0 και 1 αυτή καθαυτή. Το μήκος της είναι πεπερασμένο και ενδέχεται να μην επαρκεί για να κωδικοποιήσει όλη την πληροφορία των γονιδίων.

Παρόλα αυτά η εν λόγω κωδικοποίηση αποδίδει όταν οι μεταβλητές του προβλήματος είναι οι ίδιες σε δυαδική μορφή, όπως για παράδειγμα οι μεταβλητές απόφασης (1: Ναι, 0: Όχι). Κλασσική εφαρμογή του γενετικού αλγορίθμου με δυαδική κωδικοποίηση, είναι στο πρόβλημα επιλογής μεταβλητών στην πολλαπλή γραμμική παλινδρόμηση, όπου η απόφαση για το αν μια μεταβλητή θα συμπεριληφθεί στο μοντέλο κωδικοποιείται με 0 ή 1 αντίστοιχα.

Η κωδικοποίηση των γονιδίων με πραγματικούς αριθμούς, λύνει αρκετά από τα προβλήματα της προηγούμενης επιλογής. Αρχικά, αφού το κάθε γονίδιο αναπαριστάται με πραγματικό



Διάγραμμα 1.6: Διάγραμμα ροής γενετικού αλγορίθμου.

αριθμό, είναι ικανό να περιγράψει πολλές περισσότερες καταστάσεις, διατηρώντας σταθερό το μέγεθος του χρωμοσώματος. Δηλαδή το χρωμόσωμα μπορεί με μεγαλύτερη ακρίβεια να κωδικοποιήσει την πληροφορία, όσο περίπλοκος και αν είναι ο χώρος των πιθανών λύσεων. Επιπλέον, η αναζήτηση στον εν λόγω χώρο γίνεται με ομοιόμορφο τρόπο, αφού αποφεύγονται τυχόν σφάλματα διακριτοποίησης που θα προέκυπταν αν μια συνεχής μεταβλητή έπρεπε να απεικονιστεί με διακριτό τρόπο σε δυαδική μορφή. Τέλος, προσφέρει μεγαλύτερη ευελιξία στην υλοποίηση των γενετικών τελεστών, επιτρέποντας στον γενετικό αλγόριθμο να λύσει μεγαλύτερη ποικιλία προβλημάτων.

1.4.1 Επιλογή χρωμοσωμάτων

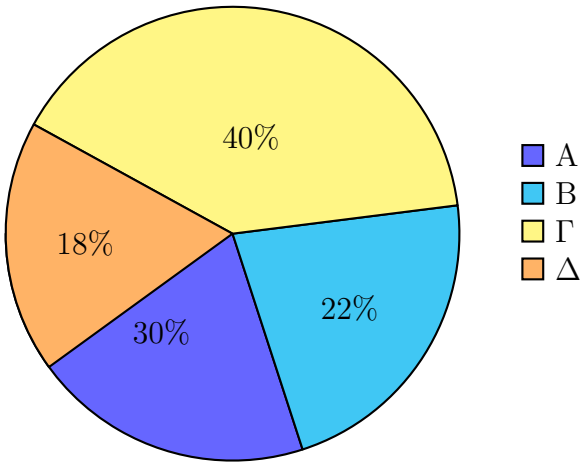
Ο τελεστής της επιλογής, έχει ως στόχο να διαλέξει υποψήφιες λύσεις από τον πληθυσμό για διασταύρωση. Στην φύση η διαδικασία της επιλογής αποτελεί σημαντικό χαρακτηριστικό για την επιβίωση του είδους, καθώς οι «ικανότεροι» οργανισμοί είναι πιθανότερο να αφήσουν απογόνους που θα φέρουν «ικανά» γονίδια σε μεγαλύτερο κομμάτι του πληθυσμού. Ο τελεστής επιλογής στον γενετικό αλγόριθμο, εμπνέεται από την φυσική επιλογή, και οι συχνότερες υλοποιήσεις είναι οι ακόλουθες:

1. **Επιλογή Ρουλέτας:** Στην επιλογή ρουλέτας, κάθε υποψήφια λύση αντιστοιχίζεται με μία πιθανότητα επιλογής για διασταύρωση. Πιο συγκεκριμένα, η πιθανότητα αυτή είναι ανάλογη της «καλής προσαρμογής» φ_i που έχει ο εν λόγω υποψήφιος i στον πληθυσμό. Η προσαρμογή υπολογίζεται μέσω της αντικειμενικής συνάρτησης $f(\boldsymbol{\theta})$, και η ζητούμενη πιθανότητα για τον i-οστό υποψήφιο σε πληθυσμό μεγέθους N_p υπολογίζεται ως:

$$p_i = \frac{\varphi_i}{\sum_{i=1}^{N_p} \varphi_i} \quad \forall i = 1, \ldots, N_p. \tag{1.4.1}$$

   Η ποσότητα 1.4.1 είναι φραγμένη στο $[0, 1]$ και δίνει μεγαλύτερη πιθανότητα στους υποψηφίους με καλύτερη προσαρμογή στην αντικειμενική συνάρτηση. Η τιμή της καλής προσαρμογής φ_i προκύπτει άμεσα από την αντικειμενική συνάρτηση ή έμμεσα όταν για παράδειγμα η τελευταία λαμβάνει και αρνητικές τιμές. Αξίζει να σημειωθεί ότι λιγότερο «ικανοί» υποψήφιοι εξακολουθούν να έχουν την δυνατότητα να επιλεγούν για διασταύρωση. Με αυτόν τον τρόπο διατηρείται η ποικιλομορφία στις λύσεις και αποφεύγεται η πρόωρη σύγκλιση σε κάποιο τοπικό ελάχιστο. Κλείνοντας, στο Διάγραμμα 1.7 οι υποψήφιοι Β, Δ έχουν όμοια πιθανότητα επιλογής, το οποίο υποδεικνύει ότι έχουν παρόμοια γονίδια. Το πρόβλημα που προκύπτει είναι στην περίπτωση που επιλεχθούν και οι δύο για διασταύρωση, θα παράξουν περισσότερες λύσεις που θα μοιράζονται τα ίδια γονίδια, μειώνοντας έτσι την ποικιλομορφία του πληθυσμού.

2. **Επιλογή βάση τάξης:** Οι υποψήφιες λύσεις στον πληθυσμό ταξινομούνται κατά φθίνουσα σειρά βάση της αντικειμενικής συνάρτησης, και αντιστοιχίζεται σε κάθε υ-



Διάγραμμα 1.7: Τομεόγραμμα πιθανότητας επιλογής ρουλέτας τεσσάρων υποψηφίων.

ποψήφιο η τάξη του (rank). Για παράδειγμα, ο πρώτος καλύτερος έχει τάξη N_p, ο δεύτερος τάξη $N_p - 1$, κ.ο.κ. Τότε η πιθανότητα επιλογής δίνεται από την σχέση:

$$p_i = \frac{rank_i}{\sum_{i=1}^{N_p} rank_i} \quad \forall i = 1, \ldots, N_p. \tag{1.4.2}$$

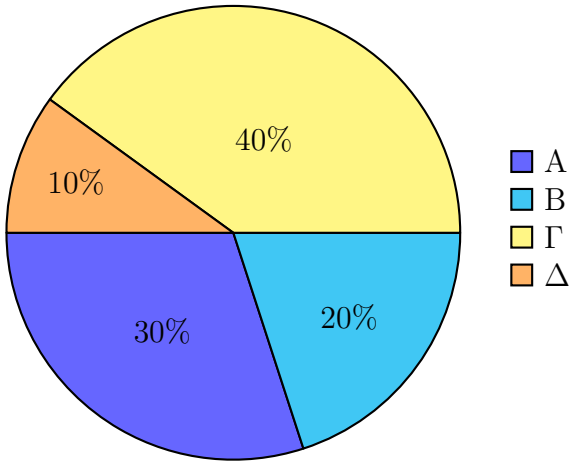
Η επιλογή βάση τάξης, σε αντίθεση με την επιλογή ρουλέτας, δουλεύει ακόμα και στην περίπτωση που η αντικειμενική συνάρτηση λάβει αρνητικές τιμές. Επιπλέον, στο ακόλουθο Διάγραμμα 1.8 παρατηρούμε ότι οι υποψήφιες λύσεις Δ,Β έχουν σημαντική διαφοροποίηση στην πιθανότητα επιλογής, σε αντίθεση με το Διάγραμμα 1.7 της προηγούμενης μεθόδου. Αυτό επιλύει το πρόβλημα της επιλογής ρουλέτας, καθώς είναι λιγότερο πιθανό δύο όμοιες λύσεις να επιλεχθούν για διασταύρωση, και συμβάλλει στην διατήρηση της ποικιλομορφίας.

3. **Τουρνουά:** Η επιλογή με τουρνουά μεγέθους κ, χωρίζει τον πληθυσμό σε ομάδες των κ-χρωμοσωμάτων. Στην συνέχεια, ο «ικανότερος» υποψήφιος σε κάθε τουρνουά έχει μία πιθανότητα επιλογής (selection probability) p_s, ο δεύτερος καλύτερος $p_s\,(1 - p_s)$ κ.ο.κ. Δηλαδή η πιθανότητα επιλογής σε ένα τουρνουά δίνεται από την σχέση:

$$p_i = p_s\,(1 - p_s)^{rank_i - 1} \quad \forall i = 1, \ldots, \kappa - 1 \tag{1.4.3}$$

όπου $rank_i$ η τάξη του υποψηφίου i στο αντίστοιχο τουρνουά. Προκειμένου να αθροίζουν οι πιθανότητες του δειγματικού χώρου στην μονάδα, η πιθανότητα επιλογής του χειρότερου υποψηφίου, εκφράζεται ως η πιθανότητα να μην επιλεγεί κανείς από τους κ-1 προηγούμενους:

$$p_{worst} = (1 - p_s)^{\kappa - 1}\,. \tag{1.4.4}$$



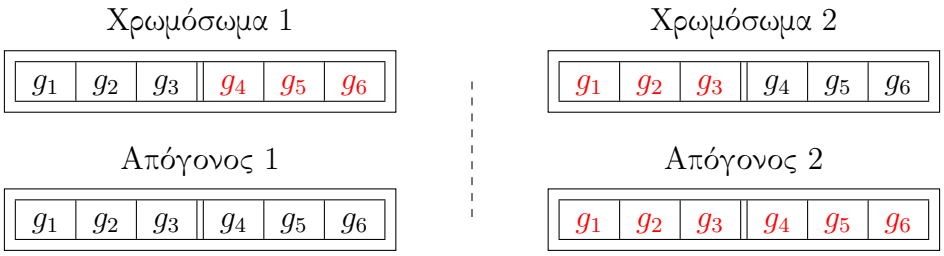
Διάγραμμα 1.8: Τομεόγραμμα πιθανότητας επιλογής βάση της τάξης των τεσσάρων υποψηφίων.

Συνήθως η πιθανότητα επιλογής είναι αρκετά μεγάλη, δηλαδή $p_s > 0.9$. Οι νικητές σε κάθε τουρνουά αποτελούν τους «γονείς» που θα παράξουν τα νέα χρωμοσώματα. Το κύριο πλεονέκτημα αυτής της μεθόδου είναι η δυνατότητα να υπάρξουν νικητές σε τουρνουά όπου όλοι οι υποψήφιοι είχαν κακή τιμή προσαρμογής, σύμφωνα με την αντικειμενική συνάρτηση. Αυτό δίνει την δυνατότητα, να διασταυρωθούν τα «λιγότερο ικανά» γονίδια με τα «περισσότερο ικανά», αυξάνοντας την ποικιλομορφία του πληθυσμού. Τα τουρνουά συνήθως δεν έχουν μέγεθος μεγαλύτερο 2 ή 3, διότι τότε μεγαλώνει το ενδεχόμενο να βρεθούν στο ίδιο τουρνουά «ικανά» χρωμοσώματα τα οποία θα επικρατήσουν.

1.4.2 Διασταύρωση χρωμοσωμάτων

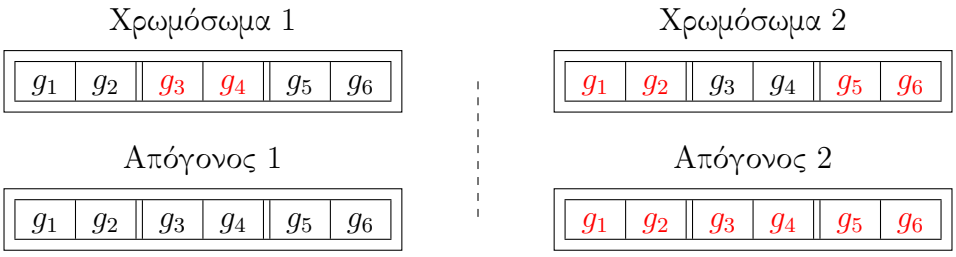
Ο γενετικός τελεστής της διασταύρωσης, αποτελεί την κύρια μέθοδο εξερεύνησης του χώρου των λύσεων. Εφαρμόζεται στους υποψηφίους του πληθυσμού που προκύπτουν έπειτα από το στάδιο της επιλογής, και έχει ως στόχο την δημιουργία νέων υποψηφίων «ικανών» λύσεων, διατηρώντας ταυτόχρονα την ποικιλομορφία του πληθυσμού. Αυτό επιτυγχάνεται συνδυάζοντας τα γονίδια από τους δύο γονείς, και η υλοποίηση του εν λόγω τελεστή συνήθως είναι κάποια από τις ακόλουθες:

1. **Διασταύρωση ενός σημείου:** Επιλέγεται τυχαία ένα σημείο στα γονίδια του χρωμοσώματος. Στην συνέχεια, τα χρωμοσώματα ανταλλάσουν γονίδια πριν και μετά από το τυχαία επιλεγμένο σημείο, και προκύπτουν δύο νέοι πιθανοί συνδυασμοί. Στο ακόλουθο Διάγραμμα 1.9, απεικονίζεται η παραπάνω διαδικασία για χρωμόσωμα μεγέθους 6, και τυχαία επιλεγμένο σημείο διασταύρωσης στην μέση.



Διάγραμμα 1.9: Διασταύρωση ενός σημείου ανάμεσα σε δύο χρωμοσώματα.

2. **Διασταύρωση πολλών σημείων:** Η λογική είναι παρόμοια με την διασταύρωση ενός σημείου όμως εδώ επιλέγονται παραπάνω από ένα σημεία, σχηματίζοντας «τμήματα» από γονίδια τα οποία εναλλάσσονται μεταξύ των δύο γονέων και σχηματίζουν τους απογόνους (βλ. Διάγραμμα 1.10). Εάν ο αριθμός των σημείων γίνει αρκετά μεγάλος τότε το καθένα από τα σχηματιζόμενα «τμήματα» στην πραγματικότητα θα τείνει να περιέχει μόνο ένα γονίδιο. Επομένως ο αλγόριθμος διασταύρωσης εκφυλίζεται σε τυχαία επιλογή γονιδίων ανάμεσα στους δύο γονείς, το οποίο δεν είναι επιθυμητό διότι τότε η πληροφορία που μεταφέρει το κάθε χρωμόσωμα κατακερματίζεται και οι απόγονοι δεν κληρονομούν κάποιο από τα χαρακτηριστικά εκείνα που ευνόησαν του γονείς τους κατά την διάρκεια της επιλογής.



Διάγραμμα 1.10: Διασταύρωση δύο σημείων ανάμεσα σε δύο χρωμοσώματα.

Αξίζει να σημειωθεί ότι στις παραπάνω μεθόδους τα γονίδια που κληρονομούν οι απόγονοι, διεκδικούν τις ίδιες θέσεις στο χρωμόσωμά τους, με αυτές των γονέων τους. Επιπλέον, η διασταύρωση πραγματοποιείται με κάποια πιθανότητα (crossover probability, P_c) την οποία ορίζει ο χρήστης και συνήθως είναι αρκετά μεγάλη.

1.4.3 Μετάλλαξη χρωμοσωμάτων

Στην φύση, κατά την διάρκεια της αντιγραφής των γονιδίων μπορεί να συμβούν λάθη τα οποία ονομάζονται μεταλλάξεις. Οι εν λόγω μεταλλάξεις δεν είναι απαραίτητο να έχουν αρνητικό αντίκτυπο στην ικανότητα επιβίωσης ενός οργανισμού, αντίθετα μπορεί να αποκτήσει νέα χαρακτηριστικά τα οποία αυξάνουν την ποικιλομορφία του πληθυσμού και ενδεχομένως βελτιώνουν την ικανότητα επιβίωσης.

Η παραπάνω διαδικασία σε έναν γενετικό αλγόριθμο επιτυγχάνεται με τον γενετικό τελεστή της μετάλλαξης, και είναι ιδιαίτερα χρήσιμος όταν ο αλγόριθμος έχει εγκλωβιστεί σε κάποιο τοπικό ακρότατο. Σε αυτή την περίπτωση, όλα τα χρωμοσώματα έχουν παρόμοια γονίδια και ο τελεστής της διασταύρωσης παράγει επίσης παρόμοιους απογόνους, αποτυγχάνοντας να βελτιώσει την λύση του προβλήματος βελτιστοποίησης. Με την βοήθεια του τελεστή μετάλλαξης, προκύπτουν νέα χρωμοσώματα με γονίδια τα οποία ήταν αδύνατο να προκύψουν από την διαδικασία της διασταύρωσης, αυξάνοντας την ποικιλομορφία του πληθυσμού, ενώ παράλληλα υπάρχει το ενδεχόμενο οι νέοι υποψήφιοι να έχουν απεγκλωβιστεί από το τοπικό ακρότατο, βελτιώνοντας την ποιότητα της λύσης στο πρόβλημα βελτιστοποίησης. Ο ρυθμός μετάλλαξης ελέγχεται μέσω της πιθανότητας μετάλλαξης (Mutation Probability, P_m) και λαμβάνει μικρές τιμές κοντά στο 0.1, διότι σε αντίθετη περίπτωση ο γενετικός αλγόριθμος θα λειτουργούσε πραγματοποιώντας εντελώς τυχαίες αναζητήσεις, μειώνοντας την αποδοτικότητά του. Συνήθως ο εν λόγω τελεστής τροποποιεί την τιμή του γονιδίου που μεταλλάσσεται, είτε σε μία εντελώς τυχαία τιμή, είτε σε ένα εύρος εντός της αρχικής.

1.4.4 Δημιουργία νέου πληθυσμού

Το τελευταίο βήμα στον γενετικό αλγόριθμο είναι η δημιουργία του πληθυσμού της επόμενης γενιάς, όπου θα δράσουν οι γενετικοί τελεστές στην επόμενη επανάληψη. Ανάλογα με τον τρόπο που συνδυάζονται τα ήδη υπάρχοντα χρωμοσώματα στον πληθυσμό με τα νέα, που προκύπτουν έπειτα από την επιλογή, διασταύρωση και μετάλλαξη, οι γενετικοί αλγόριθμοι χωρίζονται στις ακόλουθες κατηγορίες (Goswami et al., 2023):

1. **Σταθερής κατάστασης (Steady State):** Σε αυτή την κατηγορία γενετικών αλγορίθμων δεν υφίσταται η έννοια της «γενιάς», καθώς ανάμεσα στους δύο γονείς και τα δύο παιδιά, επιλέγονται οι 2 «ικανότεροι» και επιστρέφουν πίσω στον αρχικό πληθυσμό. Με αυτόν τον τρόπο το μέγεθος του πληθυσμού διατηρείται σταθερό, και αλλάζει μόνο ένα κομμάτι του πληθυσμού σε κάθε επανάληψη. Το κύριο πλεονέκτημα του γενετικού αλγορίθμου σταθερής κατάστασης αποτελεί η ικανότητα να διατηρήσει την ποικιλομορφία του πληθυσμού και να συγκλίνει ταχύτερα στην βέλτιστη λύση.

2. **Γενεαλογικοί (Generational):** Οι γενεαλογικοί γενετικοί αλγόριθμοι αντικαθιστούν σε κάθε επανάληψη τον υπάρχοντα πληθυσμό, με τους απογόνους που προκύπτουν έπειτα από την δράση των γενετικών τελεστών. Αυτό συμβάλλει στην ταχύτητα σύγκλισης, όταν ο χώρος των πιθανών λύσεων είναι μικρός και σχετικά απλός, χωρίς πολλά τοπικά ελάχιστα.

1.5 Επίλογος

Στα προβλήματα παλινδρόμησης όπου το μοντέλο είναι γραμμικό, ή αντίστοιχα η συνάρτηση υπολοίπων είναι γραμμική, είναι εύκολο να εκτιμηθούν οι παράμετροι του μοντέλου με

την μέθοδο των ελάχιστων τετραγώνων, αφού οι επαναληπτικοί μέθοδοι συγκλίνουν λόγω κυρτότητας της αντικειμενικής συνάρτησης. Σε μη γραμμικά μοντέλα, όπως αυτά στα οποία ενδέχεται να καταλήξει ένας αλγόριθμος συμβολικής παλινδρόμησης, οι τοπικές προσεγγίσεις που αναφέρθηκαν, για να συγκλίνουν θα πρέπει να δοθεί καλή αρχική συνθήκη και η αντικειμενική συνάρτηση να είναι τοπικά κυρτή. Για τα μη γραμμικά ελάχιστα τετράγωνα, η μέθοδος των Gauss-Newton εκ πρώτης όψεως φαίνεται περισσότερο ελκυστική σε σχέση με την μέθοδο του Newton, καθώς έχει μικρότερο υπολογιστικό κόστος, δεν απαιτεί την ύπαρξη της δεύτερης παραγώγου, και γραμμικοποιεί τοπικά την αντικειμενική συνάρτηση, ανάγοντας το πρόβλημα σε αυτό των γραμμικών τετραγώνων. Τέλος, για το πρόβλημα της συμβολικής παλινδρόμησης, όπου ο χώρος αναζήτησης είναι αρκετά περίπλοκος, φαίνεται να ταιριάζουν περισσότερο οι γενετικοί αλγόριθμοι σταθερής κατάστασης, ενώ η επιλογή τουρνουά, με βάση όσα ειπώθηκαν, αναμένεται να βοηθήσει περισσότερο στην διατήρηση της ποικιλομορφίας, σε σχέση με οποιαδήποτε άλλη μέθοδο επιλογής.

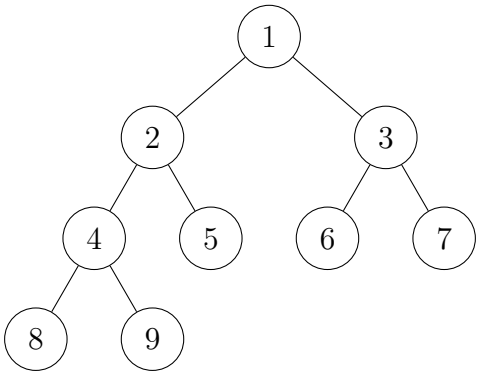
2 Αναλυτική Ανασκόπηση

2.1 Πρόλογος

Το παρόν κεφάλαιο έχει ως στόχο να κάνει μια σύντομη εισαγωγή στα δυαδικά δέντρα, τα οποία μπορούν να αξιοποιηθούν ως δομή αποθήκευσης συμβολικών εκφράσεων. Στην συνέχεια παρουσιάζονται αλγόριθμοι διαπέρασης του συμβολικού δέντρου, για τον υπολογισμό της τιμής του, καθώς και την εκτύπωση την συμβολικής έκφρασης που κωδικοποιούν. Τέλος γίνεται αναφορά σε δύο πρόσφατες υλοποιήσεις αλγορίθμων συμβολικής παλινδρόμησης. Επεξηγείται ο τρόπος λειτουργίας τους, με τα πλεονεκτήματα και τα μειονεκτήματα της κάθε επιλογής.

2.2 Δυαδικά Δέντρα

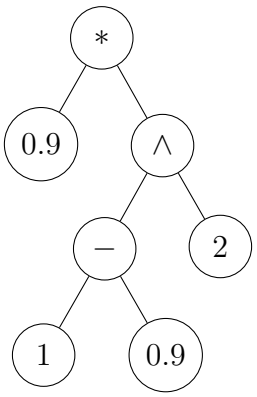
Τα δέντρα αποτελούν μία δομή δεδομένων, η οποία αποθηκεύει πληροφορίες σε ένα σύνολο από κόμβους. Οι εν λόγω κόμβοι σχετίζονται μεταξύ τους με την σχέση γονέα παιδιού. Εξαίρεση αποτελεί η ρίζα του δέντρου, η οποία δεν έχει κάποιο κόμβο - πατέρα. Τα δέντρα γενικότερα αποτελούν άκυκλα γραφήματα, δηλαδή κάθε κόμβος c, εκτός της ρίζας, έχει μοναδικό πατέρα p και ενώνονται μεταξύ τους μέσω μίας ακμής (p, c). Το σύνολο των κόμβων στο οποίο οι διαδοχικοί κόμβοι ενώνονται με ακμή, καλείται μονοπάτι. Τα δυαδικά δέντρα διαφοροποιούνται σε σχέση με τα απλά δέντρα, ως προς το μέγιστο επιτρεπτό πλήθος παιδιών ανά κόμβο. Πιο συγκριμένα, στα εν λόγω δέντρα κάθε κόμβος έχει το πολύ δύο παιδιά.



Διάγραμμα 2.1: Δυαδικό δέντρο.

Όπως φαίνεται και στο Διάγραμμα 2.1, υπάρχουν κόμβοι που δεν έχουν κανένα παιδί, γι΄ αυτό καλούνται εξωτερικοί ή φύλλα, ενώ οι υπόλοιποι είναι εσωτερικοί. Το μέγεθος του δέντρου καθορίζεται από το πλήθος των κόμβων του. Στην περίπτωση που κάθε εσωτερικός κόμβος έχει ακριβώς δύο παιδιά, τότε το δέντρο χαρακτηρίζεται ως κανονικό ή πλήρως δυαδικό.

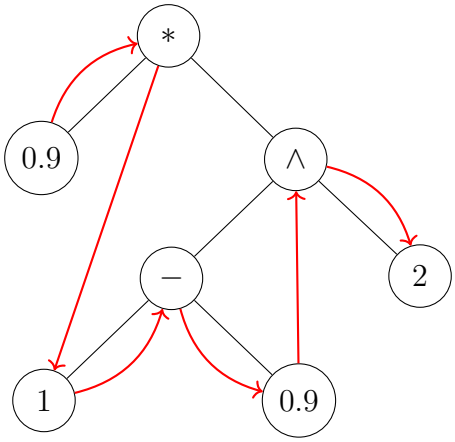
Μία εφαρμογή των δυαδικών δέντρων ως δομή δεδομένων, αποτελεί η αποθήκευση συμβολικών εκφράσεων στον υπολογιστή. Προκειμένου η εν λόγω έκφραση να είναι έγκυρη, θα πρέπει οι τελεστές να βρίσκονται αποθηκευμένοι στους εσωτερικούς κόμβους, ενώ οι μεταβλητές και οι αριθμοί στα φύλλα του δέντρου. Ενδέχεται οι τελεστές να μην είναι όλοι δυαδικοί, δηλαδή ο αντίστοιχος κόμβος να έχει δύο παιδιά, αλλά να είναι μοναδιαίοι, όπως λ.χ. το απόλυτο, ο λογάριθμος, το εκθετικό. Επομένως το αντίστοιχο δυαδικό δέντρο εν γένει δεν είναι κανονικό.



Διάγραμμα 2.2: Δέντρο της έκφρασης $0.9 * (1 - 0.9)^2$.

Για τον υπολογισμό της τιμής ενός δέντρου συμβολικής έκφρασης, όπως αυτού του Διαγράμματος 2.2, είναι αναγκαίο να αναπτυχθεί ένα αλγόριθμος που θα διαπεράσει όλους τους κόμβους του δέντρου (tree traversal). Κάποιοι από τους τρόπους με τους οποίους μπορεί να γίνει η διαπέραση του δέντρου με αφετηρία την ρίζα, είναι οι εξής (Goodrich et al., 2016):

- Ενδοδιατεταγμένη (inorder): Οι κόμβοι επισκέπτονται με την σειρά από αριστερά προς τα δεξιά. Δηλαδή ο αλγόριθμος διαπερνά αναδρομικά, πρώτα όλους τους κόμβους του αριστερού υποδέντρου, στην συνέχεια τον κόμβο αφετηρία και τέλος τους κόμβους του δεξιού υποδέντρου (βλ. Διάγραμμα 2.3 και Αλγόριθμο 1).

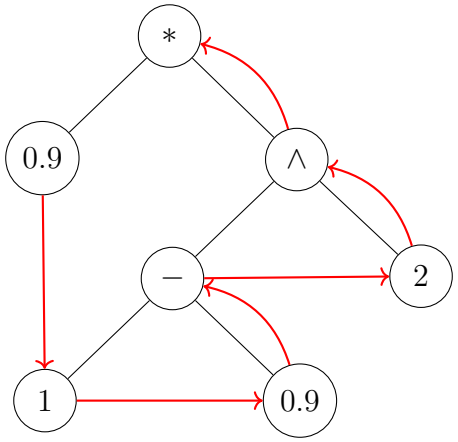


Διάγραμμα 2.3: Ενδοδιατεταγμένη διαπέραση δέντρου συμβολικής έκφρασης.

Αλγόριθμος 1 Εκτύπωση έκφρασης με ενδοδιατεταγμένη διαπέραση δυαδικού δέντρου.

```
function inorder(p)
    if p has left child l
        print "("
        inorder(l)
    print p
    if p has right child r
        inorder(r)
        print ")"
```

- Μεταδιατεταγμένη (postorder): Κατά την Μεταδιατεταγμένη διάσχιση του δέντρου, πρώτα επισκέπτονται οι κόμβοι του αριστερού υποδέντρου, εν συνεχεία οι κόμβοι του δεξιού υποδέντρου, και στο τέλος ο αρχικός κόμβος (βλ. Διάγραμμα 2.4 και Αλγόριθμο 2).

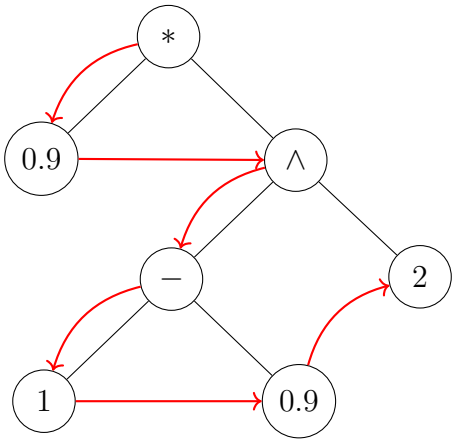


Διάγραμμα 2.4: Μεταδιατεταγμένη διαπέραση δέντρου συμβολικής έκφρασης.

Αλγόριθμος 2 Εκτύπωση έκφρασης με μεταδιατεταγμένη διαπέραση δυαδικού δέντρου.

```
function postorder(p)
    if p has left child l
        postorder(l)
    if p has right child r
        postorder(r)
    print p
```

- Προδιατεταγμένη (preorder): Ο αλγόριθμος επισκέπτεται τον κόμβο αφετηρία και στην συνέχεια αναδρομικά, πρώτα τους κόμβους του αριστερού υποδέντρου και στην συνέχεια του δεξιού (βλ. Διάγραμμα 2.5 και Αλγόριθμο 3).



Διάγραμμα 2.5: Προδιατεταγμένη διαπέραση δέντρου συμβολικής έκφρασης.

Αλγόριθμος 3 Εκτύπωση έκφρασης με προδιατεταγμένη διαπέραση δυαδικού δέντρου.

```
function preorder(p)
    print p
    if p has left child l
        preorder(l)
    if p has right child r
        preorder(r)
```

Από τις παραπάνω εναλλακτικές διάσχισης ενός δέντρου, συγκεκριμένα για τις συμβολικές εκφράσεις αξιοποιούνται ευρέως οι πρώτες δύο επιλογές. Πιο συγκεκριμένα, η ενδοδιατεταγμένη διαπέραση του συμβολικού δέντρου στο Διάγραμμα 2.3 θα τύπωνε την έκφραση (0.9 * ((1 - 0.9) ^ 2). Στην περίπτωση της μεταδιατεταγμένης διαπέρασης, η σειρά με την οποία επισκέπτονται οι κόμβοι θα έδινε για το ίδιο δέντρο την έκφραση 0.9 1 0.9 - 2 ^ *. Η έκφραση είναι γνωστή και ως postfix expression ή reverse polish notation. Η ενδοδιατεταγμένη διαπέραση βοηθάει στην απεικόνιση της συμβολικής έκφρασης σε μορφή φιλική προς τον χρήστη, ενώ η μεταδιατεταγμένη βοηθάει στον υπολογισμό της αριθμητικής τιμής του συμβολικού δέντρου, με χρήση του Αλγορίθμου 4.

Αξίζει να σημειωθεί ότι για να υπολογισθεί η τελική τιμή του συμβολικού δέντρου, αξιοποιώντας τους παραπάνω τρόπους, ο κάθε κόμβος είναι αναγκαίο να περιέχει περισσότερες πληροφορίες, εκτός από την τιμή του. Συγκεκριμένα, για ένα συμβολικό δέντρο θα ενδιαφερόμασταν εάν ο κόμβος είναι εσωτερικός ή εξωτερικός, δηλαδή αν περιέχει τελεστή ή μεταβλητή αντίστοιχα, και το είδος του τελεστή (μοναδιαίος ή δυαδικός). Στην περίπτωση του μοναδιαίου έχει μόνο ένα παιδί το οποίο κατά σύμβαση μπορεί να τοποθετηθεί δεξιά.

Το παραπάνω προγραμματιστικά γίνεται εφικτό μέσω της ενθυλάκωσης δεδομένων (data encapsulation). Ο κόμβος αποτελεί μία κλάση η οποία περιέχει αρκετά πεδία για να φιλοξενήσει όλες τις απαραίτητες πληροφορίες και υποστηρίζει ορισμένες μεθόδους οι οποίες

Αλγόριθμος 4 Υπολογισμός της αριθμητικής τιμής μέσω PostOrder διαπέρασης.

```
function evaluate(p)
    if p is leaf node
        return value of p
    else
        if element e of p is binary operator
            left_value  = evaluate(left  Child)
            right_value = evaluate(right Child)
            return e(left_value,right_value)
        else
            right_value = evaluate(right Child)
            return e(right_value)
```

διευκολύνουν την διαπέραση του δέντρου και τον αριθμητικό υπολογισμό της έκφρασης που αντιπροσωπεύει.

2.3 Προηγούμενες υλοποιήσεις αλγορίθμων συμβολικής παλινδρόμησης

Το ενδιαφέρον ανάπτυξης ενός αλγορίθμου που να δημιουργεί μία κλειστή μαθηματική έκφραση από τα δεδομένα, υπάρχει εδώ και πολλά χρόνια. Οι πρώτες υλοποιήσεις που έγιναν, στην πραγματικότητα επιχειρούσαν μέσω εξαντλητικής αναζήτησης στον χώρο των μεταβλητών και των πιθανών τελεστών, να κατασκευάσουν το τελικό μοντέλο. Αν συνυπολογίσουμε όμως στον εν λόγω χώρο και όλους του πιθανούς συνδυασμούς τιμών για τις παραμέτρους των εν λόγω μοντέλων, ο χώρος των πιθανών λύσεων γίνεται άπειρος και είναι σχεδόν αδύνατο να βρεθεί η κατάλληλη συναρτησιακή εξάρτηση μεταξύ των επεξηγηματικών μεταβλητών και της μεταβλητής απόκρισης. Για τον σκοπό αυτό, ήταν αναγκαίο να αναπτυχθούν διαφορετικές στρατηγικές αναζήτησης του χώρου των πιθανών μοντέλων, οι οποίες με την εξέλιξη της τεχνολογίας και συγκεκριμένα την αύξηση την υπολογιστικής ισχύος που διαθέτουν οι σημερινοί υπολογιστές, έχουν γίνει αρκετά δημοφιλείς τα τελευταία χρόνια. Αυτό έχει οδηγήσει τις σύγχρονες υλοποιήσεις αλγορίθμων συμβολικής παλινδρόμησης να εντάσσονται σε κάποιες από τις κατηγορίες που θα αναφερθούν στις ακόλουθες υποενότητες.

2.3.1 Γενετικοί αλγόριθμοι

Οι γενετικοί αλγόριθμοι ως τεχνική βελτιστοποίησης, έχουν εφαρμοστεί για την επίλυση ευρείας γκάμας προβλημάτων, με το κυριότερο πλεονέκτημα ότι δεν απαιτούνται ιδιαίτερες προϋποθέσεις προκειμένου να επιλυθεί το πρόβλημα. Από τα πρώτα εργαλεία συμβολικής παλινδρόμησης που υλοποιήθηκαν με την χρήση γενετικών αλγορίθμων, είναι το λογισμικό EUREQA, το οποίο είχε υποσχόμενα αποτελέσματα και προσέλκυσε το ενδιαφέρον των

ερευνητών στην συγκεκριμένη μέθοδο. Το EUREQA ήταν λογισμικό κλειστού κώδικα επί πληρωμή και πλέον έχει σταματήσει να διατίθεται, όμως έχουν υπάρξει εναλλακτικές λύσεις όπως η σχετικά πρόσφατη υλοποίηση της βιβλιοθήκης SymbolicRegression, στην γλώσσα προγραμματισμού Julia (Cranmer, 2023).

Η παραπάνω υλοποίηση στοχεύει να επιλύσει προβλήματα φυσικής. Στα προβλήματα φυσικής οι απαιτήσεις είναι ιδιαίτερες, καθώς οι μετρήσεις ενδέχεται να περιέχουν θόρυβο και το ζητούμενο μοντέλο, εκτός από καλή προσαρμογή θα πρέπει να είναι το απλούστερο δυνατό ώστε να επεξηγεί με απλό τρόπο την σχεσιακή εξάρτηση μεταξύ των χαρακτηριστικών. Για τον σκοπό αυτό η αντικειμενική συνάρτηση του γενετικού αλγορίθμου αποδίδει ποινή στις περίπλοκες εκφράσεις. Είναι εμφανές από τα παραπάνω, ότι στην πραγματικότητα η εύρεση κατάλληλης συμβολικής έκφρασης, αποτελεί ένα πρόβλημα βελτιστοποίησης Pareto (βλ. Διάγραμμα 2.6), όπου η αντικειμενική συνάρτηση για μία έκφραση E, είναι της μορφής:

$$obj\left(E\right) = w \times l\left(E\right) + \left(1 - w\right) \times C\left(E\right) \tag{2.3.1}$$

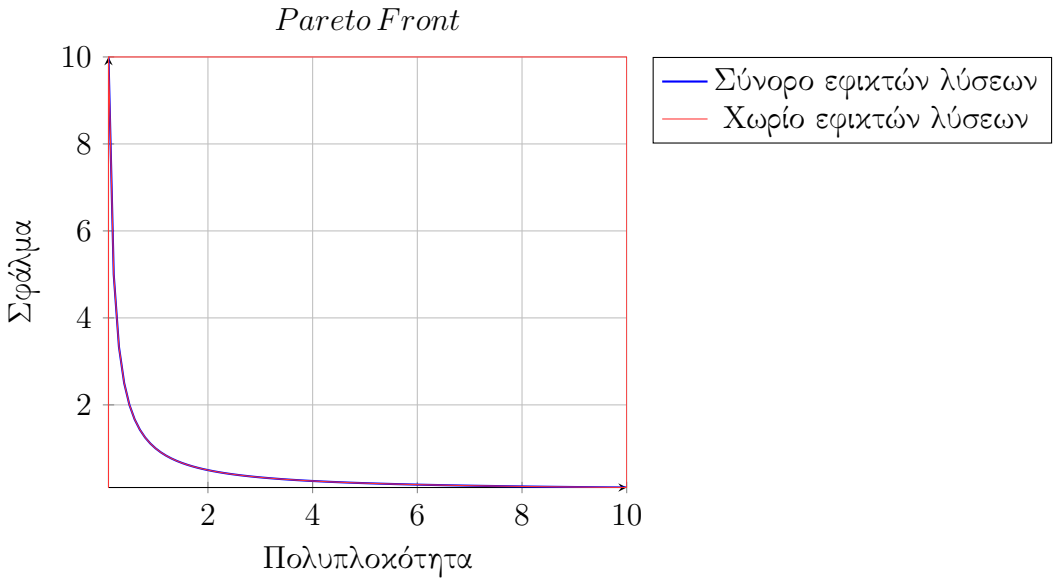
όπου:

- $l\left(E\right)$ η τιμή του «σφάλματος» για την έκφραση E, η οποία υπολογίζεται από την συνάρτηση $l\left(\cdot\right)$ που καθορίζει ο χρήστης, λ.χ. μέσο τετραγωνικό σφάλμα.
- $C\left(E\right)$ η πολυπλοκότητα της συμβολικής έκφρασης. Ορίζεται από τον χρήστη για κάθε τελεστή η σταθερά της ποινής και με αυτόν τον τρόπο αθροίζοντας την σταθερά ποινής για κάθε τελεστή έχουμε την ολική πολυπλοκότητα της συμβολικής έκφρασης. Για παράδειγμα, στο συμβολικό δέντρο του Διαγράμματος 2.2, ορίζοντας τις ποινές 1,2,3 για τους τελεστές «$-$», «$*$» και «$\wedge$» αντίστοιχα, η πολυπλοκότητα της συμβολικής είναι έκφρασης είναι 5.
- w το βάρος ανάμεσα στην καλύτερη προσαρμογή και την πολυπλοκότητα της έκφρασης, καθορίζεται από τον χρήστη.

Εκτός από τον καθορισμό της ποινής σε κάθε τελεστή, η υλοποίηση του πακέτου SymbolicRegression.jl, προσφέρει την δυνατότητα στον χρήστη να ορίσει το μέγιστο μέγεθος του υποδέντρου που μπορεί να φιλοξενήσει ο κόμβος του κάθε τελεστή. Αυτό συμβαίνει διότι σε διαφορετικά πεδία της επιστήμης αναμένεται κανείς να συναντήσει διαφορετικές σχέσεις ανάμεσα στα μεγέθη. Για παράδειγμα σε κάποιο πεδίο μπορεί η έκφραση $log\left(log\left(x\right)\right)$ να είναι αναμενόμενη, όμως σε κάποιο άλλο η ύπαρξη εμφολευμένων λογαρίθμων να είναι προβληματική.

Η ιδιαιτερότητα του γενετικού αλγορίθμου που υλοποιεί η παραπάνω βιβλιοθήκη, αρχικά οφείλεται στην αντικειμενική συνάρτηση. Αντί για την μορφή 2.3.1, χρησιμοποιείται η ακόλουθη:

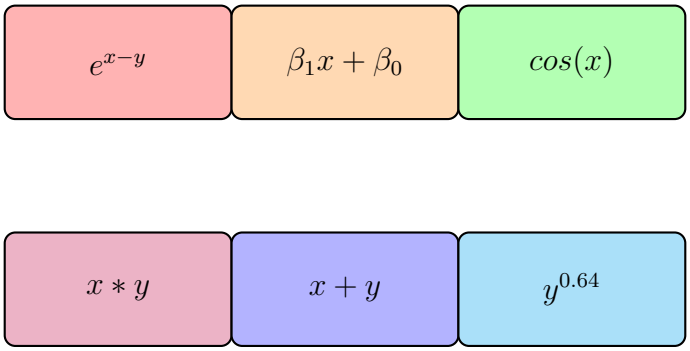
$$obj = l\left(E\right) \cdot e^{f[C(E)]} \tag{2.3.2}$$



Διάγραμμα 2.6: Καμπύλη Pareto ανάμεσα στα δύο μεγέθη. Οι εφικτές λύσεις βρίσκονται πάνω από το σύνορο.

όπου $f(\cdot)$ μία συνάρτηση που μετράει την συχνότητα εμφάνισης και την μέση ηλικία των εκφράσεων με πολυπλοκότητα $C(E)$. Αυτό συμβάλλει στο να διατηρηθούν και να εξελιχθούν στον πληθυσμό υποψήφιοι με απλή συμβολική έκφραση και μέτρια προσαρμοστικότητα, βοηθώντας τον αλγόριθμο να αποφύγει την πρόωρη σύγκληση, και δίνει παραπάνω χρόνο στην εξερεύνηση του χώρου.

Μία δεύτερη ιδιαιτερότητα, αποτελεί ο τρόπος με τον οποίο χωρίζονται οι υποψήφιες λύσεις στον πληθυσμό. Πιο συγκεκριμένα, αντί να υπάρχει ένας πληθυσμός που να διατηρεί όλες τις συμβολικές εκφράσεις, εκείνες χωρίζονται σε υπό πληθυσμούς. Κάθε υποπληθυσμός αρχικοποιείται με συμβολικές εκφράσεις που περιλαμβάνουν έναν μόνο τελεστή, όπως φαίνεται στο Διάγραμμα 2.7, και στην συνέχεια μόνο οι καλύτερες εκφράσεις του εκάστοτε πληθυσμού «μεταναστεύουν» σε άλλους υποπληθυσμούς δημιουργώντας απογόνους με συνδυασμούς τελεστών.



Διάγραμμα 2.7: Υποπληθυσμοί που αναπαριστούν οικογένειες συναρτήσεων.

Στην συνέχεια, η επιλογή των υποψηφίων για διασταύρωση γίνεται μέσω επιλογής τουρνουά, και πραγματοποιείται διασταύρωση ενός σημείου. Τα χρωμοσώματα, δηλαδή οι συμβολικές εκφράσεις, κωδικοποιούνται στην μορφή ενός δυαδικού δέντρου και ο αντίστοιχος τελεστής της μετάλλαξης, δρα με κάποιον από τους ακόλουθους τρόπους:

1. Μετάλλαξη σταθεράς του μοντέλου.
2. Τυχαία μετάλλαξη ενός τελεστή, αντικαθιστώντας τον με κάποιον άλλο παρόμοιας περιπλοκότητας.
3. Προσθήκη νέας ρίζας ή φύλλου στο δέντρο.
4. Αντικατάσταση υποδέντρου με τυχαία μεταβλητή ή σταθερά.
5. Απλοποίηση της συμβολικής έκφρασης.
6. Αντικατάσταση του δέντρου με ένα νέο.

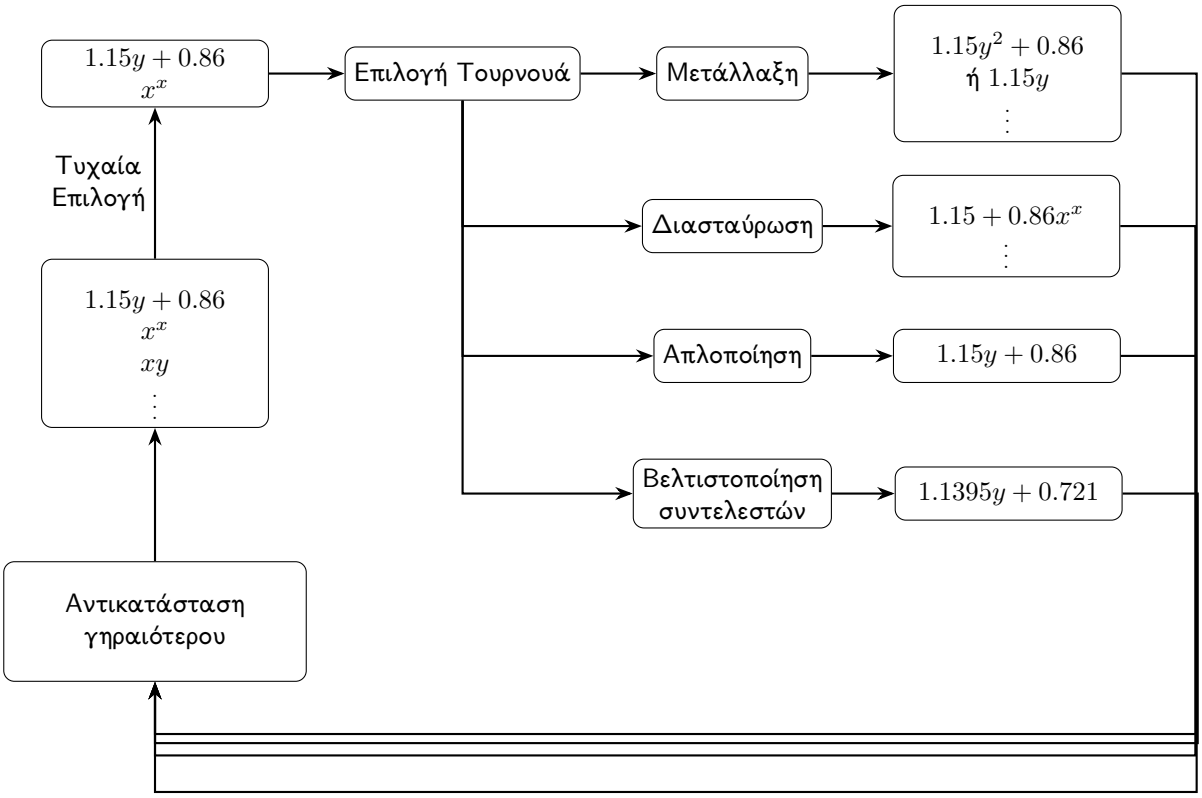
Κάθε ένα από τα παραπάνω σενάρια μετάλλαξης μπορεί να συμβεί με διαφορετική πιθανότητα, ανάλογα το πρόβλημα που καλούμαστε να επιλύσουμε. Η 5η περίπτωση της απλοποίησης της συμβολικής έκφρασης αποτελεί ένα ιδιαίτερα δύσκολο κομμάτι, το οποίο βοηθά τα δέντρα να διατηρούν μικρό μέγεθος και ταυτόχρονα να μην υπάρχει «φλυαρία» στο τελικό μοντέλο.

Η μετάλλαξη γίνεται αποδεκτή όταν η τιμή της αντικειμενικής συνάρτησης είναι καλύτερη στο δέντρο που προκύπτει, έναντι του αρχικού. Εν αντιθέσει, όταν είναι χειρότερη υπάρχει μία πιθανότητα να γίνει και πάλι αποδεκτή. Πιο συγκεκριμένα, η εν λόγω πιθανότητα καθορίζεται από το πόσο μεγάλη είναι η μεταβολή στην τιμή της αντικειμενικής συνάρτησης και αν ο αλγόριθμος βρίσκεται σε κύκλο μεγάλης ή μικρής θερμοκρασίας (περισσότερα στο επόμενο κεφάλαιο).

Τέλος, οι νέοι υποψήφιοι που προκύπτουν εισάγονται στον πληθυσμό όπου ανήκουν οι γονείς τους, αντικαθιστώντας τους παλαιότερους ηλικιακά υποψηφίους. Επομένως διατηρείται μέρος του πληθυσμού από γενιά σε γενιά, συνεπώς το είδος του γενετικού αλγορίθμου είναι σταθερής κατάστασης. Η επαναληπτική διαδικασία του πακέτου SymbolicRegression.jl, συνοψίζεται στο Διάγραμμα 2.8.

2.3.2 Μαρκοβιανή αλυσίδα Μόντε Κάρλο

Το πρόβλημα του γενετικού αλγορίθμου, είναι ότι δεν εγγυάται πως θα επισκεφτεί με μεγαλύτερη συχνότητα τα μοντέλα που περιγράφουν καλύτερα τα δεδομένα, λόγω του ενδεχόμενου πρόωρης σύγκλισης. Αυτό επιχειρεί να επιλύσει η μέθοδος συμβολικής παλινδρόμησης με χρήση Μαρκοβιανών Αλυσίδων Μόντε Κάρλο (Markov Chain Monte Carlo, MCMC). Η εν λόγω μέθοδος, επιλύει το θεμελιώδες πρόβλημα ανάμεσα στην καλή προσαρμογή και περιπλοκότητα του μοντέλου, αξιοποιώντας μία εναλλακτική, Μπεϋζιανή προσέγγιση. Πιο



Διάγραμμα 2.8: Διάγραμμα ροής SymbolicRegression.jl (Cranmer, 2023).

συγκεκριμένα, υπολογίζεται για τα μοντέλα ανάμεσα στα οποία έχει να επιλέξει ο αλγόριθμος, μία εκ των υστέρων πιθανότητα (posterior ή ύστερη) η οποία περιλαμβάνει την καλή προσαρμογή και την πολυπλοκότητα του μοντέλου. Η πολυπλοκότητα στην ύστερη πιθανότητα, προκύπτει έχοντας μελετήσει εκ των προτέρων άλλα μοντέλα, όπως μοντέλα φυσικής, και αποδίδει μία εκ των προτέρων (prior ή πρότερη) πιθανότητα για το εξεταζόμενο μοντέλο, ότι αποτελεί κατάλληλο υποψήφιο.

Η φυσική σημασία της «εκ των υστέρων» πιθανότητας για κάποιο μοντέλο f_i, είναι να «εκτιμήσει» το ενδεχόμενο η μεταβλητή απόκρισης να περιγράφεται από την συναρτησιακή εξάρτηση f_i:

$$\boldsymbol{y} = f_i\left(\boldsymbol{x}, \boldsymbol{\theta_i}\right) + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2 I_n\right)$$

δεδομένου ότι έχει προηγηθεί το πείραμα και έχουν συλλεχθεί τα δεδομένα (γεγονός D). Από τον τύπο του Bayes, η παραπάνω πιθανότητα γράφεται:

$$P\left(f_i|D\right) = \frac{P\left(D|f_i\right) P\left(f_i\right)}{P\left(D\right)}. \tag{2.3.3}$$

Η πιθανότητα $P\left(D\right)$ είναι κοινή σε κάθε μοντέλο f_i, επομένως ορίζουμε να είναι μία σταθερά Z, και η πιθανότητα $P\left(f_i\right)$ είναι η πρότερη. Εφαρμόζοντας το θεώρημα της ολικής πιθανότητας για την δεσμευμένη πιθανότητα $P\left(D|f_i\right)$, λαμβάνουμε την εξής σχέση:

$$P\left(D|f_i\right) = \int_{\boldsymbol{\Theta_i}} P\left(D|f_i, \boldsymbol{\theta_i}\right) P\left(\boldsymbol{\theta_i}|f_i\right) d\boldsymbol{\theta_i} \tag{2.3.4}$$

όπου $\boldsymbol{\theta_i} \in \boldsymbol{\Theta_i}$ οι τιμές των παραμέτρων του μοντέλου f_i. Συνδυάζοντας τώρα τις σχέσεις 2.3.3 και 2.3.4:

$$P\left(f_i|D\right) = \frac{P\left(f_i\right) \int_{\boldsymbol{\Theta_i}} P\left(D|f_i, \boldsymbol{\theta_i}\right) P\left(\boldsymbol{\theta_i}|f_i\right) d\boldsymbol{\theta_i}}{Z} \tag{2.3.5}$$

$$= \frac{e^{-\mathcal{L}(f_i)}}{Z}. \tag{2.3.6}$$

Η ποσότητα $\mathcal{L}\left(f_i\right)$ καλείται ελάχιστο περιγραφικό μήκος, και μετράει το μήκος της μικρότερης δυνατής κωδικοποίησης των δεδομένων που απαιτείται, χωρίς να χαθεί πληροφορία.

$$\mathcal{L}\left(f_i\right) = -ln\left(P\left(f_i\right) \int_{\boldsymbol{\Theta_i}} P\left(D|f_i, \boldsymbol{\theta_i}\right) P\left(\boldsymbol{\theta_i}|f_i\right) d\boldsymbol{\theta_i}\right) \tag{2.3.7}$$

$$= -ln\left(\int_{\boldsymbol{\Theta_i}} P\left(D|f_i, \boldsymbol{\theta_i}\right) P\left(\boldsymbol{\theta_i}|f_i\right) d\boldsymbol{\theta_i}\right) - ln\left(P\left(f_i\right)\right). \tag{2.3.8}$$

Η έννοια του ελάχιστου περιγραφικού μήκους (minimum description length, MDL) είναι αρκετά αφηρημένη και έχει αποδειχθεί πως δεν υπάρχει αλγόριθμος που να αναγνωρίζει σε πολυωνυμικό χρόνο, αν μία κωδικοποίηση ελαχιστοποιεί το MDL. Παρόλαυτα μπορούν να γίνουν προσεγγίσεις, όπως η ακόλουθη:

$$\mathcal{L}\left(f_i\right) \approx \frac{BIC\left(f_i\right)}{2} - ln\left(P\left(f_i\right)\right). \tag{2.3.9}$$

Το BIC, αποτελεί ένα από τα κριτήρια ποινικοποιημένης πιθανοφάνειας (Φουσκάκης, 2013), και υπολογίζεται ως:

$$BIC\left(f_i\right) = -2 \cdot ln\left(L\left(f_i\right)\right) + p \cdot ln\left(n\right) \tag{2.3.10}$$

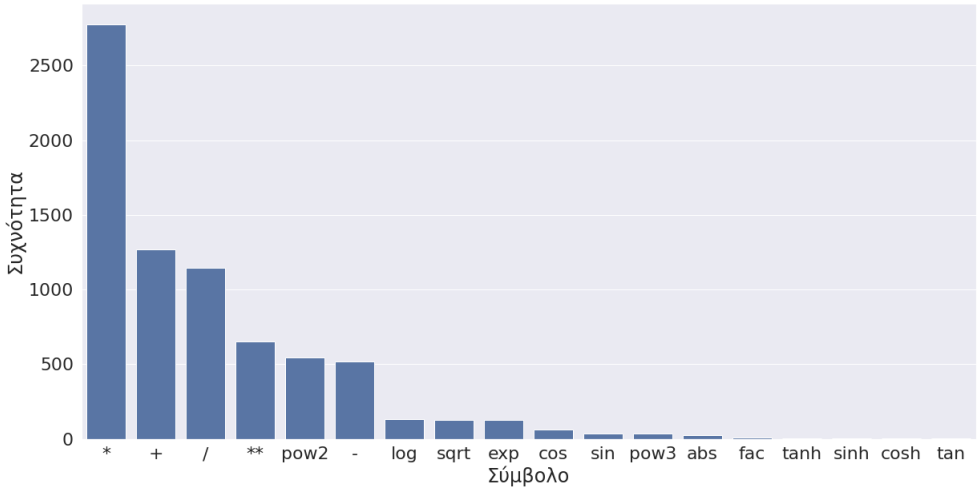
όπου $L\left(f_i\right)$ η μεγιστοποιημένη πιθανοφάνεια του μοντέλου f_i, με παραμέτρους $\boldsymbol{\theta_i^*} \in \mathbb{R}^p$, p η διάσταση του $\boldsymbol{\theta_i}$ και n το πλήθος των δεδομένων.

Από την Εξίσωση 2.3.5, είναι εμφανές ότι επιλέγονται τα μοντέλα με την μεγαλύτερη ύστερη πιθανότητα. Η εν λόγω πιθανότητα αυξάνει μικραίνοντας το ελάχιστο περιγραφικό μήκος, το οποίο με την βοήθεια της Εξίσωσης 2.3.9, συμβαίνει όταν αποδίδεται στο μοντέλο μικρή τιμή του BIC και μεγάλη πρότερη πιθανότητα. Πιο συγκεκριμένα, το BIC οδηγεί σε μοντέλα με καλή προσαρμογή, λόγω της μεγιστοποιημένης πιθανοφάνειας, και ταυτόχρονα

ποινικοποιεί το μοντέλο όταν αυτό περιλαμβάνει πολλές παραμέτρους. Επιπλέον, η εκ των προτέρων κατανομή αποδίδει χαμηλή πιθανότητα στα μοντέλα τα οποία εμφανίζονται τελεστές με «ασυνήθιστη» συχνότητα, σε σχέση με μια πρότερη γνώση άλλων μοντέλων. Επομένως η περιπλοκότητα του μοντέλου της MCMC, ποινικοποιείται σε επίπεδο παραμέτρων από το BIC, και σε επίπεδο δομής από την πρότερη πιθανότητα.

Η εκ των προτέρων πιθανότητα $P(f_i)$ θα μπορούσε να είναι ίδια για όλα τα πιθανά μοντέλα που απαρτίζουν τον χώρο των τυχαίων συμβολικών εκφράσεων. Τότε, η εκ των υστέρων πιθανότητα επιλογής ενός μοντέλου θα εξαρτάται αποκλειστικά και μόνο από την τιμή του BIC. Δίχως την συμβολή της πρότερης πιθανότητας, ενδέχεται το τελικό μοντέλο να μην είναι επεξηγήσιμο, λόγω της περίπλοκης δομής του, επομένως το εν λόγω μοντέλο είναι υπερεκπαιδευμένο, όχι με την κλασσική έννοια των παραμέτρων, αλλά ως προς την συμβολική έκφραση η οποία περιέχει όρους που δεν επεξηγούν τα μοτίβα στα δεδομένα, αλλά την έκτροπη συμπεριφορά ορισμένων παρατηρήσεων.

Με βάση τα παραπάνω, ο Roger Guimerà, (Guimerà et al., 2020), μελετώντας μία λίστα που αποτελείται από 4080 συμβολικές εκφράσεις της ιστοσελίδας Wikipedia[1], κατασκεύασε το αντίστοιχο ιστόγραμμα (βλ. Διάγραμμα 2.9) με την συχνότητα εμφάνισης κάθε τελεστή.



Διάγραμμα 2.9: Ιστόγραμμα συχνοτήτων εμφάνισης κάθε τελεστή από την λίστα της Wikipedia.

Στην συνέχεια, όρισε την εκ των προτέρων πιθανότητα, ώστε τα επιλεγμένα μοντέλα να έχουν παρόμοια συχνότητα, καθώς και το τετράγωνο της συχνότητας εμφάνισης κάθε τελεστή με την εμπειρική λίστα του Wikipedia[1]. Συγκεκριμένα, η πρότερη πιθανότητα που πληροί τις παραπάνω προϋποθέσεις υπολογίζεται ως εξής:

$$P(f_i) = \sum_{o \in \mathcal{O}} a_o n_o(f_i) + \beta_o n_o^2(f_i) \tag{2.3.11}$$

[1] https://en.wikipedia.org/wiki/List_of_scientific_equations_named_after_people

όπου $\mathcal{O} = \{+, *, \cos, \ldots\}$ το σύνολο των τελεστών και $n_o(f_i)$ η συνάρτηση που μετράει πόσες φορές εμφανίζεται στο δέντρο ο τελεστής $o \in \mathcal{O}$. Οι συντελεστές a_o και β_o αποτελούν υπερπαραμέτρους οι οποίες εκτιμούνται κατάλληλα, ώστε οι συμβολικές εκφράσεις που παράγει η μέθοδος MCMC, να περιέχουν τελεστές που ακολουθούν την κατανομή του Διαγράμματος 2.9.

Από πρόβλημα σε πρόβλημα οι υπερπαράμετροι a_o και β_o διαφέρουν, καθώς στο κριτήριο με βάση το οποίο εξελίσσεται η συμβολική έκφραση, δηλαδή την ύστερη πιθανότητα, υπεισέρχεται και το κριτήριο του BIC μαζί με την πρότερη. Επιπλέον σε διαφορετικά προβλήματα, ενδέχεται ο στόχος της πρότερης κατανομής να μην είναι να προσεγγίσει αυτή του Διαγράμματος 2.9, αλλά ένα υποσύνολο από τις 4080 εκφράσεις του Wikipedia. Γι΄ αυτό τον λόγο τα a_o και β_o της Εξίσωσης 2.3.11 καθορίζονται «εκ του αποτελέσματος»:

1. Αρχικοποιούνται οι τιμές των a_o και β_o.

2. Δημιουργείται ένας πολύ μεγάλος πληθυσμός από συμβολικές εκφράσεις MCMC.

3. Στην συνέχεια υπολογίζεται η συχνότητα εμφάνισης n_o του κάθε τελεστή $o \in \mathcal{O}$, και ελέγχεται αν η κατανομή της προσεγγίζει αυτή της πρότερης κατανομής.

   (α΄) Αν ναι τότε οι τιμές που επιλέχθηκαν είναι ικανοποιητικές.

   (β΄) Διαφορετικά με βάση τις δειγματικές τιμές n_o και n_o^2 του πληθυσμού που προέκυψε από το βήμα 2, ενημερώνονται οι υπερπαράμετροι:

$$\alpha_o^{(i+1)} = max\left(0, a_o^{(i)} + \varepsilon\lambda\frac{n_o - n_{o,target}}{n_{o,target}}\right) \tag{2.3.12}$$

$$\beta_o^{(i+1)} = max\left(0, \beta_o^{(i)} + \varepsilon\lambda\frac{n_o^2 - n_{o,target}^2}{n_{o,target}^2}\right) \tag{2.3.13}$$

   όπου λ ο ρυθμός εκμάθησης, με τιμές συνήθως κοντά στο 0.01 και $\varepsilon \sim N(0, 1)$.

4. Τερματισμός αν η μεταβολή των ενημερωμένων τιμών είναι μικρή σε σχέση με την αρχική. Διαφορετικά επιστροφή στο βήμα 2.

Η υλοποίηση της Μαρκοβιανής Αλυσίδας, έχει ως στόχο να δειγματίσει συμβολικές εκφράσεις από την κατανομή που ακολουθεί η εκ των υστέρων πιθανότητα $P(f_i|D)$ (Εξίσωση 2.3.5), επομένως σε κάθε βήμα θα πρέπει να εγγυάται ότι η μέθοδος MCMC με το πέρας του χρόνου, θα καταλήξει στην στάσιμη κατανομή $\pi_i = P(f_i|D)$. Αυτό μπορεί να γίνει εάν σε κάθε βήμα η πιθανότητα αποδοχής της μετάβασης από μια συμβολική έκφραση f_i σε κάποια f_j, ακολουθεί τον κανόνα του Metropolis-Hastings:

$$p_{accept}\left(f_i \to f_j\right) = \min\left\{1, \frac{P(f_j|D)/g(f_j|f_i)}{P(f_i|D)/g(f_i|f_j)}\right\} \tag{2.3.14}$$

$$= \min\left\{1, \frac{\pi_j}{\pi_i} \times \frac{g\left(f_i|f_j\right)}{g\left(f_j|f_i\right)}\right\}, \tag{2.3.15}$$

όπου $g\left(f_j|f_i\right)$ η κατανομή εισήγησης της συμβολικής έκφρασης f_j δεδομένου ότι στην τωρινή κατάσταση η MCMC βρίσκεται στο μοντέλο f_i. Η εν λόγω κατανομή, καθορίζει την ταχύτητα σύγκλισης της αλυσίδας στην στάσιμη κατανομή. Σε αρκετές εφαρμογές επιλέγεται να είναι συμμετρική, δηλαδή $g\left(f_i|f_j\right) = g\left(f_j|f_i\right)$. Στην Συμβολική Παλινδρόμηση, η κατανομή εισήγησης στην πιθανότητα αποδοχής του κανόνα Metropolis-Hastings 2.3.15, επιλέγεται να είναι η ομοιόμορφη κατανομή (Häggström, 2002), δηλαδή:

$$p_{accept}\left(f_i \to f_j\right) = \min\left\{1, \frac{\pi_j}{\pi_i} \times \frac{1/d_j}{1/d_i}\right\} \tag{2.3.16}$$

$$= \min\left\{1, \frac{\pi_j}{\pi_i} \times \frac{d_i}{d_j}\right\} \tag{2.3.17}$$

όπου d_i και d_j το πλήθος των γειτονικών συμβολικών εκφράσεων που είναι προσβάσιμες από τις εκφράσεις f_i και f_j αντίστοιχα, με κάποιο από τα πιθανά βήματα της αλυσίδας (βλ. παρακάτω). Η Μαρκοβιανή Αλυσίδα που πραγματοποιεί τις μεταβάσεις ακολουθώντας αυτόν τον κανόνα, ικανοποιεί την ακόλουθη σχέση:

$$\pi_i P\left(f_i \to f_j\right) = \pi_j P\left(f_j \to f_i\right) \tag{2.3.18}$$

για κάθε f_i, f_j. Άρα οι αντίστοιχες πιθανότητες μετάβασης και η αναλλοίωτη κατανομή π βρίσκονται σε ακριβή ισορροπία και είναι χρονικά αντιστρέψιμες. Το τελευταίο συνεπάγεται ότι η MCMC που ακολουθεί τις παραπάνω πιθανότητες μετάβασης έχει ως αναλλοίωτη την κατανομή $\pi = P\left(f|D\right)$ (Λουλάκης, 2019), δηλαδή την κατανομή της ζητούμενης ύστερης πιθανότητας.

Τα πιθανά βήματα που ορίζονται για την μετάβαση από μία συμβολική έκφραση f_i σε κάποια f_j, είναι τα εξής:

- Αντικατάσταση κόμβου (node replacement): Επιλέγεται τυχαία κάποιος κόμβος της συμβολικής έκφρασης, και αντικαθίσταται με άλλον, αρκεί να είναι συμβατός. Δηλαδή αν ο κόμβος που επιλέχθηκε είναι φύλλο, τότε περιέχει μεταβλητή και θα αντικατασταθεί από κάποια άλλη μεταβλητή. Αντίθετα, εάν είναι εσωτερικός κόμβος, τότε περιέχει κάποιον τελεστή, ο οποίος αν είναι δυαδικός θα πρέπει αντικατασταθεί από άλλον δυαδικό τελεστή, αλλιώς με μοναδιαία πράξη.

- Προσθήκη ή αντικατάσταση ρίζας (root addition ή root removal):

- Στην περίπτωση της προσθήκης ρίζας, επιλέγεται ομοιόμορφα στην τύχη ένας τελεστής ως νέα ρίζα, και το παρόν δέντρο προσαρτίζονται, κατά σύμβαση, στα αριστερά της νέας ρίζας, ενώ δεξιά υπάρχει ως φύλλο μία τυχαία επιλεγμένη μεταβλητή.
- Η αντίθετη κίνηση είναι ντετερμινιστική, δηλαδή αφαιρείται η ρίζα του δέντρου, και κατά σύμβαση το δεξί υποδέντρο. Επομένως η νέα συμβολική έκφραση που προκύπτει, είναι το αριστερό υποδέντρο της αρχικής.

- Αντικατάσταση απλού δέντρου (elementary tree replacement): Ως απλό δέντρο καλείται το δέντρο έκφρασης το οποίο περιέχει ως ρίζα:
  - Έναν τελεστή, και τα παιδιά του είναι φύλλα (δέντρο μεγέθους 2 ή 3 ανάλογα τον τύπο του τελεστή στην ρίζα).
  - Μία μεταβλητή (δηλαδή το δέντρο έχει μέγεθος 1).

  Στην αντικατάσταση απλού δέντρου, επιλέγεται τυχαία ένα απλό δέντρο από το δέντρο της συμβολικής έκφρασης, και αντικαθίσταται με ένα τυχαία επιλεγμένο ομοιόμορφα, από όλα τα πιθανά απλά δέντρα.

Για τον υπολογισμό της πιθανότητας μετάβασης 2.3.17, πρέπει να διακρίνουμε περιπτώσεις ανάλογα με το βήμα που πραγματοποιεί η Μαρκοβιανή Αλυσίδα:

- Αντικατάσταση κόμβου: Η επιλογή του κόμβου προς αντικατάσταση γίνεται τυχαία με ομοιόμορφη πιθανότητα. Επιπροσθέτως είναι ομοιόμορφα τυχαία η επιλογή του αντικαταστάτη, και η αντίστροφη διαδικασία ανάκτησης της αρχικής συμβολικής έκφρασης f_i από την f_j μέσω της αντικατάστασης κόμβου είναι συμμετρική. Άρα $d_i = d_j$ και η πιθανότητα αποδοχής είναι:

$$P_{accept}^{ReplaceNode}(f_i \to f_j) = min\left\{1, \frac{\pi_j}{\pi_i}\right\} \tag{2.3.19}$$

$$= min\left\{1, \frac{P(f_j|D)}{P(f_i|D)}\right\} \tag{2.3.20}$$

$$= min\left\{1, e^{-(\mathcal{L}(f_j)-\mathcal{L}(f_i))}\right\}. \tag{2.3.21}$$

- Προσθήκη ή αφαίρεση ρίζας: Η επιλογή της νέας ρίζας, γίνεται επιλέγοντας τυχαία με ίση πιθανότητα κάποιο από τα $d_i = N$ δέντρα που αποτελούνται από έναν μοναδιαίο τελεστή ως ρίζα, ή έναν δυαδικό τελεστή ως ρίζα με δεξί παιδί-φύλλο κάποια μεταβλητή. Εν αντιθέσει, η μετάβαση από το τελικό δέντρο f_j στο αρχικό f_i, γίνεται ντετερμινιστικά, δηλαδή $d_j = 1$. Οι ζητούμενες πιθανότητες για κάθε μια από τις αντίστοιχες περιπτώσεις είναι:

$$P_{accept}^{\text{AddRoot}}(f_i \to f_j) = min\left\{1, \frac{\pi_j \cdot N}{\pi_i \cdot 1}\right\} \tag{2.3.22}$$

$$= min\left\{1, N \cdot e^{-(\mathcal{L}(f_j)-\mathcal{L}(f_i))}\right\}, \tag{2.3.23}$$

$$P_{accept}^{RemoveRoot}(f_i \to f_j) = min\left\{1, \frac{\pi_j \cdot 1}{\pi_i \cdot N}\right\} \tag{2.3.24}$$

$$= min\left\{1, \frac{1}{N} \cdot e^{-(\mathcal{L}(f_j)-\mathcal{L}(f_i))}\right\}. \tag{2.3.25}$$

- Αντικατάσταση απλού δέντρου: Αρχικά πρέπει να επιλεχθεί ομοιόμορφα, αν το απλό υποδέντρο της συμβολικής έκφρασης προς αντικατάσταση, θα είναι φύλλο ή τελεστής δυαδικός ή τελεστής μοναδιαίος, και αντίστοιχα τί εκ των τριών θα είναι το απλό δέντρο που θα πάρει την θέση του παλιού. Έστω n_{ij} οι αντίστοιχοι συνδυασμοί. Στην συνέχεια, για την τάξη του απλού δέντρου που επιλέχθηκε να αντικατασταθεί στο αρχικό δέντρο, θα πρέπει να μετρήσει κανείς όλα τα πιθανά σημεία όπου μπορεί να γίνει η αλλαγή, έστω ότι είναι Ω_i. Τέλος, υπάρχουν s_j επιλογές διαφορετικών απλών υποδέντρων για την τάξη του νέου απλού δέντρου που επιλέχθηκε. Άρα συνολικά λαμβάνουμε ότι $d_i = n_{ij} \cdot \Omega_i \cdot s_j$ γείτονες της έκφρασης f_i προσβάσιμοι μέσω της αντικατάστασης απλού δέντρου. Για την αντίστροφη μετάβαση λαμβάνουμε με ίδια λογική $d_j = n_{ji} \cdot \Omega_j \cdot s_i$. Επομένως η πιθανότητα αποδοχής της μετάβασης γράφεται:

$$P_{accept}^{ReplaceTree}(f_i \to f_j) = min\left\{1, \frac{\pi_j \cdot n_{ij} \cdot \Omega_i \cdot s_j}{\pi_i \cdot n_{ji} \cdot \Omega_j \cdot s_i}\right\} \tag{2.3.26}$$

$$= min\left\{1, \frac{n_{ij} \cdot \Omega_i \cdot s_j}{n_{ji} \cdot \Omega_j \cdot s_i} e^{-(\mathcal{L}(f_j)-\mathcal{L}(f_i))}\right\}. \tag{2.3.27}$$

Από όλες τις παραπάνω πιθανές κινήσεις της Μαρκοβιανής Αλυσίδας, η αντικατάσταση απλού δέντρου αποτελεί την κύρια μέθοδο εξερεύνησης του χώρου, καθώς δεν προκαλεί μεγάλη μεταβολή στο συμβολικό δέντρο. Η προσθήκη ή αφαίρεση ρίζας αποτελεί την μέθοδο με βάση την οποία τα δέντρα συμβολικών εκφράσεων μεγαλώνουν ή μικραίνουν, προσθέτοντας ή αφαιρώντας αντίστοιχα όρους στην συμβολική έκφραση. Τέλος, η αντικατάσταση κόμβου εν γένει αποτελεί μία επιπόλαια κίνηση, αφού προκαλεί τεράστια μεταβολή στην συμβολική έκφραση, και συνήθως δεν είναι αποτελεσματική κίνηση.

Η πιθανότητα επιλογής της κάθε μιας από τις εν λόγω μεθόδους μεταβολής του συμβολικού δέντρου, επιλέγεται αυθαίρετα από τον χρήστη. Για τους λόγους που αναφέρθηκαν παραπάνω, η αντικατάσταση κόμβου λαμβάνει μικρή πιθανότητα, ενώ οι υπόλοιπες μεταβολές

είναι σχεδόν ισοπίθανες.

Αξίζει να σημειωθεί πως η εν λόγω μέθοδος έχει αρκετές ομοιότητες σε σχέση με τον γενετικό αλγόριθμο. Συγκεκριμένα, η τυχαία αντικατάσταση κόμβου επιφέρει ίδια μεταβολή με τον τελεστή της μετάλλαξης, και οι αντίστοιχες πιθανότητες για να γίνει είναι πολύ μικρές. Επιπλέον, η προσθήκη ή αφαίρεση ρίζας, καθώς και η αντικατάσταση απλού δέντρου, προκαλούν μεταβολές στο συμβολικό δέντρο, οι οποίες θα μπορούσε να είναι αποτέλεσμα της δράσης του τελεστή της διασταύρωσης ανάμεσα σε δύο δέντρα του πληθυσμού. Τέλος, και οι δύο μέθοδοι απαιτούν τον υπολογισμό των κατάλληλων παραμέτρων προκειμένου, από την μία ο γενετικός αλγόριθμος για να βελτιώσει την προσαρμογή του μοντέλου με βάση την αντικειμενική συνάρτηση, από την άλλη η MCMC για τον υπολογισμό του BIC της ύστερης πιθανότητας.

Σε δεδομένα με θόρυβο, η τιμή του κριτηρίου BIC θα αναμένεται να είναι αρκετά μεγάλη. Σε αυτή την περίπτωση υπάρχει το ενδεχόμενο στο ελάχιστο περιγραφικό μήκος 2.3.9 η συνεισφορά της εκ των προτέρων πιθανότητας να είναι αμελητέα. Γι αυτό η εν λόγω έκφραση μπορεί να τροποποιηθεί και να γραφτεί ως εξής:

$$\mathcal{L}\left(f_i\right) \approx \frac{BIC\left(f_i\right)}{2T} - ln\left(P\left(f_i\right)\right), \tag{2.3.28}$$

όπου T μια σταθερά. Μεγάλες τιμές του T οδηγούν σε απλούστερα μοντέλα, ενώ μικρές τιμές ευνοούν την καλή προσαρμογή. Η παραπάνω λύση, είναι παρόμοια με την αντικειμενική συνάρτηση 2.3.1 στους γενετικούς αλγορίθμους.

Στην πράξη, η μέθοδος MCMC, όπως και ο γενετικός αλγόριθμος, ξεκινώντας από ένα απλό δέντρο, προσπαθεί να μεταβεί με τις κινήσεις που περιγράφηκαν παραπάνω σε γειτονικές συμβολικές εκφράσεις. Επομένως σε κάθε επανάληψη κινείται τοπικά αναζητώντας της καταλληλότερη έκφραση, και υπάρχει κίνδυνος εγκλωβισμού σε τοπικό ακρότατο. Επιπλέον είναι αυθαίρετη η επιλογή του συνόλου εκφράσεων μέσα από το οποίο καθορίζεται η εκ των προτέρων πιθανότητα $P\left(f_i\right)$. Αν και παρέχει σχετική ευελιξία καθώς μπορούν να χρησιμοποιηθούν εκφράσεις που επιλύουν παρεμφερή προβλήματα με αυτό που μας ενδιαφέρει, στην πράξη η επιλογή τους γίνεται με εμπειρικά κριτήρια ή κρίνοντας εκ του αποτελέσματος, δίχως να υπάρχει κάποιος κανόνας. Τέλος, η επιλογή κατάλληλων υπερπαραμέτρων της εκ των προτέρων πιθανότητας 2.3.11, γίνεται επαναληπτικά το οποίο αποτελεί εξαιρετικά χρονοβόρα διαδικασία, καθώς προαπαιτεί ένα ικανά μεγάλο μέγεθος πληθυσμού.

2.4 Επίλογος

Η κωδικοποίηση των συμβολικών εκφράσεων στον υπολογιστή με χρήση δυαδικών δέντρων, αποτελεί το θεμέλιο των υλοποιήσεων αλγορίθμων συμβολικής παλινδρόμησης, καθώς πάνω σε αυτό χτίζονται οι λειτουργίες του κάθε αλγορίθμου για την εύρεση της ζητούμενης συμβολικής έκφρασης. Ο τρόπος λειτουργίας του γενετικού αλγορίθμου εκ πρώτης όψεως φαίνεται να είναι αρκετά διαφορετικός σε σχέση με αυτόν της Μαρκοβιανής αλυσίδας Μόντε Κάρλο.

Εν τέλει όμως και οι δύο μέθοδοι λύνουν το πρόβλημα με παρόμοιες μεταβολές στα δέντρα συμβολικών εκφράσεων. Επιπλέον, η εύρεση κατάλληλης συμβολικής έκφρασης με καλή προσαρμογή και ταυτόχρονα χαμηλή περιπλοκότητα, αποτελεί ένα σύνθετο πρόβλημα το οποίο επιχειρούν με διαφορετικό τρόπο να αντιμετωπίσουν η κάθε μέθοδος. Σε κάθε περίπτωση η εκτίμηση των παραμέτρων του υποψήφιου μοντέλου, για τον υπολογισμό της προσαρμογής του, παραμένει αναγκαία, και μπορεί να γίνει με τις μεθόδους που αναπτύχθηκαν στο Κεφάλαιο 1.3.3, ή με την μέθοδο που θα παρουσιαστεί στο επόμενο κεφάλαιο. Αξίζει να σημειωθεί ότι υπάρχουν και άλλοι αλγόριθμοι συμβολικής παλινδρόμησης πέρα από αυτούς που αναφέρθηκαν, όπως για παράδειγμα το πακέτο AI-Feynman (Udrescu & Tegmark, 2020), το οποίο χρησιμοποιεί νευρωνικά δίκτυα για να παρεμβάλει τα δεδομένα και να ανιχνεύσει τυχόν συμμετρίες, όπως συμβαίνει συνήθως στους τύπους της φυσικής, διαιρώντας το αρχικό πρόβλημα σε υποπροβλήματα.

3 Εισαγωγή στο Πρόβλημα - Μεθοδολογία & Θεωρία

3.1 Πρόλογος

Το κεφάλαιο έρχεται σε συνέχεια του προηγούμενου, όπου έγινε μια εισαγωγή στα συμβολικά δέντρα και αναφέρθηκαν υλοποιήσεις που επιλύουν το πρόβλημα της συμβολικής παλινδρόμησης. Εστιάζοντας στους γενετικούς αλγορίθμους, γίνεται αναλυτική παρουσίαση των κυριότερων εναλλακτικών που υπάρχουν στους γενετικούς τελεστές της επιλογής, διασταύρωσης, μετάλλαξης, και εξέλιξης του πληθυσμού από επανάληψη σε επανάληψη. Στην συνέχεια, παρουσιάζεται η μέθοδος Levenberg-Marquardt για την εκτίμηση των παραμέτρων του μοντέλου, ως βελτίωση των μεθόδων που παρουσιάστηκαν στο Κεφάλαιο 1. Τέλος, η υλοποίηση του γενετικού αλγορίθμου της εργασίας, για την επίλυση του προβλήματος της συμβολικής παλινδρόμησης, εφαρμόζεται σε ένα πρόβλημα με γνωστή λύση και παρουσιάζονται τα αποτελέσματα.

3.2 Υλοποίηση γενετικού αλγορίθμου

3.2.1 Αρχικοποίηση πληθυσμού

Ο πληθυσμός του γενετικού αλγορίθμου στο πρόβλημα της συμβολικής παλινδρόμησης, αποτελείται από συμβολικά δυαδικά δέντρα, τα οποία αντιπροσωπεύουν μία συμβολική έκφραση. Η αρχικοποίηση του πληθυσμού γίνεται τυχαία, με απλά δέντρα ύψους 1, ως εξής:

1. Επιλέγεται τυχαία και ισοπίθανα ένας τελεστής ως κόμβος-ρίζα του συμβολικού δέντρου, από το σύνολο των τελεστών που έχει καθορίσει ο χρήστης.

2. Αν ο τελεστής είναι δυαδικός, τότε επιλέγονται τυχαία δύο μεταβλητές ως παιδιά της ρίζας, διαφορετικά ο τελεστής είναι μοναδιαίος και επιλέγεται μία μεταβλητή αντίστοιχα.

Η διαδικασία δημιουργίας νέων υποψηφίων, όπως αυτών του Διαγράμματος 3.1, επαναλαμβάνεται μέχρι ο πληθυσμός να έχει το μέγεθος που έχει καθορίσει ο χρήστης.

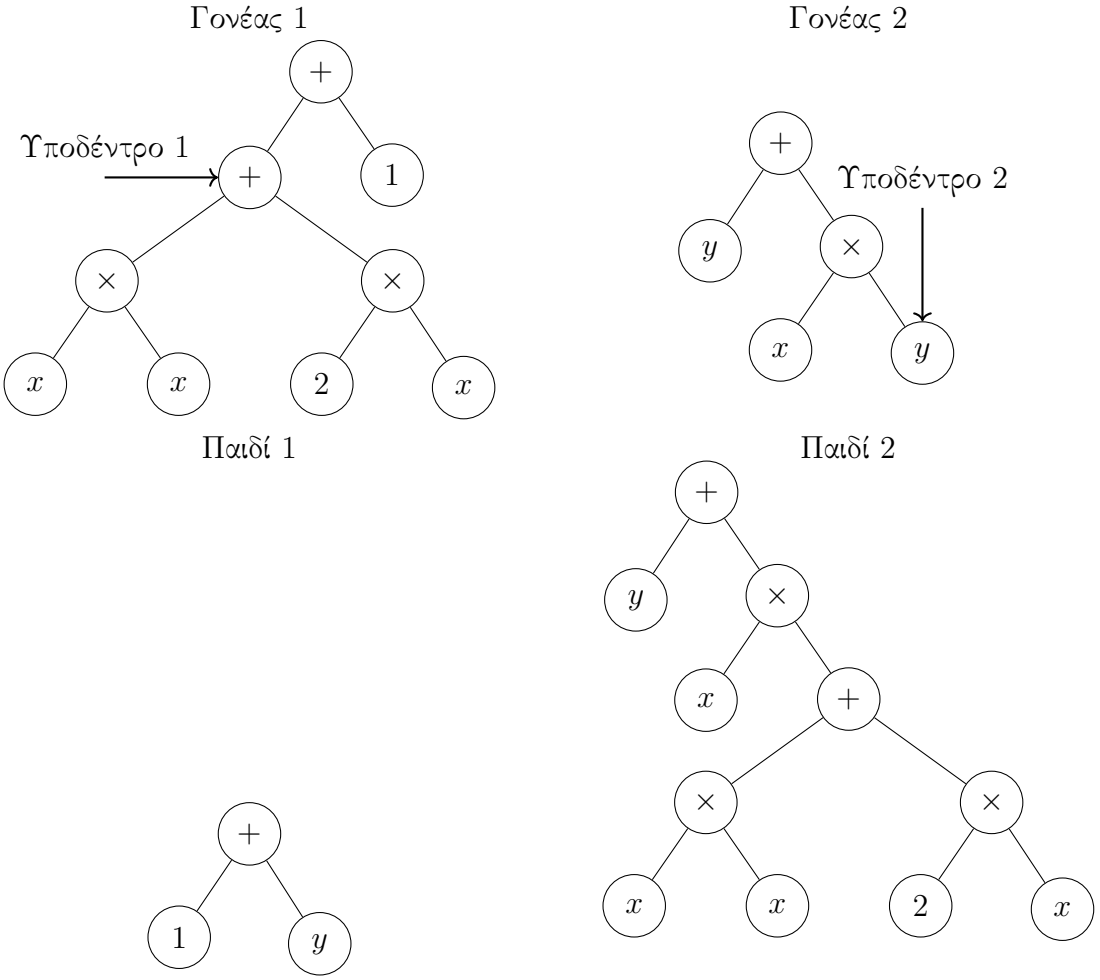
Διάγραμμα 3.1: Τυχαία δέντρα ύψους 1.

3.2.2 Επιλογή για διασταύρωση

Ο τελεστής της επιλογής υποψηφίων για διασταύρωση, στην περίπτωση της συμβολικής παλινδρόμησης δεν διαφέρει σε κάτι από τους κλασικούς γενετικούς αλγορίθμους. Επομένως τα πλεονεκτήματα και μειονεκτήματα που αναφέρθηκαν στην Υποενότητα 1.4.1 για κάθε πιθανή εναλλακτική ισχύουν. Συνεπώς, ο εν λόγω τελεστής επιλέγεται να λειτουργεί με τουρνουά ανάμεσα σε 2 χρωμοσώματα, προκειμένου να διατηρηθεί η ποικιλομορφία του πληθυσμού και να αποφευχθεί κατά το δυνατόν η πρόωρη σύγκλιση.

3.2.3 Τελεστής διασταύρωσης

Ο τελεστής διασταύρωσης αποτελεί το σημαντικότερο χαρακτηριστικό του γενετικού αλγορίθμου, καθώς οι επιλεγμένοι υποψήφιοι καλούνται να συνδυάσουν τα καλύτερα χαρακτηριστικά τους. Συνδυάζοντας την γενετική πληροφορία που κωδικοποιεί το χρωμόσωμα τους, παράγονται νέες λύσεις που θα καθοδηγήσουν τον πληθυσμό στην επίλυση του προβλήματος.



Διάγραμμα 3.2: Απλή διασταύρωση ενός σημείου ανάμεσα σε δύο υποδέντρα.

Όλες οι μορφές διασταύρωσης που θα παρουσιαστούν σε αυτή την υποενότητα, αποτε-

λούν παραλλαγή της απλής διασταύρωσης σε ένα σημείο μεταξύ δύο υποδέντρων. Για την απλή διασταύρωση επιλέγονται τυχαία και ισοπίθανα δύο κόμβοι, ένας από κάθε δέντρο, και ανταλλάζουν θέσεις τα αντίστοιχα υποδέντρα.

Πραγματοποιώντας τυχαία απλή διασταύρωση, οι πιθανότητες οι νέοι υποψήφιοι να φέρουν καλύτερα χαρακτηριστικά από τους γονείς είναι μικρές. Αυτό οφείλεται στο γεγονός ότι το ενδεχόμενο τα εν λόγω δέντρα να μοιράζονται κάποια ομοιότητα είναι σχεδόν μηδαμινή και μετατρέπει τον γενετικό αλγόριθμο, σε αλγόριθμο τυχαίας αναζήτησης. Στο Διάγραμμα 3.2, φαίνεται μία ακραία περίπτωση όπου τα σημεία διασταύρωσης, επιλέχθηκαν να είναι κοντά στην ρίζα και σε κάποιο φύλλο αντίστοιχα, με αποτέλεσμα την ριζική αλλαγή των συμβολικών εκφράσεων που κωδικοποιεί το κάθε παιδί.

Για την βελτίωση της αποτελεσματικότητας του τελεστή της διασταύρωσης, το ενδιαφέρον αρχικά στρέφεται στην αναζήτηση όμοιων δομικών χαρακτηριστικών μεταξύ των δύο υποδέντρων. Συγκεκριμένα θα θέλαμε η επιλογή του σημείου διασταύρωσης να μην είναι εντελώς τυχαία, αλλά να αποδίδει μεγαλύτερη πιθανότητα σε ορισμένους κόμβους.

Σε περίπτωση που μας ενδιαφέρει η ανταλλαγή μεγαλύτερης γενετικής πληροφορίας μεταξύ των δέντρων, τότε θα πρέπει να αποδίδεται μεγαλύτερη πιθανότητα επιλογής στους κόμβους που βρίσκονται πιο κοντά στην ρίζα, και κατ΄ επέκταση έχουν μεγαλύτερο υποδέντρο. Αυτό όμως θα είχε σαν αποτέλεσμα να προκύπτουν εξαιρετικά μεγάλες συμβολικές εκφράσεις, οι οποίες δεν θα κωδικοποιούν ουσιώδη πληροφορία, αλλά θα περιέχουν «φλύαρους» όρους που έχουν αποστηθίσει τα δεδομένα εκπαίδευσης.

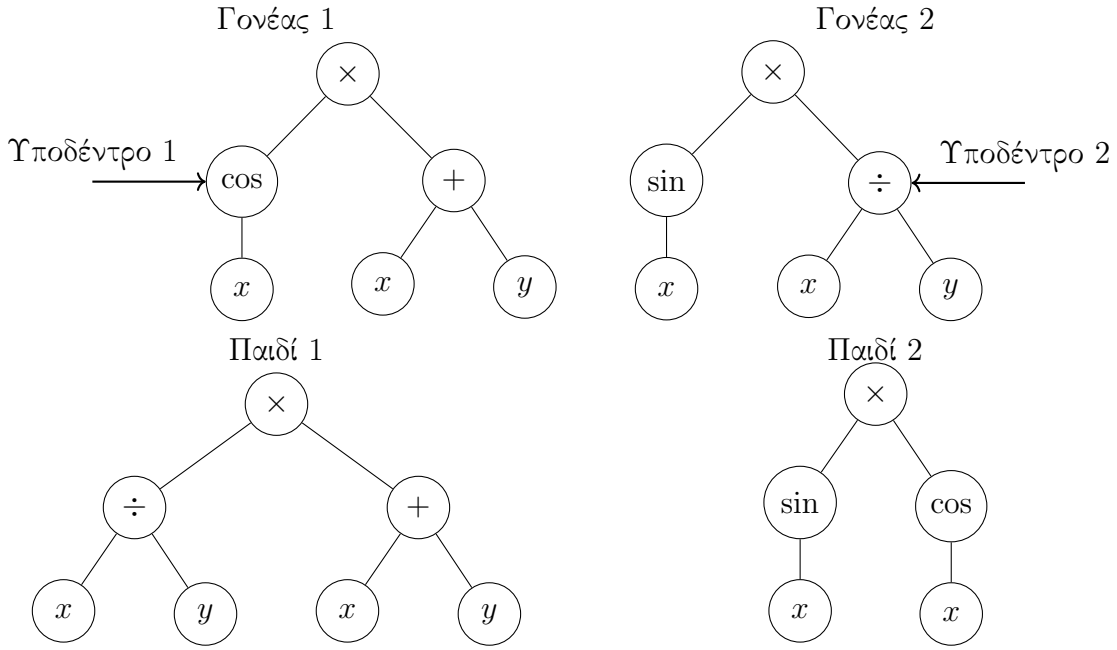
Εναλλακτική επιλογή, αποτελεί η ανταλλαγή υποδέντρων ίδιου ύψους. Ύψος ενός δέντρου ορίζεται να είναι το μέγεθος του μέγιστου μονοπατιού μεταξύ της ρίζας και των φύλλων του, και υπολογίζεται με βάση με τον Αλγόριθμο 5. Ως μέγεθος μονοπατιού, αναφερόμαστε στο πλήθος των ακμών που ενώνουν τους διαδοχικούς κόμβους στο ζητούμενο μονοπάτι.

Αλγόριθμος 5 Υπολογισμός του ύψους ενός δέντρου.

```
function height(p)
    if p is leaf node
        return 0
    else
        hl = 0
        hr = 0
        if p has left child l
            hl = 1+height(l)
        if p has right child r
            hr = 1+height(r)
    return max(hl,hr)
```

Κατά την διασταύρωση με υποδέντρα ίδιου ύψους, αρχικά υπολογίζεται το ύψος όλων των υποδέντρων που περιέχει ο κάθε γονιός. Η ρίζα του αρχικού δέντρου έχει το μέγιστο

ύψος h_{root}, επομένως τα πιθανά ύψη είναι $\{0, \ldots, h_{root}\}$. Στην συνέχεια επιλέγεται τυχαία μια έγκυρη τιμή ύψους $h \in \{0, \ldots, min\,(h_1, h_2)\}$, όπου h_1, h_2 τα αντίστοιχα ύψη του πρώτου και δεύτερου γονιού. Τέλος, επιλέγεται τυχαία ως σημείο διασταύρωσης, εκεί όπου τα αντίστοιχα υποδέντρα σε κάθε γονιό έχουν ίδια τιμή ύψους ίση με h. Όπως φαίνεται και στο Διάγραμμα 3.3, τα δέντρα που προκύπτουν έχουν ίδιο ύψος με το αρχικό. Ενδεχομένως όμως να ήταν προτιμότερο ο αλγόριθμος να επιλέξει ως σημείο διασταύρωσης τον κόμβο sin στο δεύτερο υποδέντρο, αφού θα είχαν περισσότερα κοινά μεταξύ τους τα δύο υποδέντρα με ρίζα τριγωνομετρική συνάρτηση.



Διάγραμμα 3.3: Διασταύρωση ενός σημείου ανάμεσα σε δύο υποδέντρα ίδιου ύψους $h = 1$.

Για τον σκοπό αυτό αντί να εστιάζει ο τελεστής της διασταύρωσης σε δομικά χαρακτηριστικά του συμβολικού δέντρου θα μπορούσε να ελέγχει την συνάφεια (context) του δέντρου. Η συνάφεια κατά την διάρκεια της διασταύρωσης, μπορεί να μελετηθεί με τους εξής τρόπους:

1. Συνάφεια του συμβολικού δέντρου σε σχέση με τα υποδέντρα του:

   Σε αυτή την περίπτωση, εξετάζεται για κάθε υποδέντρο η μεταβολή στην τιμή της αντικειμενικής συνάρτησης, που θα επιφέρει η αντικατάστασή του με κάποιον τυχαίο κόμβο-φύλλο. Μικρές μεταβολές σημαίνουν μικρή συνεισφορά του υποδέντρου στην ικανότητα προσαρμογής του εξεταζόμενου μοντέλου, κατά συνέπεια το χειρότερο υποδέντρο επιλέγεται για αντικατάσταση μέσω της διασταύρωσης.

2. Συνάφεια του συμβολικού δέντρου σε σχέση με το προς αντικατάσταση υποδέντρο του δεύτερου γονιού:

   Επιλέγεται τυχαία ένα σημείο διασταύρωσης στον δεύτερο γονιό, και το αντίστοιχο υποδέντρο διασταυρώνεται σε όλες τις πιθανές θέσεις του πρώτου γονιού. Από τους

απογόνους που προκύπτουν επιλέγεται εκείνος με την καλύτερη τιμή της αντικειμενικής συνάρτησης του γενετικού αλγορίθμου.

Κάθε μία από τις παραπάνω επιλογές, περιλαμβάνει μια εξαντλητική αναζήτηση, είτε σε καθένα από τα επιλεγμένα για διασταύρωση δέντρα, είτε στους συνδυασμούς απογόνων που μπορούν να δημιουργήσουν. Αυτό σε αρκετά προβλήματα ενδέχεται να είναι απαγορευτικό, είτε επειδή η διάσταση των δεδομένων είναι μεγάλη, είτε επειδή τα δέντρα συμβολικών εκφράσεων είναι αρκετά μεγάλα, αυξάνοντας εκθετικά τους συνδυασμούς που μπορούν να γίνουν μεταξύ των σημείων διασταύρωσης ώστε να προκύψουν ικανοί απόγονοι.

Αν εστιάσει κανείς στις παραπάνω μεθόδους, οι περισσότερες έχουν ως κοινό χαρακτηριστικό να βελτιώσουν την αποτελεσματικότητα του γενετικού τελεστή της διασταύρωσης. Άλλοτε αυτό επιχειρείται μεγιστοποιώντας την ανταλλαγή γενετικής πληροφορίας μεταξύ των συμβολικών δέντρων του πληθυσμού, άλλοτε μέσω του περιορισμού των σημείων διασταύρωσης ώστε να μην προκληθεί τεράστια μεταβολή, και άλλες φορές αναζητώντας εξαντλητικά τον συνδυασμό που θα δώσει τους καλύτερους απογόνους στον πληθυσμό. Επομένως, η απάντηση στο πρόβλημα ενδέχεται να μην βρίσκεται ούτε στην δομή, ούτε στην συνάφεια, αλλά στην σημασιολογία (semantics) των συμβολικών δέντρων.

Η έννοια της σημασιολογίας είναι αρκετά αφηρημένη, όμως για τα συμβολικά δέντρα έχει δοθεί ένας προσεγγιστικός ορισμός:

> *Έστω F μια συνάρτηση που κωδικοποιείται από το δέντρο T, με πεδίο ορισμού D και $P = \{p_1, \ldots, p_N\} \subseteq D$. Τότε η δειγματοληπτική σημασιολογία (sampling semantics) του T στο P είναι το σύνολο $S = \{s_1, \ldots, s_N\}$, όπου $s_i = F(p_i)$ με $i = 1, \ldots, N$.(Uy et al., 2011)*

Η δειγματοληπτική σημασιολογία βάση του παραπάνω ορισμού, στην πραγματικότητα δεν είναι τίποτα περισσότερο, παρά η αριθμητική τιμή του εκάστοτε δέντρου T σε ένα σύνολο σημείων P, του πεδίου ορισμού του. Αυτό δίνει την δυνατότητα, να ελέγξει κανείς την σημασιολογική διαφορά ανάμεσα σε δύο δέντρα, και να αποφανθεί κατά πόσο είναι κοντά το ένα στο άλλο. Αρκεί λοιπόν να γίνει κατάλληλη επιλογή ενός κοινού συνόλου P από N στοιχεία, τότε η σημασιολογική διαφορά υπολογίζεται ως:

$$\delta = \frac{1}{N} \left\| S^{(1)} - S^{(2)} \right\|_1 \tag{3.2.1}$$

$$= \frac{1}{N} \left(\left| s_1^{(1)} - s_1^{(2)} \right| + \ldots + \left| s_N^{(1)} - s_N^{(2)} \right| \right), \tag{3.2.2}$$

όπου $s_i^{(1)}$ και $s_i^{(2)}$ οι αντίστοιχες τιμές των συμβολικών δέντρων σε καθένα από τα σημεία του P, για $i = 1, \ldots, N$. Η παραπάνω διαφορά θα μπορούσε να υπολογιστεί και με άλλη νόρμα, όπως την νόρμα 2 και τότε η Εξίσωση 3.2.1 θα λάμβανε την μορφή της μέσης τετραγωνικής απόστασης. Το σύνολο P, θα πρέπει να περιλαμβάνει αρκετά σημεία ώστε

το αποτέλεσμα να μην είναι παραπλανητικό, και ταυτόχρονα να μην προκαλεί μεγάλη υπολογιστική επιβάρυνση στην διαδικασία της διασταύρωσης. Επομένως, ορίζεται να είναι μια τυχαία επιλογή παρατηρήσεων από το αρχικό σύνολο δεδομένων που δίνεται στον γενετικό αλγόριθμο.

Η σημασιολογική διαφορά στον τελεστή της διασταύρωσης, έχει ως στόχο να βρει δύο σημεία διασταύρωσης σε κάθε γονιό, τέτοια ώστε τα αντίστοιχα υποδέντρα να μην είναι εντελώς διαφορετικά το ένα από το άλλο. Δηλαδή, για κάποια σταθερά $\delta_{ub} > 0$ πρέπει να ισχύει:

$$\delta \leq \delta_{ub}. \tag{3.2.3}$$

Όταν ισχύει η παραπάνω ανίσωση τα δύο υποδέντρα είναι σημασιολογικά ισοδύναμα (semantic equivalence). Με αυτό τον τρόπο, η μεταβολή των υποδέντρων είναι μικρότερη, μειώνοντας το ενδεχόμενο οι απόγονοι που θα προκύψουν να έχουν πολύ χειρότερη προσαρμογή, αφού ο χώρος των συμβολικών εκφράσεων εξερευνάται με μικρότερες μεταβολές. Από την άλλη, αν τα δύο υποδέντρα ταυτίζονται σε μεγάλο βαθμό, τότε ανταλλάσσεται πολύ μικρή γενετική πληροφορία, με αποτέλεσμα ο πληθυσμός να παραμένει σχεδόν στάσιμος χωρίς να βελτιώνεται η λύση. Γι αυτό η ανισότητα 3.2.3 δεν επαρκεί ως κριτήριο για την διασταύρωση των υποδέντρων, αλλά χρησιμοποιείται η ακόλουθη:

$$\delta_{lb} \leq \delta \leq \delta_{ub} \tag{3.2.4}$$

όπου $\delta_{lb} > 0$ μία σταθερά ως κάτω όριο. Δηλαδή, αν ισχύει η 3.2.4, τα αντίστοιχα υποδέντρα είναι σημασιολογικά όμοια (semantic similarity).

Η παραπάνω διαδικασία είναι εμπνευσμένη από την βιολογία, καθώς υπάρχουν γονίδια στα χρωμοσώματα των θηλαστικών, τα οποία ενεργοποιούν το ανοσοποιητικό σύστημα σε απότομες γενετικές μεταβολές, ενώ φροντίζουν τα νέα χρωμοσώματα που δημιουργούνται να έχουν παρόμοια, όχι όμως ολόιδια γονίδια με τους γονείς, προκειμένου να εισάγεται νέα γενετική πληροφορία στον πληθυσμό.

Αλγοριθμικά η σημασιολογική διαφορά καθοδηγεί τον τελεστή της διασταύρωσης με τον εξής τρόπο:

1. Γίνεται τυχαία επιλογή N παρατηρήσεων από το αρχικό σύνολο δεδομένων, τα οποία προστίθενται στο σύνολο P.

2. Επιλέγεται ένα τυχαίο υποδέντρο στον πρώτο γονιό, διαλέγοντας με ομοιόμορφη πιθανότητα κάποιον από τους κόμβους του αντίστοιχου συμβολικού δέντρου.

3. Επιλέγεται τυχαία ένα υποδέντρο στον δεύτερο γονιό.

   (α′) Αν η δειγματοληπτική σημασιολογική διαφορά τους (βλ. Εξίσωση 3.2.1) ικανοποιεί την ανισότητα 3.2.4, τότε πραγματοποιείται διασταύρωση των δύο υπο-

δέντρων.

(β′) Αν δεν ικανοποιείται η ανισότητα 3.2.4, τότε τα υποδέντρα δεν είναι σημασιολογικά όμοια, και επαναλαμβάνεται το βήμα 3, με νέο τυχαίο υποδέντρο του δεύτερου γονιού

Η διαδικασία εύρεσης κατάλληλου υποδέντρου, επαναλαμβάνεται το πολύ 10 φορές, καθώς υπάρχει το ενδεχόμενο για τα όρια δ_{lb} και δ_{ub} που έχουν επιλεχθεί να μην ικανοποιείται η ανισότητα 3.2.4. Τότε πραγματοποιείται τυχαία διασταύρωση ενός σημείου.

3.2.4 Τελεστής μετάλλαξης

Ο τελεστής της μετάλλαξης έχει ως στόχο να προκαλέσει τυχαία αλλαγή σε κάποιο γονίδιο του χρωμοσώματος του κάθε απογόνου που προκύπτει μετά την δράση του τελετή της διασταύρωσης. Αυτό δίνει την δυνατότητα να εξερευνηθούν λύσεις οι οποίες δεν θα ήταν δυνατό να δημιουργηθούν μέσω της διασταύρωσης, συμβάλλοντας στην διατήρηση της γενετικής ποικιλομορφίας του πληθυσμού. Η μετάλλαξη μπορεί να φέρει θετικά αποτελέσματα, στην περίπτωση που ο αλγόριθμος έχει εγκλωβιστεί σε τοπικό ακρότατο, όμως θα πρέπει η αντίστοιχη πιθανότητα δράσης του να παραμείνει μικρή, καθώς όπως ειπώθηκε και στην εισαγωγή, μεγάλες τιμές μετατρέπουν τον γενετικό αλγόριθμο, σε αλγόριθμο τυχαίας αναζήτησης, μειώνοντας την αποτελεσματικότητά του.

Η πιο απλή μορφή μετάλλαξης που μπορεί να συμβεί σε γενετικό αλγόριθμο συμβολικής παλινδρόμησης με συμβολικά δέντρα ως χρωμοσώματα, είναι η τυχαία επιλογή με ομοιόμορφη πιθανότητα ενός κόμβου, και αντικατάστασή του με κάποιον τυχαίο. Ο νέος κόμβος θα πρέπει να είναι συμβατός με τον προηγούμενο, επομένως διακρίνουμε τις εξής περιπτώσεις για τον κόμβο που επιλέγεται:

1. Αν ο κόμβος είναι φύλλο, τότε φιλοξενεί κάποια μεταβλητή. Επομένως γίνεται τυχαία αντικατάσταση από κάποια άλλη μεταβλητή.

2. Αν είναι εσωτερικός, τότε περιέχει κάποιον τελεστή:

   (α′) Αν είναι δυαδικός, αντικαθίσταται από άλλον δυαδικό τελεστή.

   (β′) Αλλιώς είναι μοναδιαίος και πραγματοποιείται η ανάλογη αντικατάσταση.

Παραδοσιακά η μετάλλαξη στο χρωμόσωμα γίνεται αποδεκτή, αρκεί ο νέος υποψήφιος να έχει καλύτερη τιμή της αντικειμενικής συνάρτησης. Όταν όμως οι υποψήφιοι του πληθυσμού κωδικοποιούν παρόμοια γενετική πληροφορία, ενδέχεται μία μετάλλαξη να ωφελήσει την επόμενη γενιά, ακόμα και όταν το μεταλλαγμένο χρωμόσωμα έχει χειρότερη προσαρμογή. Για αυτό τον λόγο ορίζεται η πιθανότητα αποδοχής της μετάλλαξης:

$$p_{accept} = e^{-\Delta L/T} \tag{3.2.5}$$

όπου ΔL η μεταβολή στην αντικειμενική συνάρτηση και $T \in [0,1]$ η θερμοκρασία. Μικρές τιμές της θερμοκρασίας σημαίνουν μεγάλη ποικιλομορφία του πληθυσμού, ενώ μεγάλες τιμές, μικρή ποικιλομορφία. Η ποικιλομορφία του πληθυσμού μπορεί να υπολογιστεί μετρώντας την σχετική συχνότητα $\hat{p_o}$ εμφάνισης του κάθε τελεστή στον πληθυσμό, $o \in \mathcal{O} = \{+, \times, e^x, \ldots\}$. Στην συνέχεια, υπολογίζεται η εντροπία του πληθυσμού ακολούθως:

$$\mathcal{H} = -\sum_{o \in \mathcal{O}} \hat{p_o} \log\left(\hat{p_o}\right). \tag{3.2.6}$$

Αποδεικνύεται με χρήση των πολλαπλασιαστών Lagrange, ότι η μέγιστη τιμή της εντροπίας προκύπτει όταν οι σχετικές συχνότητες ταυτίζονται. Πιο συγκεκριμένα, έχουμε ότι:

$$g\left(p_{o,1}, \ldots p_{o,N}, \lambda\right) = -\sum_{o \in \mathcal{O}} \hat{p_o} \log\left(\hat{p_o}\right) + \lambda \left(\sum_{o \in \mathcal{O}} \hat{p_o} - 1\right) \tag{3.2.7}$$

$$\frac{\partial g}{\partial p_o} = -\log\left(\hat{p_o}\right) - 1 + \lambda = 0 \rightarrow \hat{p_o} = e^{\lambda - 1}, \qquad \forall o \in \mathcal{O} \tag{3.2.8}$$

$$\frac{\partial g}{\partial \lambda} = \sum_{o \in \mathcal{O}} \hat{p_o} - 1 = 0 \rightarrow \sum_{o \in \mathcal{O}} \hat{p_o} = 1. \tag{3.2.9}$$

Η Εξίσωση 3.2.8 δείχνει ότι όλες οι πιθανότητες είναι σταθερές και ίσες με κάποιο p, άρα αντικαθιστώντας στην 3.2.9, η αντίστοιχη τιμή της πιθανότητας είναι:

$$\sum_{o \in \mathcal{O}} \hat{p_o} = 1 \rightarrow p = \frac{1}{N} \tag{3.2.10}$$

όπου N το πλήθος των τελεστών στο σύνολο $\mathcal{O}$. Αντικαθιστώντας το παραπάνω αποτέλεσμα στην εντροπία 3.2.6, η μέγιστη τιμή της υπολογίζεται να είναι:

$$\mathcal{H}_{max} = +\log\left(N\right). \tag{3.2.11}$$

Επομένως η θερμοκρασία της Εξίσωσης 3.2.5, ορίζεται βάση της εντροπίας να είναι:

$$T = 1 - \frac{\mathcal{H}}{\mathcal{H}_{max}} = 1 - \frac{\mathcal{H}}{\log\left(N\right)}. \tag{3.2.12}$$

Η παραπάνω έκφραση της θερμοκρασίας είναι πράγματι φραγμένη στο διάστημα $[0,1]$ και αποδίδει μεγαλύτερη πιθανότητα αποδοχής μιας «κακής» μετάλλαξης όταν η εντροπία του πληθυσμού είναι μικρή, άρα υπάρχει ανάγκη για νέο γενετικό υλικό.

3.2.5 Εισαγωγή των νέων λύσεων στον πληθυσμό

Μετά την διασταύρωση και τη μετάλλαξη, οι νέοι υποψήφιοι πρέπει να ενταχθούν στον υπάρχοντα πληθυσμό. Οι θέσεις όμως είναι περιορισμένες, δηλαδή κάποιες λύσεις, είτε από αυτές που προϋπήρχαν, είτε από εκείνες που προέκυψαν μετά την διασταύρωση, θα πρέπει να αποχωρήσουν.

Τα κριτήρια με βάση τα οποία γίνεται αυτή η επιλογή, πρέπει να είναι προσεκτικά επιλεγμένα, προκειμένου να διατηρηθούν στον πληθυσμό οι «καλοί» υποψήφιοι. Ο χαρακτηρισμός ενός υποψηφίου ως «καλός», μπορεί να γίνει με διάφορους τρόπους, όπως:

- Την καλή προσαρμογή.
- Την ποικιλομορφία που προσφέρουν στον πληθυσμό τα γονίδιά του.
- Την περιπλοκότητά του.
- Την ηλικία του.

Ανεξάρτητα της επιλογής του κριτηρίου, ο στόχος παραμένει κοινός, δηλαδή να διατηρηθούν οι καλύτεροι υποψήφιοι ώστε στην επόμενη επανάληψη του γενετικού αλγορίθμου, να συνεχίσει να βελτιώνει την λύση και να μην παραμένει στάσιμη η διαδικασία της εξέλιξης. Το κλασσικό κριτήριο επιλογής στους γενετικούς αλγορίθμους, είναι να μένουν εκτός πληθυσμού οι υποψήφιοι με την χειρότερη προσαρμογή. Παρόλα αυτά θα μπορούσε κανείς να συνδυάσει παραπάνω από ένα κριτήρια για να συνθέσει το πληθυσμό της επόμενης γενιάς.

Στα προβλήματα συμβολικής παλινδρόμησης, συνηθίζεται να αξιοποιείται η ηλικία του υποψηφίου. Για παράδειγμα στο πακέτο συμβολικής παλινδρόμησης SymbolicRegression.jl, οι νέοι υποψήφιοι εισέρχονται στις θέσεις των γηραιότερων χρωμοσωμάτων του πληθυσμού (Cranmer, 2023). Ο λόγος πίσω από αυτή την απόφαση, κρύβεται στα θετικά αποτελέσματα που έχει η χρήση της ηλικίας ως ελεγκτικός μηχανισμός σε άλλες μεθόδους μηχανικής μάθησης. Σε αυτό το σημείο, αξίζει να σημειωθούν οι τρόποι με τους οποίους μπορεί να καθοριστεί η ηλικία ενός υποψηφίου:

- Με την κυριολεκτική έννοια της λέξης: Για κάθε υποψήφιο καταγράφεται η χρονική στιγμή που δημιουργήθηκε, επομένως η ηλικία του σε μετέπειτα χρονική στιγμή καθορίζεται με αφαίρεση της αρχικής.
- Με την έννοια της γενιάς: Εφαρμόζεται μόνο σε γενεαλογικούς γενετικούς αλγορίθμους. Η ηλικία των χρωμοσωμάτων του πληθυσμού αυξάνεται κατά ένα σε κάθε επανάληψη του αλγορίθμου.
- Με την έννοια της συμμετοχής του χρωμοσώματος στους γενετικούς τελεστές: Εφαρμόζεται μόνο σε γενετικούς αλγορίθμους σταθερής κατάστασης. Κάθε φορά που το χρωμόσωμα επιλέγεται για διασταύρωση ή πραγματοποιείται μετάλλαξη, η ηλικία του αυξάνεται κατά ένα.

Στις δύο τελευταίες περιπτώσεις, όταν αρχικοποιείται ο πληθυσμός όλοι ο υποψήφιοι έχουν ηλικία ένα. Κατά την διάρκεια της διασταύρωσης η μέγιστη ηλικία των δύο γονιών μεταφέρεται στα παιδιά.

Όταν θέλει κανείς να εστιάσει τόσο στην ηλικία, όσο και στην καλή προσαρμογή για την δημιουργία του επόμενου πληθυσμού, θα πρέπει η επιλογή να γίνεται έτσι ώστε οι μικροί ηλικιακά υποψήφιοι να μην κυριαρχούνται από τους γηραιότερους, οι οποίοι αναμένεται να έχουν καλύτερη προσαρμογή.

Μία αρκετά γνωστή υλοποίηση, αποτελεί ο αλγόριθμος ALPS (Age Layered Population Structure) (Hornby, 2006). Πιο συγκεκριμένα, ο πληθυσμός χωρίζεται σε ξεχωριστούς υποπληθυσμούς, στους οποίους κατατάσσονται οι υποψήφιοι ανάλογα την ηλικία τους. Σε κάθε υποπληθυσμό τρέχει παράλληλα ένας γενετικός αλγόριθμος, ο οποίος μέσα από τους γενετικούς τελεστές δημιουργεί νέες λύσεις. Μόλις ολοκληρωθεί μια επανάληψη του γενετικού αλγορίθμου σε κάθε υποπληθυσμό, ενημερώνονται οι ηλικίες των χρωμοσωμάτων και γίνονται οι απαραίτητες μετακινήσεις. Το σημαντικότερο πλεονέκτημα αυτής της μεθόδου αποτελεί η ύπαρξη ενός υποπληθυσμού που φιλοξενεί τα νεότερα σε ηλικία χρωμοσώματα. Εκεί σε κάθε επανάληψη εισάγονται νέες τυχαίες λύσεις, οι οποίες επειδή συνυπάρχουν με άλλους υποψηφίους παρόμοιας ηλικίας, δεν κινδυνεύουν να κυριαρχηθούν από τα χρωμοσώματα με καλύτερη προσαρμογή και μεγαλύτερη ηλικία. Αυτό συμβάλλει στην διατήρηση της ποικιλομορφίας και στην πραγματικότητα δίνει την δυνατότητα στον γενετικό αλγόριθμο να ανανεώνει τον πληθυσμό του δραπετεύοντας από τοπικά ελάχιστα.

Φυσικά, το μειονέκτημα του αλγορίθμου ALPS, είναι η εισαγωγή νέων παραμέτρων στο πρόβλημα, όπως τον αριθμό των υποπληθυσμών και τα ηλικιακά όρια ανάμεσα σε κάθε υποπληθυσμό. Μία παραλλαγή του ALPS, αντιμετωπίζει τα παραπάνω μειονεκτήματα, παραλείποντας το κομμάτι της διαμέρισης του πληθυσμού (Schmidt & Lipson, 2011). Αντίθετα τα νεότερα ηλικιακά χρωμοσώματα συνυπάρχουν με τα παλαιότερα, όμως δεν κυριαρχούνται από την καλύτερη προσαρμογή των γηραιότερων, καθώς το κριτήριο επιλογής τους εκφράζεται σαν πρόβλημα Pareto:

$$criteria = w \times loss + (1 - w) \times age. \tag{3.2.13}$$

Σε κάθε επανάληψη εισάγεται ένας νέος υποψήφιος στον πληθυσμό, και η Εξίσωση 3.2.13 εξασφαλίζει ότι σε περίπτωση εγκλωβισμού σε τοπικό ελάχιστο, όπου θα αυξάνεται η ηλικία των λύσεων χωρίς να βελτιώνεται η προσαρμογή, οι νεότερες λύσεις θα παραμείνουν ανανεώνοντας τον πληθυσμό.

Εν συνεχεία, αφού υπολογιστεί η τιμή του κριτηρίου, μέσω επιλογής τουρνουά ανάμεσα σε κ-χρωμοσώματα, επιλέγεται το καλύτερο βάση της Εξίσωσης 3.2.13 και προστίθεται στον πληθυσμό της επόμενης γενιάς. Η επιλογή τουρνουά επαναλαμβάνεται μέχρι ο νέος πληθυσμός να έχει το επιθυμητό μέγεθος.

Ο παραπάνω αλγόριθμος για την επιλογή του πληθυσμού της επόμενης επανάληψης στον

γενετικό αλγόριθμο, καλείται Age-Fitness Pareto (Schmidt & Lipson, 2011), και σε συγκριτικές δοκιμές με τον ALPS, η επίδοση είναι παρόμοια. Το υπολογιστικό κόστος και η ταχύτητα σύγκλισης είναι ελαφρώς καλύτερη σε απλούστερα προβλήματα με τον αλγόριθμο ALPS, όμως σε πιο περίπλοκα ο Age-Fitness Pareto υπερτερεί με μικρή διαφορά.

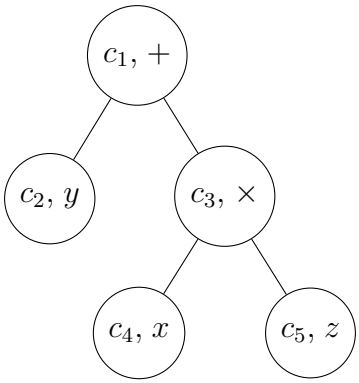
3.3 Εκτίμηση των παραμέτρων του μοντέλου

Η προσαρμογή του μοντέλου της συμβολικής παλινδρόμησης, εξαρτάται από την ύπαρξη των κατάλληλων τελεστών μέσα στο μοντέλο, αλλά και την εκτίμηση των αντίστοιχων παραμέτρων του μοντέλου. Στον γενετικό αλγόριθμο που υλοποιήθηκε σε αυτή την εργασία, η εύρεση των κατάλληλων παραμέτρων, γίνεται κατά την δημιουργία, διασταύρωση, και μετάλλαξη των χρωμοσωμάτων, εξασφαλίζοντας την καλύτερη δυνατή προσαρμογή στην αντικειμενική συνάρτηση.

3.3.1 Κωδικοποίηση των παραμέτρων στο συμβολικό δέντρο

Επειδή ένα μοντέλο ενδέχεται να έχει πολλές παραμέτρους (ή σταθερές) αναμένεται το αντίστοιχο μέγεθος του δέντρου να είναι αρκετά μεγάλο. Αυτό έχει ως αποτέλεσμα οι γενετικοί τελεστές να δυσλειτουργούν, αφού ο μοναδικός τρόπος για να προκύψει στο κατάλληλο σημείο, κόμβος-σταθερά, είναι μέσω της διασταύρωσης ή της μετάλλαξης.

Προκειμένου να αποφευχθεί η ύπαρξη ξεχωριστών κόμβων στα συμβολικά δέντρα για τις σταθερές, προστίθεται ένα ακόμη πεδίο σε κάθε κόμβο, το οποίο φιλοξενεί την τιμή της σταθεράς με την οποία πολλαπλασιάζεται το αριθμητικό αποτέλεσμα της συμβολικής έκφρασης που κωδικοποιεί το αντίστοιχο υποδέντρο (βλ. Διάγραμμα 3.4).



Διάγραμμα 3.4: Τροποποιημένο δέντρο του οποίου οι κόμβοι ενθυλακώνουν την σταθερά.

Η έκφραση που κωδικοποιεί το παραπάνω δέντρο είναι η:

$$f\left([\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}]\right) = c_1 \cdot (c_2 \cdot \boldsymbol{y} + c_3 \cdot (c_4 \cdot \boldsymbol{x} + c_5 \cdot \boldsymbol{z})) \tag{3.3.1}$$

$$= c_1 c_2 \boldsymbol{y} + c_1 c_3 c_4 \boldsymbol{x} + c_1 c_3 c_5 \boldsymbol{z}. \tag{3.3.2}$$

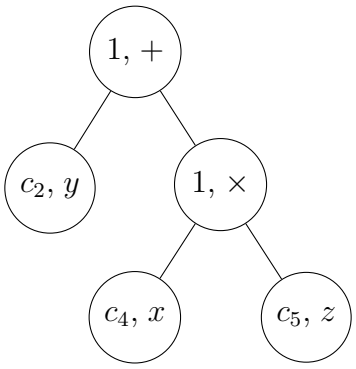
Παρατηρούμε ότι η έκφραση περιέχει πέντε σταθερές, ενώ στην πραγματικότητα χρειάζεται μόλις τρεις, μία σε κάθε φύλλο του δέντρου. Για παράδειγμα η ρίζα του δέντρου δεν χρειάζεται να έχει κάποια σταθερά, καθώς η συνάρτηση που προσθέτει δύο στοιχεία μεταξύ τους, είναι ομογενής πρώτης τάξης. Επίσημα, μια συνάρτηση g λέγεται ομογενής πρώτης τάξης όταν για κάθε $c \neq 0$:

$$c \cdot g\left(\boldsymbol{x}, \boldsymbol{y}\right) = g\left(c \cdot \boldsymbol{x}, c \cdot \boldsymbol{y}\right). \tag{3.3.3}$$

Επομένως κατά την εκτίμηση των συντελεστών του μοντέλου, το διάνυσμα $\boldsymbol{\theta}$ των παραμέτρων του αποτελείται από:

- Τις σταθερές των εσωτερικών κόμβων που περιλαμβάνουν μη ομογενή τελεστή.
- Τις σταθερές των φύλλων του.

Οι υπόλοιποι κόμβοι θεωρείται ότι έχουν σταθερά ίση με ένα. Σε κάθε περίπτωση η αριθμητική τιμή του εν λόγου δέντρου, σαν αυτό του Διαγράμματος 3.5, υπολογίζεται σύμφωνα με τον Αλγόριθμο 6.



Διάγραμμα 3.5: Επαναδιατύπωση σταθερών του δέντρου σύμφωνα με το αν οι κόμβοι περιέχουν ομογενής τελεστές.

3.3.2 Ελαχιστοποίηση των υπολοίπων με την μέθοδο Levenberg-Marquardt

Στο Κεφάλαιο 1.3.3 έγινε μία σύντομη εισαγωγή στα μη γραμμικά ελάχιστα τετράγωνα. Για το πρόβλημα της συμβολικής παλινδρόμησης, δεχόμαστε την υπόθεση ότι τα υπόλοιπα προέρχονται από την κανονική κατανομή με μέση τιμή μηδέν. Διαφορετικά, το μοντέλο που έχει εκτιμηθεί δεν είναι κατάλληλο για να περιγράψει την συναρτησιακή σχέση μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών.

Επομένως οι εν λόγω μέθοδοι του πρώτου κεφαλαίου θα μπορούσαν να εφαρμοστούν όπως περιγράφηκαν, για την εύρεση των παραμέτρων $\boldsymbol{\theta}^* \in \mathbb{R}^p$ που επιλύουν το πρόβλημα των ελαχίστων τετραγώνων 1.3.7. Όμως στα μοντέλα που απαρτίζουν τον πληθυσμό του γενετικού αλγορίθμου, η συνάρτηση των υπολοίπων 1.3.1 εν γένει μπορεί να μην είναι κυρτή

Αλγόριθμος 6 Υπολογισμός της αριθμητικής τιμής του τροποποιημένου δέντρου με σταθερές, μέσω PostOrder διαπέρασης.

```
function evaluate(p)
    c = constant of p
    if p is leaf node
        return value of c*p
    else
        if element e of p is binary operator
            left_value  = evaluate(left  Child)
            right_value = evaluate(right Child)
            return c*e(left_value,right_value)
        else
            right_value = evaluate(right Child)
            return c*e(right_value)
```

ή να περιέχει πολλές απότομες «κοιλάδες». Σε αυτές τις περιπτώσεις το ενδεχόμενο να βρεθεί κάποιο ελάχιστο είναι μικρό, και οφείλεται στο γεγονός ότι οι προσεγγίσεις της συνάρτησης των υπολοίπων, τόσο της πρώτης όσο και της δεύτερης τάξης, ισχύουν τοπικά γύρω από το αρχικό σημείο $\boldsymbol{\theta_0}$.

Η μέθοδος των Levenberg-Marquardt, αποτελεί μια στρατηγική γενίκευσης που καθοδηγεί την λύση του προβλήματος βελτιστοποίησης ακόμα και σε «δύσβατες» περιοχές μακριά από το αρχικό σημείο. Η κύρια ιδέα είναι σε κάθε επανάληψη του αλγορίθμου να επιλύεται το πρόβλημα:

$$\min_{d\in\mathbb{R}^p}\left\{f\left(\boldsymbol{\theta_\kappa}+\boldsymbol{d}\right)\right\} \tag{3.3.4}$$

υπό τον περιορισμό $\|\boldsymbol{d}\|_2 \leq \Delta_k$, όπου $\Delta_k > 0$ το μέγιστο επιτρεπτό βήμα και $f(\cdot)$ η συνάρτηση αθροίσματος των τετραγωνικών υπολοίπων 1.3.4.

Το πρόβλημα 3.3.4, στην μέθοδο Levenberg-Marquardt λύνεται επαναληπτικά, επιλύοντας το ακόλουθο αλγεβρικό σύστημα ως προς $\boldsymbol{d}$:

$$\left(J\left(\boldsymbol{\theta_k}\right)^T J\left(\boldsymbol{\theta_k}\right) - \lambda I_{p\times p}\right)\boldsymbol{d} = -J\left(\boldsymbol{\theta_k}\right)^T \boldsymbol{r}\left(\boldsymbol{\theta_k}\right). \tag{3.3.5}$$

Για μικρές τιμές της παραμέτρου λ, το πρόβλημα είναι το ίδιο με αυτό της προσέγγισης πρώτης τάξης των Gauss-Newton, ενώ για μεγάλες τιμές το επαναληπτικό σχήμα λαμβάνει την μορφή της μεθόδου καθόδου βάση της κλίσης (gradient descend).

Στην συνέχεια υπολογίζεται το πηλίκο της πραγματικής μείωσης της αντικειμενικής συνάρτησης f, έναντι της προσέγγισης αυτής με χρήση του αναπτύγματος Taylor πρώτης τάξης:

$$\rho\left(\boldsymbol{d}\right)=\frac{f\left(\boldsymbol{\theta_{\kappa}}\right)-f\left(\boldsymbol{\theta_{\kappa}}+\boldsymbol{d}\right)}{f\left(\boldsymbol{\theta_{\kappa}}\right)-\left(f\left(\boldsymbol{\theta_{\kappa}}\right)+\nabla f\left(\boldsymbol{\theta_{\kappa}}\right)^{\mathrm{T}}\boldsymbol{d}\right)}. \tag{3.3.6}$$

Για τις διάφορες τιμές που μπορεί να λάβει η 3.3.6, διακρίνονται οι εξής περιπτώσεις με βάση τις οποίες αλλάζει η περιοχή εμπιστοσύνης Δ_k:

- Αν το κλάσμα είναι κοντά στο ένα, τότε η τοπική προσέγγιση είναι σχεδόν το ίδιο καλή με την πραγματική τιμή της συνάρτησης. Αποδεχόμαστε την μεταβολή $\boldsymbol{d}$, και επειδή έχουμε ισχυρότερη πεποίθηση ότι η προσέγγιση είναι ρεαλιστική αυξάνουμε το μέγιστο επιτρεπτό βήμα Δ_k κατά έναν παράγοντα.

- Εάν το ρ είναι πολύ μικρό, τότε η προσέγγιση απέχει πολύ από την κανονική τιμή. Άρα δεν μπορούμε να εμπιστευτούμε το βήμα $\boldsymbol{d}$ και το απορρίπτουμε, ενώ μικραίνει το Δ_k κατά έναν παράγοντα.

- Τέλος, αν το κλάσμα δεν είναι ούτε πολύ μικρό, ούτε πολύ κοντά στο ένα, τότε το βήμα γίνεται αποδεκτό και το εύρος εμπιστοσύνης Δ_k παραμένει σταθερό.

Η επιλογή των ορίων βάση των οποίων διαχωρίζονται οι παραπάνω περιπτώσεις, καθώς και το μέγεθος της μεταβολής του Δ_k, δεν είναι προφανής ούτε καθολική για κάθε πρόβλημα. Επιπλέον, οι πολλοί και διαφορετικοί τρόποι που μπορεί κανείς να επιλέξει την σταθερά λ στην Εξίσωση 3.3.5, έχουν οδηγήσει στην δημιουργία πολλών παραλλαγών της εν λόγω μεθόδου. Στην βιβλιοθήκη scipy[2] της python, χρησιμοποιείται η υλοποίηση που προτάθηκε από τον Jorge J. More (Moré, 1978), η οποία αποτέλεσε την τελική επιλογή για την εκτίμηση των παραμέτρων των μοντέλων στον πληθυσμό του γενετικού αλγορίθμου.

3.4 Εφαρμογή

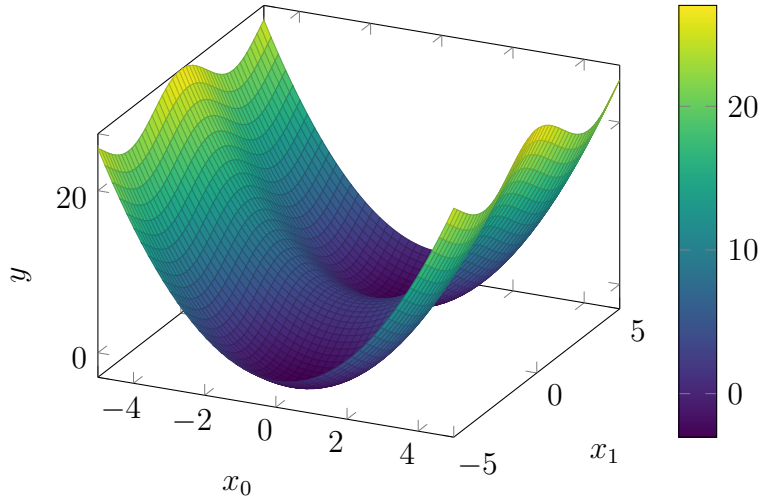
Σε αυτή την υποενότητα, θα λυθεί ένα «γνωστό» πρόβλημα παλινδρόμησης μέσω του γενετικού αλγορίθμου συμβολικής παλινδρόμησης. Τα δεδομένα θα αποτελούνται από δύο επεξηγηματικές μεταβλητές, $\boldsymbol{x_0}$, $\boldsymbol{x_1}$ και μία μεταβλητή εξάρτησης $\boldsymbol{y}$. Οι επεξηγηματικές μεταβλητές, ορίζεται να αποτελούνται από 1000 τυχαίες τιμές ομοιόμορφα επιλεγμένες από το διάστημα $[-3, 3]$ και $[0, 10]$ αντίστοιχα. Η συναρτησιακή εξάρτηση μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών είναι:

$$y=2.5382\cdot\cos\left(x_1\right)+x_0^2-0.5+\varepsilon \tag{3.4.1}$$

όπου $\varepsilon\sim N\left(0,0.2^2\right)$ το τυχαίο σφάλμα σε κάθε παρατηρήση. Στο Διάγραμμα 3.6, φαίνεται το μη στοχαστικό μέρος του μοντέλου.

Προκειμένου να προκύψει ο σταθερός όρος «-0.5» της Εξίσωσης 3.4.1, στις επεξηγηματικές μεταβλητές προστίθεται μία ακόμα εικονική μεταβλητή $\boldsymbol{x_2}=[1,\ldots,1]^T$.

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.least_squares.html



Διάγραμμα 3.6: Τρισδιάστατη επιφάνεια της καμπύλης $y = 2.5382 \cdot \cos(x_1) + x_0^2 - 0.5$.

Χρησιμοποιώντας την παραμετροποίηση του Πίνακα 1, το τελικό μοντέλο στο οποίο κατέληξε ο γενετικός αλγόριθμος είναι το ακόλουθο:

$$y = (0.93 \times x_0 \times 1.08 \times x_0 - 2.16 \times \sin(0.23 \times x_2) + 2.54 \times \cos(x_1)). \tag{3.4.2}$$

Συχνά χρειάζεται το τελικό μοντέλο να επαναδιατυπωθεί, με κάποιες απλοποιήσεις ή πράξεις μεταξύ των σταθερών του, προκειμένου να αποκτήσει μια πιο συμπτυγμένη μορφή:

$$y = 1.00 \times x_0 \times x_0 + 2.54 \times \cos(x_1) - 0.49 + x_0^2. \tag{3.4.3}$$

Παρατηρούμε ότι ο αλγόριθμος της συμβολικής παλινδρόμησης, κατάφερε να ανακτήσει από τα δεδομένα τον κλειστό τύπο 3.4.1 που περιγράφει την συναρτησιακή εξάρτηση μεταξύ των μεταβλητών. Συγκεκριμένα, αν και δεν είχε δοθεί ο τελεστής ύψωσης στο τετράγωνο, το τελικό μοντέλο κατασκεύασε την εν λόγω σχέση πολλαπλασιάζοντας την μεταβλητή x_0 με τον εαυτό της. Επιπροσθέτως, αν και δόθηκε προς διευκόλυνση η εικονική μεταβλητή x_2 προκειμένου ο γενετικός αλγόριθμος να κατασκευάσει τον σταθερό όρο της εξίσωσης 3.4.1, εν τέλει ο ζητούμενος όρος προέκυψε συνδιάζοντας τον τελεστή $sin(x)$.

Το μέσο τετραγωνικό σφάλμα μεταξύ της εκτίμησης του μοντέλου και της πραγματικής τιμής είναι 4.3×10^{-2}, ενώ η μέση τιμή των υπολοίπων είναι 2.1×10^{-7} και η αντίστοιχη τυπική απόκλιση 0.2. Επίσης, σύμφωνα με το Διάγραμμα 3.7 η υπόθεση ότι τα σφάλματα προέρχονται από κανονική κατανομή με μέση τιμή μηδέν δεν φαίνεται παράλογη.

Στις πρώτες επαναλήψεις όπου τα δέντρα είναι μικρά, παρατηρούμε ότι η προσαρμογή δεν είναι καλή. Μετά το πέρας ορισμένων επαναλήψεων, πραγματοποιούνται διασταυρώσεις μεταξύ των χρωμοσωμάτων που βελτιώνουν την λύση του γενετικού αλγορίθμου, και στην 10η επανάληψη ο αλγόριθμος τερματίζει.

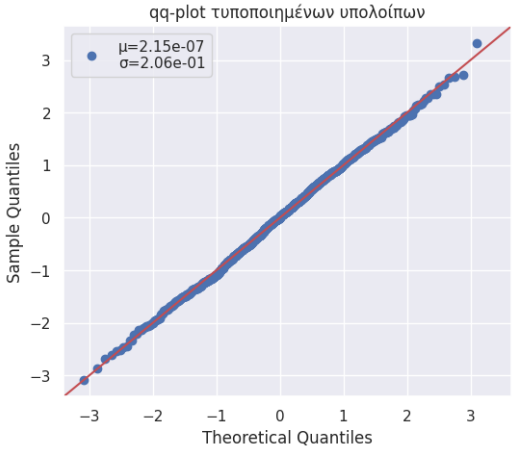
Παράμετρος	Σύμβολο	Τιμή	Σχόλιο
Μέγεθος Πληθυσμού	N_p	400	-
Πιθανότητα Διασταύρωσης	P_c	0.8	-
Πιθανότητα Μετάλλαξης	P_m	0.1	-
Πιθανότητα Επιλογής	P_S	0.9	Πιθανότητα επιλογής του καλύτερου στο τουρνουά, για τον καθορισμό των υποψηφίων προς διασταύρωση.
Κάτω όριο σημασιολογικής διαφοράς	δ_{lb}	0.01	Όριο κάτω από το οποίο τα υποδέντρα θεωρούνται ίδια και δεν διασταυρώνονται
Άνω όριο σημασιολογικής διαφοράς	δ_{ub}	0.5	Όριο πάνω από το οποίο τα υποδέντρα θεωρούνται εντελώς διαφορετικά και δεν διασταυρώνονται
Επιλογή υποψηφίων επόμενης επανάληψης	w	0.70	Age-Fitness Pareto
Λίστα τελεστών	$\mathcal{O}$	$\{+, \times, /, \cos, \sin\}$	-
Αντικειμενική συνάρτηση	f	$\frac{1}{n}\left\|y - \hat{y}\right\|_2^2$	Μέσο τετραγωνικό σφάλμα
Κριτήριο τερματισμού	-	$f < 0.045$	-

Πίνακας 1: Παραμετροποίηση του γενετικού αλγορίθμου.

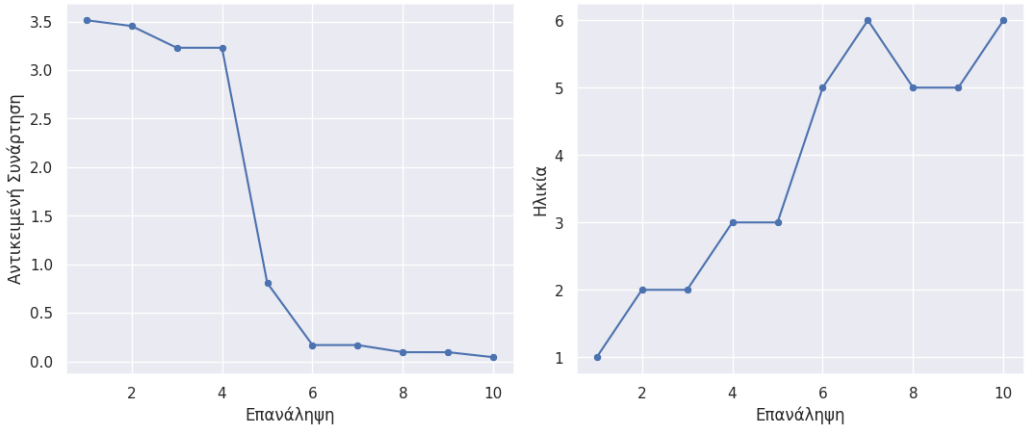
Με μία πρώτη ματιά, στο Διάγραμμα 3.8 δεν φαίνεται να υπάρχει ουσιαστική βελτίωση της λύσης στις πρώτες 4 επαναλήψεις. Παρόλα αυτά, η ηλικία του βέλτιστου χρωμοσώματος στην αρχή του αυξάνεται, το οποίο σημαίνει ότι το καλύτερο χρωμόσωμα στις εν λόγω επαναλήψεις αλλάζει κάθε φορά, ή συμμετέχει στους γενετικούς τελεστές της μετάλλαξης και διασταύρωσης, επομένως ο πληθυσμός εξελίσσεται και δεν μένει στάσιμος.

Το συμπέρασμα ότι ο πληθυσμός εξελίσσεται επιβεβαιώνει και το Διάγραμμα 3.9 της θερμοκρασίας. Πιο συγκεκριμένα, η θερμοκρασία (βλ. Εξίσωση 3.2.12) στην αρχή του γενετικού αλγορίθμου είναι μικρή, καθώς η αρχικοποίηση του πληθυσμού γίνεται τυχαία. Επομένως εμφανίζεται κάθε τελεστής στον αρχικό πληθυσμό και υπάρχει μεγάλη ποικιλομορφία. Καθώς προχωράει ο αλγόριθμος, η θερμοκρασία αυξάνεται, το οποίο σημαίνει ότι μειώνεται η ποικιλομορφία καθώς οι υποψήφιοι του πληθυσμού συγκλίνουν προς το βέλτιστο μοντέλο. Συγκεκριμένα, μετά την 4η επανάληψη όπου έγινε σημαντική βελτίωση στην τιμή της αντικειμενικής συνάρτησης, η κλίση της καμπύλης της θερμοκρασίας αυξάνεται, καθώς οι υποψήφιοι με τα ικανά γονίδια κυριαρχούν τον πληθυσμό.

Δημιουργώντας ένα πλέγμα $[-3, 3] \times [0, 10]$ μπορούμε στο Διάγραμμα 3.10 να δούμε γραφικά το μοντέλο εκτίμησης 3.4.3 σε σχέση με την πραγματική λύση 3.4.1.



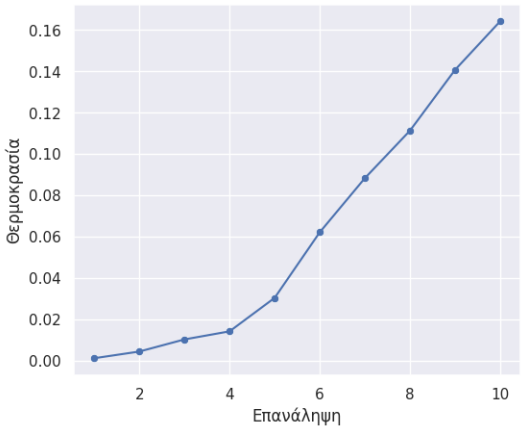
Διάγραμμα 3.7: Γραφική παράσταση των δειγματικών ποσοστημορίων ως προς τα θεωρητικά ποσοστημόρια της Κανονικής Κατανομής.



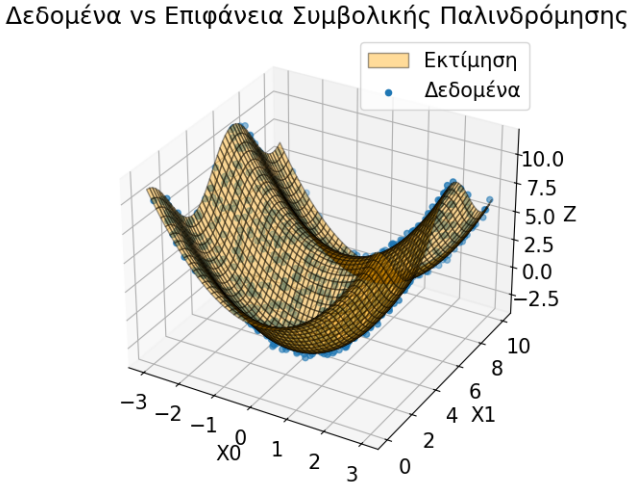
Διάγραμμα 3.8: Εξέλιξη της βέλτιστης λύσης με το πέρας των επαναλήψεων: Αριστερά η τιμή της αντικειμενικής συνάρτησης, δεξιά η ηλικία.

3.5 Επίλογος

Στόχος του κεφαλαίου ήταν η αναλυτική παρουσίαση του γενετικού αλγορίθμου για το πρόβλημα συμβολικής παλινδρόμησης. Κάθε χαρακτηριστικό του γενετικού αλγορίθμου απαιτεί κατάλληλες τροποποιήσεις προκειμένου να προσαρμοστεί στις ιδιομορφίες των χρωμοσωμάτων που είναι τα συμβολικά δέντρα. Επίσης οι διάφορες εκδοχές των γενετικών τελεστών, στοχεύουν να βελτιώσουν την ικανότητα αναζήτησης του γενετικού αλγορίθμου στον σύνθετο χώρο των συμβολικών εκφράσεων. Για κάθε γενετικό τελεστή, στην αντίστοιχη υποενότητα αναφέρεται η τελική εκδοχή του, η οποία υλοποιήθηκε στον τελικό αλγόριθμο. Ο γενετικός αλγόριθμος που προκύπτει, καλείται στην τελευταία υποενότητα να αποδείξει ότι πράγματι έχει την ικανότητα να λύσει το πρόβλημα συμβολικής παλινδρόμησης. Τα αποτελέσματα είναι ικανοποιητικά, αφού το αρχικό μοντέλο ανακτήθηκε από τα δεδομένα και με σχεδόν μηδενική απόκλιση. Επομένως, μένει να δοκιμαστεί η ικανότητα αυτής της



Διάγραμμα 3.9: Μεταβολή της θερμοκρασίας.



Διάγραμμα 3.10: Επιφάνεια 3-διαστάσεων ανάμεσα στο μοντέλο που εκτιμήθηκε και τα αρχικά δεδομένα.

μεθόδου συμβολικής παλινδρόμησης, στο πραγματικό πρόβλημα του επόμενου κεφαλαίου.

4 Προβολή και Ανάλυση Δεδομένων - Σύγκριση Μεθόδων

4.1 Πρόλογος

Στο παρόν κεφάλαιο, ο αλγόριθμος συμβολικής παλινδρόμησης του προηγούμενου κεφαλαίου εφαρμόζεται σε ένα πρόβλημα δυαδικής ταξινόμησης. Το σύνολο δεδομένων που χρησιμοποιήθηκε, αποτελεί προϊόν προσομοίωσης με την μέθοδο των Μαρκοβιανών αλυσίδων, για την καταγραφή σωματιδίων υψηλής ενέργειας από ένα τηλεσκόπιο (Bock, 2007). Σκοπός των παραπάνω δεδομένων είναι η ταξινόμηση των παρατηρήσεων ως «ακτίνα γάμμα» ή «αδρόνιο», και θα γίνει με διάφορους ταξινομητές, τα αποτελέσματα των οποίων θα συγκριθούν. Η ανάπτυξη ενός αλγόριθμου ταξινόμησης για το παραπάνω πρόβλημα έχει μεγάλη σημασία, καθώς οι ακτίνες γάμμα φέρουν σημαντικές πληροφορίες σχετικά με κοσμικά γεγονότα (ESA, 2018).

4.2 Ανάλυση δεδομένων

Τα δεδομένα αποτελούνται από 11 μεταβλητές, εκ των οποίων οι 10 έχουν επεξηγηματικό ρόλο, και η τελευταία είναι η μεταβλητή απόκρισης.

- Μεταβλητή απόκρισης:
  - Δυαδική «output»: Περιγράφει αν το σήμα που κατέγραψε το τηλεσκόπιο αποτελεί ακτίνα γάμμα ή αδρόνιο, κωδικοποιώντας με 1 ή 0 τα δύο γεγονότα αντίστοιχα.
- Επεξηγηματικές μεταβλητές (όλες συνεχείς):
  - «fLength»: Το μήκος του κύριου άξονα της έλλειψης (σε χιλιοστά), που σχηματίζεται από το σωματίδιο που ανιχνεύθηκε. Ως κύριος άξονας σε μία έλλειψη ορίζεται να είναι η μεγαλύτερη διάμετρος.
  - «fWidth»: Το μήκος του δευτερεύοντα άξονα της έλλειψης (σε χιλιοστά) που σχηματίζεται από το σωματίδιο. Η μικρότερη διάμετρος σε μια έλλειψη αποτελεί τον δευτερεύοντα άξονα.
  - «fSize»: Αντιπροσωπεύει το λογαριθμισμένο, με βάση 10, άθροισμα του περιεχομένου των εικονοστοιχείων που απαρτίζουν την εικόνα του τηλεσκοπίου. Η εν λόγω μεταβλητή υποδεικνύει την ένταση του ληφθέντος σήματος.
  - «fConc»: Ο λόγος ανάμεσα στο άθροισμα μεταξύ των δύο μεγαλύτερων εικονοστοιχείων, και της μεταβλητής fsize. Παρέχει πληροφορίες σχετικά με τον τρόπο κατανομής του περιεχομένου της εικόνας.

– «fConc1»: Ο λόγος ανάμεσα το μεγαλύτερο εικονοστοιχείο και την μεταβλητή fsize. Η μεταβλητή αυτή κωδικοποιεί παρόμοια πληροφορία με την fConc.

– «fAsym»: Η απόσταση (σε χιλιοστά) μεταξύ του μεγαλύτερου εικονοστοιχείου της εικόνας και την προβολή του κέντρου της έλλειψης στον κύριο άξονα. Δηλαδή αντιπροσωπεύει την μετατόπιση του σήματος σε σχέση με το κέντρο της έλλειψης.

– «fM3Long»: Η κυβική ρίζα της ροπής τρίτης τάξεως κατά μήκος του κύριου άξονα της έλλειψης. Η εν λόγω μεταβλητή παρέχει πληροφορία σχετικά το σχήμα της κατανομής των εικονοστοιχείων κατά μήκος του κύριου άξονα.

– «fM3Trans»: Η κυβική ρίζα της ροπής τρίτης τάξεως κατά μήκος του δευτερεύοντα άξονα της έλλειψης. Όμοια με την fM3Long, η μεταβλητή αυτή παρέχει πληροφορίες για το σχήμα της κατανομής των εικονοστοιχείων κατά μήκος του δευτερεύοντα άξονα.

– «fAlpha»: Η γωνία (σε μοίρες) που σχηματίζεται μεταξύ του κύριου άξονα της έλλειψης και ενός διανύσματος που ενώνει την αρχή των αξόνων της εικόνας με το κέντρο της έλλειψης. Με αυτό τον τρόπο υπολογίζεται ο προσανατολισμός του σήματος.

– «fDist»: Η απόσταση (σε χιλιοστά) μεταξύ του κέντρου της εικόνας και του κέντρου της έλλειψης.

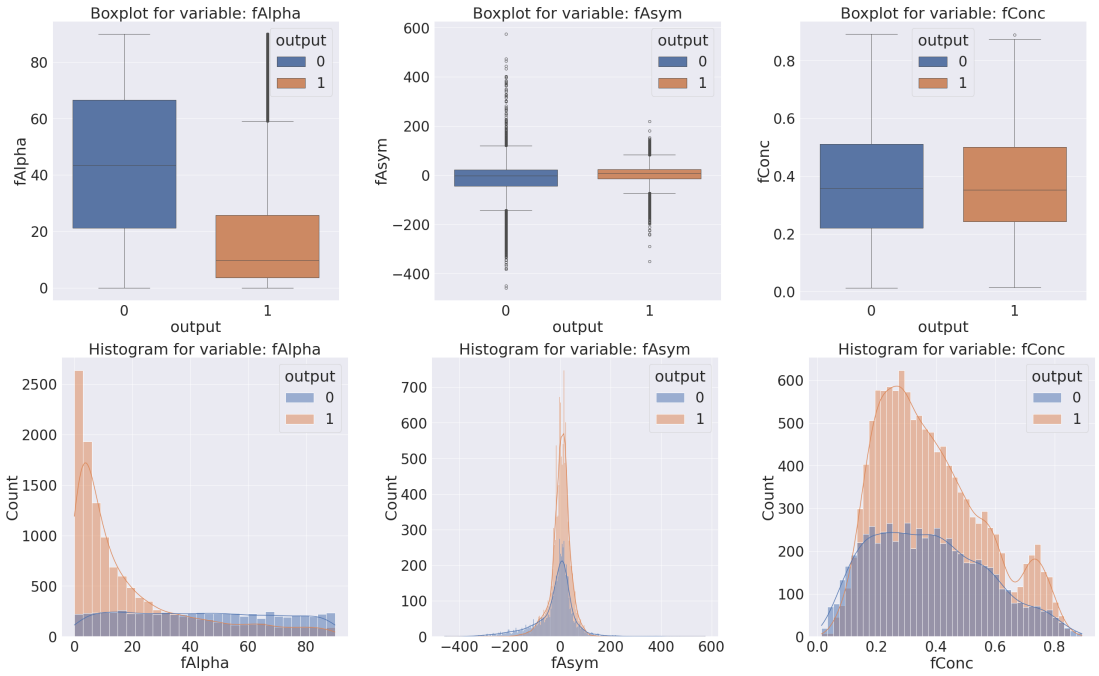
Κλείνοντας, σημειώνεται ότι στα δεδομένα δεν υπάρχουν κενές εγγραφές, ούτε διπλότυπες παρατηρήσεις.

4.2.1 Διαγράμματα

Αρχικά, για κάθε μεταβλητή σχεδιάζεται το αντίστοιχο θηκοδιάγραμμα και ιστόγραμμα, για κάθε κλάση της μεταβλητής απόκρισης αντίστοιχα (βλ. Διαγράμματα 4.1, 4.2, 4.3 & 4.4).

- fAlpha: Η διάμεσος των δεδομένων του δεύτερου δείγματος (κλάση 1), είναι μικρότερη από αυτή του πρώτου δείγματος (κλάση 0). Επιπλέον παρατηρούμε ότι η μεταβλητότητα, υπολογισμένη με βάση το ενδοτεταρτημοριακό εύρος, είναι διαφορετική στα δύο δείγματα και στην κλάση 1 υπάρχουν αρκετές έκτροπες τιμές. Τέλος, από το ιστόγραμμα φαίνεται ότι η κατανομή που ακολουθεί ο κάθε πληθυσμός είναι εντελώς διαφορετική.

- fAsym: Και τα δύο δείγματα περιέχουν αρκετές έκτροπες τιμές. Η μεταβλητότητα των παρατηρήσεων της κλάσης 0 είναι μεγαλύτερη, ενώ η αντίστοιχη διάμεσος είναι μικρότερη σε σχέση με την κλάση 1. Το ιστόγραμμα συχνοτήτων φανερώνει ότι η κατανομή που ακολουθεί το κάθε δείγμα είναι παρόμοια. Οι εν λόγω κατανομές είναι σχετικά συμμετρικές, αν και η κλάση 0 έχει μεγαλύτερη αριστερή ουρά.



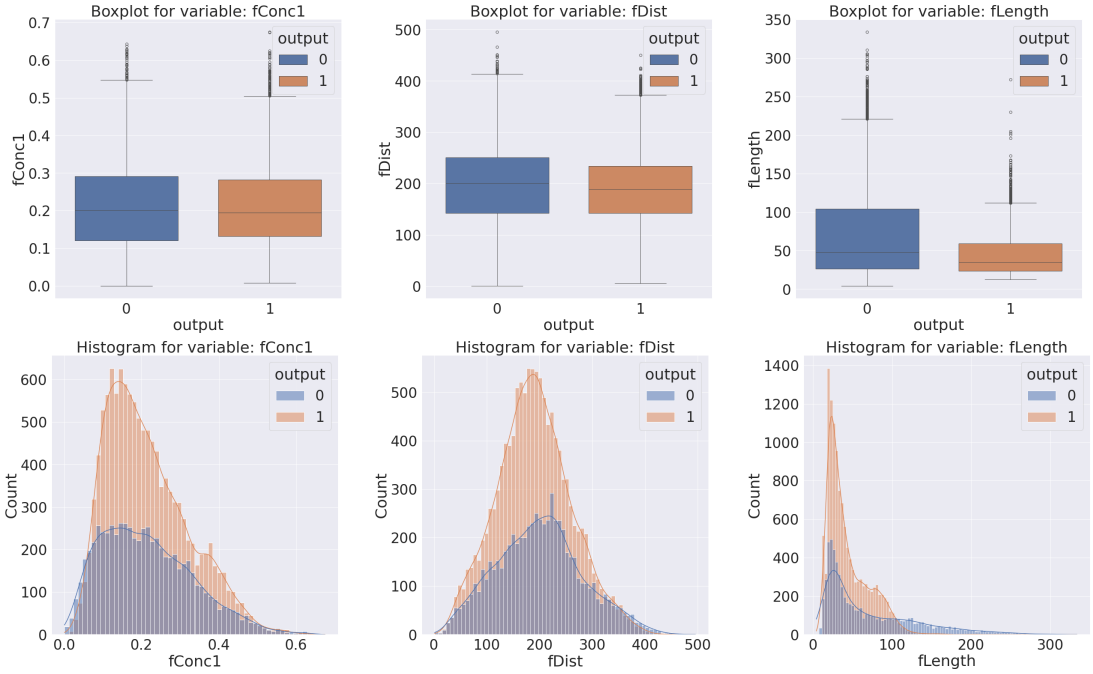
Διάγραμμα 4.1: Θηκοδιαγράμματα και Ιστογράμματα μεταβλητών fAlpha, fAsym & fConc.

- fConc: Η διάμεσος στους δύο πληθυσμούς είναι παρόμοια. Επίσης η μεταβλητότητα της κλάσης 1, υπολογισμένη με βάση το ενδοτεταρτημοριακό εύρος, είναι ελαφρώς μικρότερη σε σχέση με την κλάση 0. Κατ επέκταση, εξίσου όμοια με το θηκοδιάγραμμα είναι οι κατανομές στο ιστόγραμμα συχνοτήτων, όπου η λοξότητα των κατανομών είναι θετική.

- fConc1: Στο αντίστοιχο διάγραμμα παρατηρούμε ότι οι δύο πληθυσμοί έχουν παρόμοια διάμεσο. Επιπροσθέτως η μεταβλητότητα της κλάσης 1 είναι μικρότερη από αυτήν της κλάσης 0. Τέλος, και τα δύο δείγματα περιέχουν αρκετές έκτροπες τιμές. Για το ιστόγραμμα ισχύουν οι ίδιες παρατηρήσεις με αυτές της μεταβλητής fConc.

- fDist: Στο θηκοδιάγραμμα της εν λόγω μεταβλητής παρατηρούμε ότι η διάμεσος της κλάσης 1 είναι μικρότερη από αυτή της κλάσης 0. Εν συνεχεία, η μεταβλητότητα, υπολογισμένη με βάση το ενδοτεταρτημοριακό εύρος, είναι μικρότερη στην κλάση 1, ενώ και στα δύο δείγματα υπάρχουν αρκετές έκτροπες τιμές. Η κλάση 1 ακολουθεί μια πιο συμμετρική κατανομή, σε σχέση με την κλάση 0 η οποία έχει λίγο μεγαλύτερη ουρά στα δεξιά.

- fLength: Η διάμεσος της του δείγματος που ανήκει στην κλάση 0 είναι μεγαλύτερη σε σχέση με την κλάση 1. Επίσης μεγαλύτερη είναι η αντίστοιχη μεταβλητότητα των παρατηρήσεων που ανήκουν στην εν λόγω κλάση. Τέλος, οι δύο πληθυσμοί περιέχουν αρκετές έκτροπες παρατηρήσεις. Στο ιστόγραμμα συχνοτήτων επίσης διαφαίνεται η διαφορά στην μεταβλητότητα, και η έντονη λοξότητα στις αντίστοιχες κατανομές.



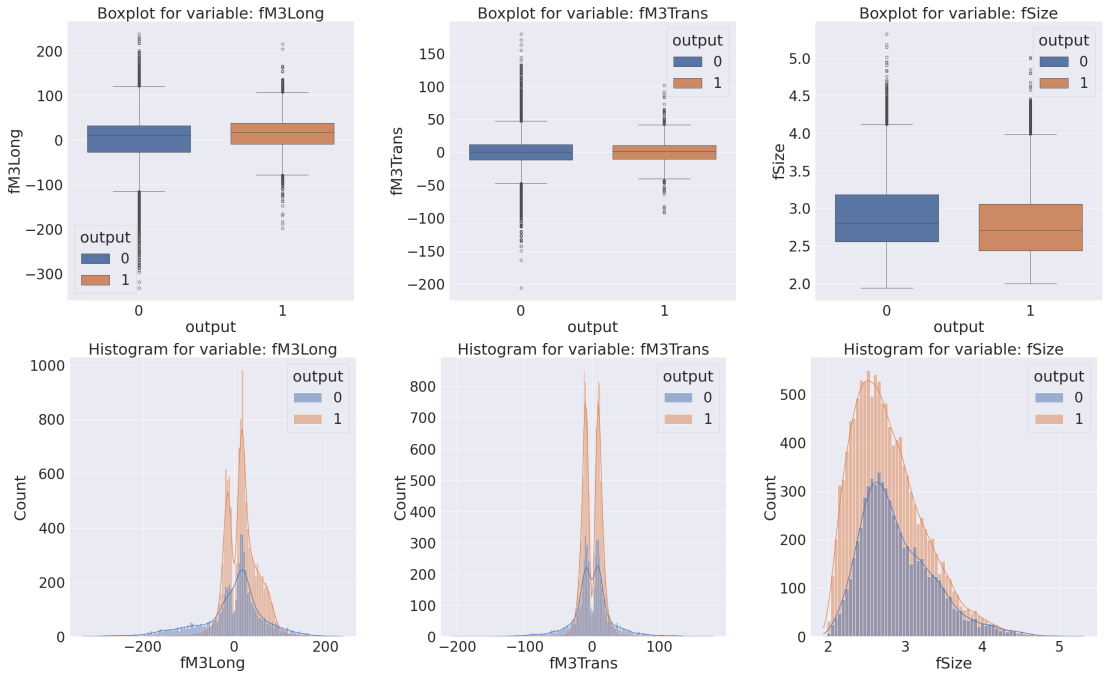
Διάγραμμα 4.2: Θηκοδιαγράμματα και Ιστογράμματα μεταβλητών fConc1, fDist & fLength.

- fM3Long: Παρατηρούνται έκτροπες τιμές και στα δύο δείγματα. Ο πληθυσμός της δεύτερης κλάσης έχει μικρότερη μεταβλητότητα σε σχέση με την πρώτη, ενώ η αντίστοιχη διάμεσος είναι μεγαλύτερη. Τέλος, παρατηρούμε στο αντίστοιχο ιστόγραμμα ότι η κατανομή της κλάσης 0 έχει μεγαλύτερη αριστερή ουρά, δηλαδή η λοξότητα είναι αρνητική.

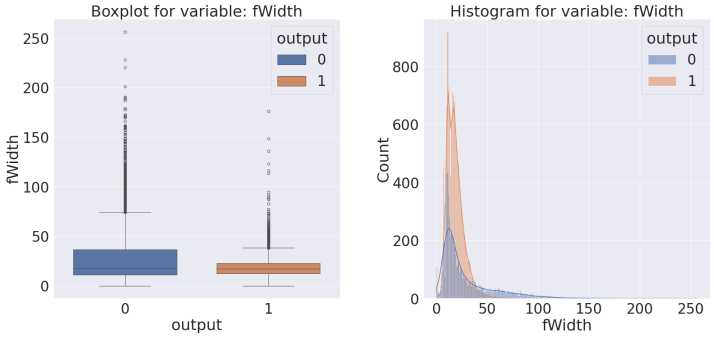
- fM3Trans: Στο αντίστοιχο διάγραμμα παρατηρούμε ότι οι δύο πληθυσμοί έχουν παρόμοια διάμεσο. Επιπροσθέτως η μεταβλητότητα της κλάσης 1 είναι μικρότερη από αυτήν της κλάσης 0. Τέλος, και τα δύο δείγματα περιέχουν αρκετές έκτροπες τιμές. Η κατανομή των δύο πληθυσμών είναι παρόμοια και σχετικά συμμετρική, αν και η κλάση 0 έχει μεγαλύτερες ουρές.

- fSize: Η διάμεσος των δεδομένων του δεύτερου δείγματος (κλάση 1), είναι μικρότερη από αυτή του πρώτου δείγματος (κλάση 0). Επιπλέον παρατηρούμε ότι η μεταβλητότητα, υπολογισμένη με βάση το ενδοτεταρτημοριακό εύρος, είναι διαφορετική στα δύο δείγματα και στην κλάση 0 υπάρχουν αρκετές έκτροπες τιμές. Κλείνοντας, παρατηρούμε στο αντίστοιχο ιστόγραμμα συχνοτήτων ότι οι παρατηρήσεις κατανέμονται με παρόμοιο τρόπο στα δεδομένα.

- fWidth: Και τα δύο δείγματα περιέχουν αρκετές έκτροπες τιμές. Η μεταβλητότητα των παρατηρήσεων της κλάσης 0 είναι μεγαλύτερη, ενώ η αντίστοιχη διάμεσος είναι παρόμοια. Οι παρατηρήσεις των αντίστοιχων κλάσεων κατανέμονται με μεγάλη λοξότητα, καθώς υπάρχει μεγάλη δεξιά ουρά.

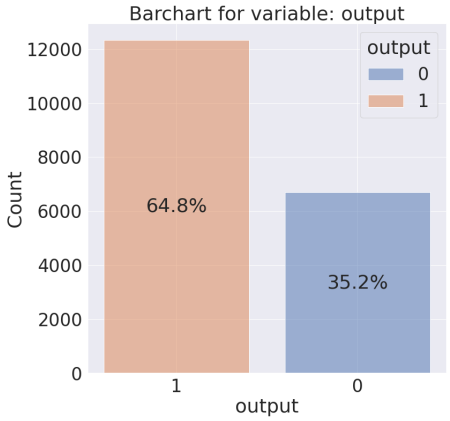


Διάγραμμα 4.3: Θηκοδιαγράμματα και Ιστογράμματα μεταβλητών fM3Long, fM3Trans & fSize.



Διάγραμμα 4.4: Θηκοδιάγραμμα και Ιστόγραμμα για την μεταβλητή fWidth.

Τέλος, στο Διάγραμμα 4.5 παρουσιάζεται το ραβδόγραμμα συχνοτήτων της μεταβλητής απόκρισης output. Το 65% των παρατηρήσεων ανήκει στην κλάση 1, δηλαδή το σήμα προέρχεται από ακτίνες γ.



Διάγραμμα 4.5: Ραβδόγραμμα σχετικών συχνοτήτων για την μεταβλητή απόκρισης.

4.2.2 Συσχετίσεις

Συχνά τα δεδομένα ελέγχονται για συσχετίσεις (correlation) πριν την διαδικασία της μάθησης. Οι συσχετίσεις εξετάζονται μεταξύ:

- Της μεταβλητής απόκρισης και των υπόλοιπων επεξηγηματικών μεταβλητών.
- Των επεξηγηματικών μεταβλητών.

Το επιθυμητό είναι η μεταβλητή απόκρισης να συσχετίζεται με κάποιες από τις επεξηγηματικές μεταβλητές. Σε αυτή την περίπτωση υπάρχει πρώιμη ένδειξη ότι πράγματι τα δεδομένα περιέχουν πληροφορία που μπορεί να αξιοποιηθεί προκειμένου να γίνουν προβλέψεις για την μεταβλητή απόκρισης.

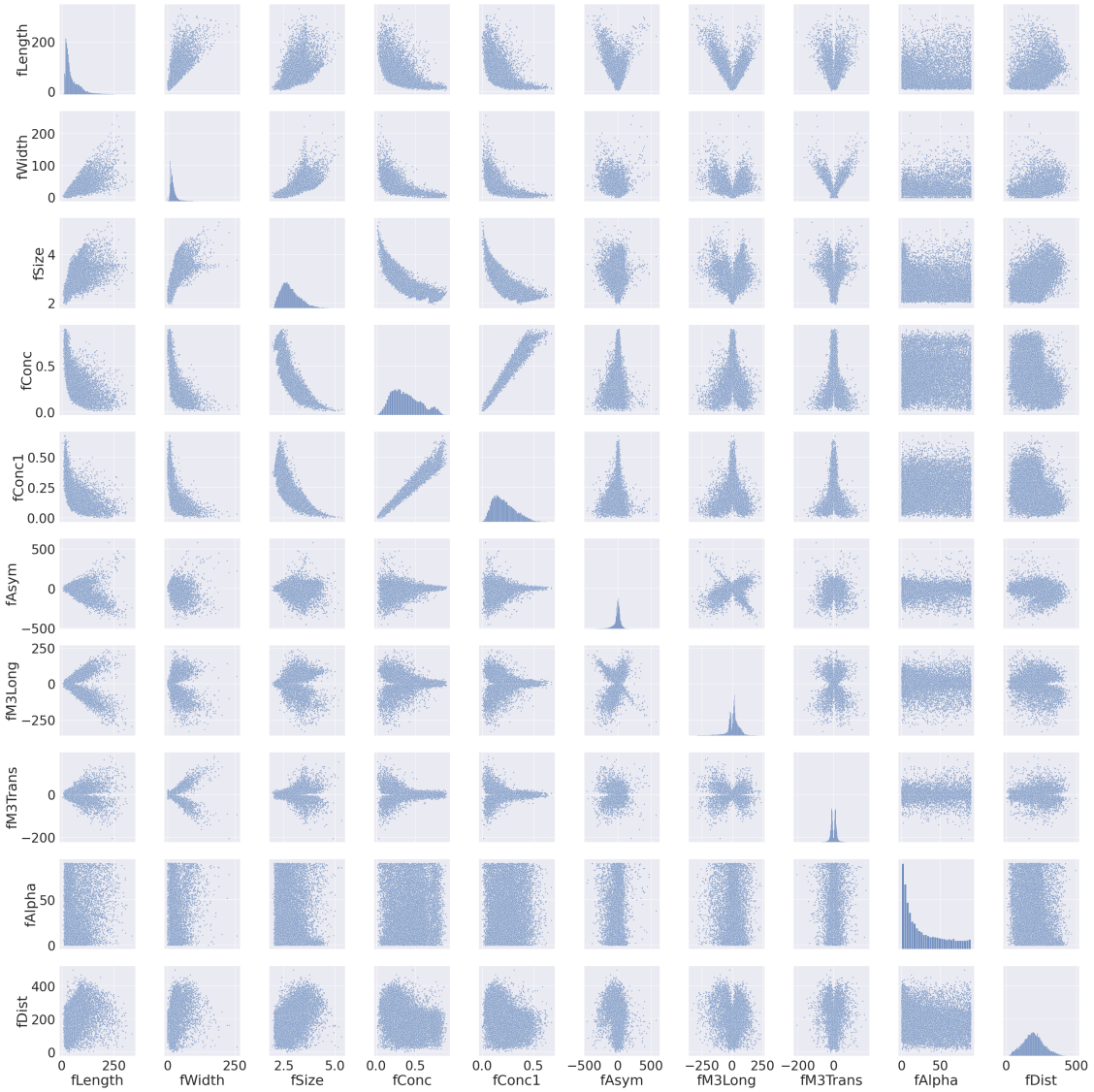
Αντίθετα, οι επεξηγηματικές μεταβλητές δεν πρέπει να συσχετίζονται ισχυρά μεταξύ τους. Σε περίπτωση που συσχετίζονται, τότε υπάρχει το ενδεχόμενο κάποια από αυτές να είναι περιττή καθώς επεξηγείται από κάποια άλλη.

Οι συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών μπορούν να γίνουν αντιληπτές κατασκευάζοντας τα αντίστοιχα διαγράμματα διασποράς ανά δύο μεταβλητές.

Στο Διάγραμμα 4.6, παρατηρούμε ότι οι παρατηρήσεις δεν απλώνονται με τυχαίο τρόπο στα αντίστοιχα διαγράμματα διασποράς των μεταβλητών:

- fLength με τις fConc, fConc1, fSize & fWidth.
- fWidth με τις fConc, fConc1 & fSize.
- fSize με τις fDist, fConc & fConc1.
- fConc με την fConc1.

Η πιο απλή μέθοδος υπολογισμού των συσχετίσεων είναι με τον συντελεστή συσχέτισης του Pearson (Φουσκάκης, 2013), όπου ελέγχεται η γραμμική συσχέτιση ανάμεσα σε δύο μεταβλητές, έστω **x**, **y**:



Διάγραμμα 4.6: Διαγράμματα διασποράς ανά ζεύγος μεταβλητών.

$$r_p = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}. \tag{4.2.1}$$

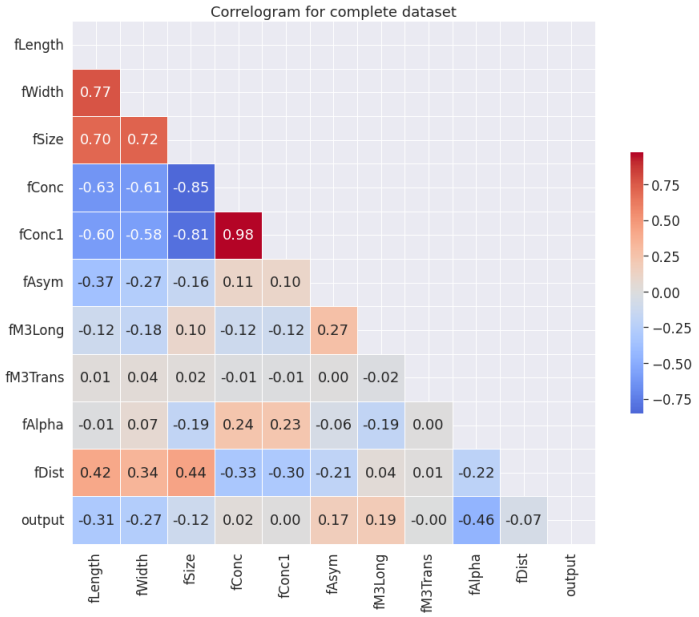
Ο συντελεστής συσχέτισης 4.2.1 λαμβάνει τιμές στο διάστημα $[-1, 1]$. Όταν $r_p \to 1$ τότε οι αντίστοιχες παρατηρήσεις των μεταβλητών $\mathbf{x}$, $\mathbf{y}$ είναι θετικά συσχετισμένες, ενώ όταν $r_p \to -1$ είναι αρνητικά συσχετισμένες. Τέλος, τιμές του συντελεστή συσχέτισης του Pearson κοντά στο 0, δείχνουν ότι δεν υπάρχει γραμμική συσχέτιση ανάμεσα στα δύο μεγέθη.

Όταν οι παρατηρήσεις ακολουθούν κανονική κατανομή, μπορεί κανείς να πραγματοποιήσει τον αμφίπλευρο έλεγχο με μηδενική υπόθεση H_0: $r_p = 0$ έναντι της εναλλακτική H_1: $r_p \neq 0$

σε επίπεδο σημαντικότητας α, με την βοήθεια της ελεγχοσυνάρτησης:

$$T_p = r_p \sqrt{\frac{n-2}{1-r_p^2}} \sim St\,(n-2)\,. \tag{4.2.2}$$

Εάν $|T_p| \geq t_{n-2,a/2}$ τότε δεν υπάρχουν ισχυρές ενδείξεις έναντι της μηδενικής υπόθεσης και δεν την απορρίπτουμε.



Διάγραμμα 4.7: Συντελεστής συσχέτισης του Pearson για τις μεταβλητές του συνόλου δεδομένων MagicTelescope.

Στην περίπτωση που τα δεδομένα δεν προέρχονται από την κανονική κατανομή, δεν μπορεί να χρησιμοποιηθεί η ελεγχοσυνάρτηση 4.2.2. Τότε, αντί του συντελεστή συσχέτισης του Pearson, χρησιμοποιείται εναλλακτικά ο βαθμολογικός δείκτης συσχέτισης του Spearman (Κοκολάκης & Φουσκάκης, 2009):

$$r_s = \frac{\sum_{i=1}^{n} \left(r_{x,i} - \overline{r_x}\right)\left(r_{y,i} - \overline{r_y}\right)}{\sqrt{\sum_{i=1}^{n} \left(r_{x,i} - \overline{r_x}\right)^2 \sum_{i=1}^{n} \left(r_{y,i} - \overline{r_y}\right)^2}} \tag{4.2.3}$$

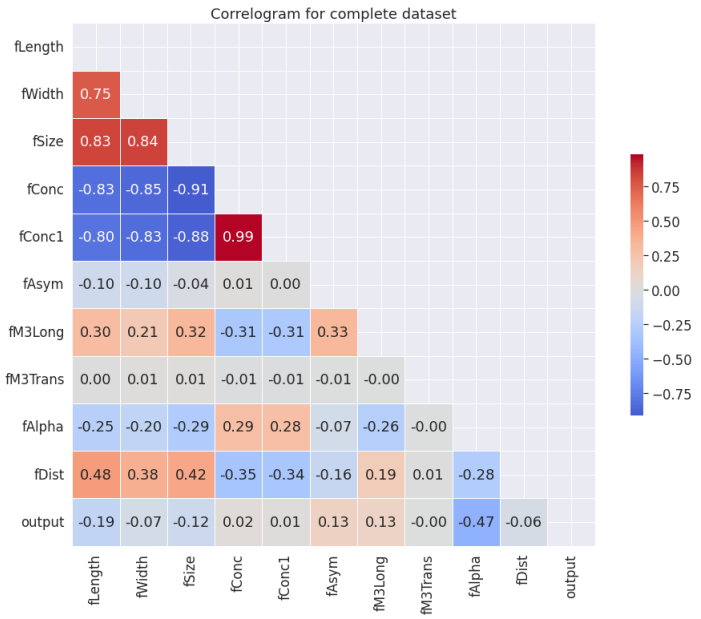
όπου $r_{\cdot,i}$ η τάξη της παρατήρησης i στο x και y αντίστοιχα.

Ο δείκτης συσχέτισης του Spearman 4.2.3, είναι επίσης φραγμένος στο $[-1, 1]$. Σε αντίθεση με αυτόν του Pearson, αντί για την γραμμική εξάρτηση ο εν λόγω συντελεστής δείχνει κατά πόσο μία μεταβλητή μπορεί να επεξηγηθεί ως μονότονη συνάρτηση της άλλης. Θετικές τιμές δείχνουν αύξουσα μονότονη συσχέτιση ανάμεσα στις δύο μεταβλητές, ενώ οι αρνητικές τιμές υποδεικνύουν φθίνουσα μονότονη συσχέτιση.

Η αντίστοιχη τιμή της ελεγχοσυνάρτησης για τον έλεγχο με μηδενική υπόθεση H_0: $r_s = 0$ έναντι της εναλλακτική H_1: $r_s \neq 0$ σε επίπεδο σημαντικότητας α, υπολογίζεται ως:

$$T_s = r_s \sqrt{\frac{n-2}{1-r_s^2}} \sim St\,(n-2)\,. \tag{4.2.4}$$

Η μηδενική υπόθεση δεν απορρίπτεται, όταν για την αντίστοιχη p-τιμή ισχύει ότι $|T_s| \geq t_{n-2,a/2}$.



Διάγραμμα 4.8: Συντελεστής συσχέτισης του Spearman για τις μεταβλητές του συνόλου δεδομένων MagicTelescope.

Συγκρίνοντας τα Διαγράμματα 4.7 & 4.8 για τους αντίστοιχους δείκτες συσχέτισης κατά Pearson & Spearman, παρατηρούμε ότι ο βαθμολογικός δείκτης συσχέτισης του Spearman είναι πιο συντηρητικός και αποδίδει μεγαλύτερη τιμή συσχέτισης ανάμεσα στις μεταβλητές. Πιο συγκεκριμένα, ο δείκτης συσχέτισης λαμβάνει μεγαλύτερες τιμές ανάμεσα στις μεταβλητές fLength, fConc, fConc1, fSize & fWidth. Τα παραπάνω αποτελέσματα έρχονται σε συμφωνία με τις παρατηρήσεις που έγιναν γραφικά από τα διαγράμματα διασποράς 4.6.

Σε αυτό το σημείο πρέπει να τονιστεί ότι οι παραπάνω έλεγχοι των συσχετίσεων πραγματοποιούνται ανάμεσα σε δύο μεταβλητές μόνο. Το τελικό μοντέλο όμως που αποσκοπούμε να φτιάξουμε θα απαρτίζεται από παραπάνω επεξηγηματικές μεταβλητές. Επομένως είναι σκόπιμο να θέλει κανείς να ελέγξει αν υπάρχουν συσχετίσεις κάποιας επεξηγηματικής μεταβλητής με συνδυασμούς των υπόλοιπων. Ένας τέτοιος δείκτης για τον εντοπισμό της πολυσυγγραμμικότητας, είναι ο παράγοντας μεγέθυνσης διασποράς (Variance Inflation Factor, VIF). Για τον υπολογισμό του εν λόγω συντελεστή προσαρμόζεται το πολλαπλό γραμμικό μοντέλο με μεταβλητή απόκρισης την $\boldsymbol{x_j}$, και επεξηγηματικές όλες τις υπόλοιπες $\boldsymbol{x_i}$, με $i \neq j$:

$$\boldsymbol{x_j} = \beta_0 + \beta_1 \boldsymbol{x_1} + \ldots + \beta_{j-1} \boldsymbol{x_{j-1}} + \beta_{j+1} \boldsymbol{x_{j+1}} + \ldots + \beta_n \boldsymbol{x_n}. \tag{4.2.5}$$

Τότε η τιμή του δείκτη VIF υπολογίζεται ως (Καρώνη & Οικονόμου, 2020):

$$VIF = \frac{1}{1 - R_j^2} \tag{4.2.6}$$

όπου R_j^2 ο συντελεστής προσδιορισμού στο αντίστοιχο πολλαπλό γραμμικό μοντέλο 4.2.5. Εάν οι επεξηγηματικές μεταβλητές είναι μόνο 2, τότε ο συντελεστής προσδιορισμού ταυτίζεται με τον δείκτη γραμμικής συσχέτισης του Pearson 4.2.1. Ο παράγοντας μεγέθυνσης διασποράς 4.2.6 δείχνει κατά πόσο αυξάνεται η διασπορά του εκτιμητή $\hat{\beta}_j$ όταν υπάρχουν προβλήματα πολυσυγγραμμικότητας μεταξύ των επεξηγηματικών μεταβλητών στο μοντέλο:

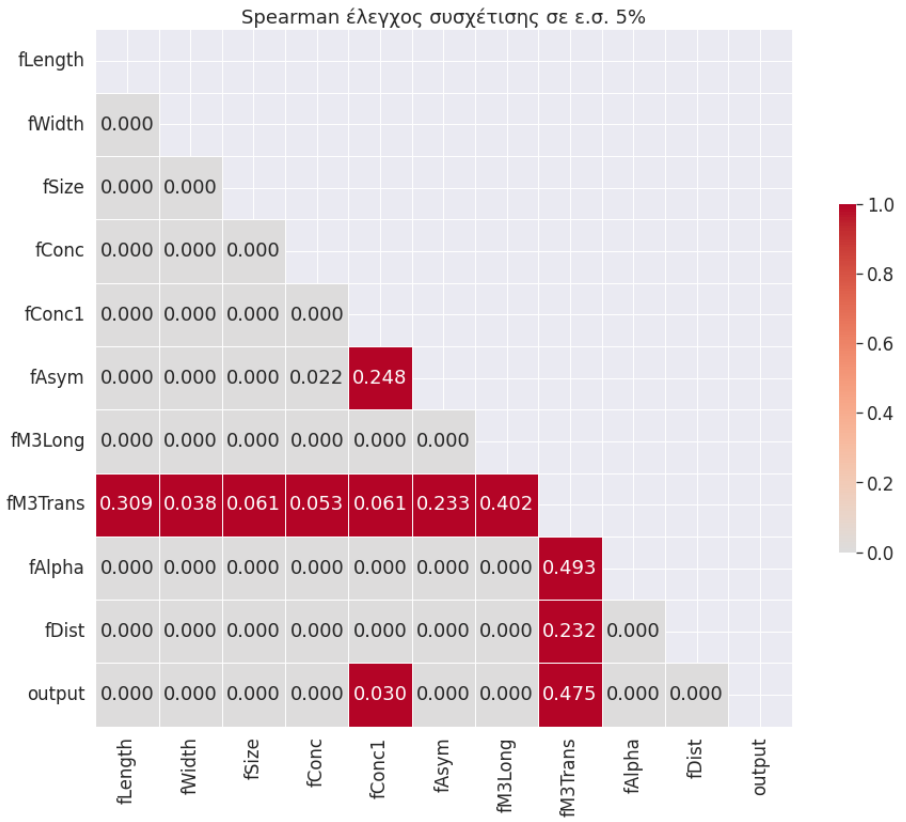
$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \ldots + \beta_j \boldsymbol{x_j} + \ldots + \beta_n \boldsymbol{x_n}. \tag{4.2.7}$$

Τονίζεται ότι η ύπαρξη πολυσυγγραμμικότητας ή συσχέτισης μεταξύ κάποιων μεταβλητών δεν επηρεάζει απαραίτητα την προβλεπτική ικανότητα του μοντέλου ούτε την προσαρμογή του μοντέλου στην μεταβλητή απόκρισης.

Μεταβλητή	VIF
fLength	8.27
fWidth	7.57
fSize	22.47
fConc	114.55
fConc1	102.04
fAsym	1.27
fM3Long	1.29
fM3Trans	1.00
fAlpha	2.56
fDist	10.57

Πίνακας 2: Τιμή του παράγοντα μεγέθυνσης διασποράς για τα δεδομένα του προβλήματος.

Αν και δεν υπάρχει απόλυτο όριο της τιμής του VIF, συνήθως εφιστάται η προσοχή σε μεταβλητές με τιμή μεγαλύτερη του 5. Ο Πίνακας 2, περιέχει τις τιμές του παράγοντα μεγέθυνσης διασποράς για τις επεξηγηματικές μεταβλητές του μοντέλου που αποσκοπούμε να κατασκευάσουμε. Οι μεταβλητές fConc, fConc1 & fSize έχουν πολύ μεγαλύτερη τιμή σε σχέση με τις υπόλοιπες. Αυτό είναι αναμενόμενο, αφού εκ των προτέρων γνωρίζαμε από την αρχή του κεφαλαίου 4.2, ότι οι εν λόγω μεταβλητές κωδικοποιούν πληροφορία που είναι γραμμική συνάρτηση η μία της άλλης.



Διάγραμμα 4.9: P-τιμές για τον έλεγχο συσχέτισης του Spearman. Με κόκκινο χρώμα η μηδενική υπόθεση γίνεται αποδεκτή, ενώ με γκρι απορρίπτεται.

Καθώς δεν έχουμε κάνει την υπόθεση προέρχονται από την κανονική κατανομή, και δεν υπάρχει κάποια τέτοια ένδειξη από τα διαγράμματα του κεφαλαίου 4.2.1, στο Διάγραμμα 4.9 παρουσιάζονται τα αποτελέσματα του μη παραμετρικού ελέγχου συσχέτισης κατά Spearman με χρήση της ελεγχοσυνάρτησης 4.2.4.

- Για την μεταβλητή fM3Trans δεν απορρίπτουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας 5%, καθώς η αντίστοιχη p-τιμή είναι αρκετά μεγάλη σε κάθε έλεγχο με τις υπόλοιπες μεταβλητές.

- Επίσης δεν απορρίπτουμε την μηδενική υπόθεση ότι είναι ασυσχέτιστες η μεταβλητή fConc1 με την fAsym, καθώς και την μεταβλητή απόκρισης output. αφού οι αντίστοιχες p-τιμές σε ε.σ. 5% είναι αρκετά μεγάλες.

- Σε κάθε άλλη περίπτωση, υπάρχουν ισχυρές ενδείξεις έναντι της μηδενικής υπόθεσης επομένως την απορρίπτουμε. Δηλαδή υπάρχουν συσχετίσεις ανάμεσα στα ζεύγη των μεταβλητών.

Κλείνοντας, αξίζει να σημειωθεί ότι η ανάλυση που έγινε σε αυτό κεφάλαιο δεν έχει στόχο να καταλήξει εάν θα πρέπει να βγει κάποια μεταβλητή από το μοντέλο. Αυτό θα ήταν επισφαλές αφού εν γένει τα μοντέλα που θα προσαρμοστούν στο τέλος του κεφαλαίου είναι μη γραμμικά,

και όπως προαναφέρθηκε και για το VIF, η ύπαρξη συσχετίσεων και η πολυσυγγραμμικότητα δεν σημαίνει απαραίτητα ότι θα έχει κακή επίδοση το τελικό μοντέλο. Αντίθετα, η παραπάνω μελέτη βοηθάει στην περαιτέρω κατανόηση των δεδομένων και θα πρέπει να ληφθεί υπόψιν σε συνδυασμό με άλλα αποτελέσματα προκειμένου να καταλήξει κανείς σε κάποιο συμπέρασμα.

4.2.3 Προβολή σε χαμηλότερη διάσταση

Η απλούστερη τεχνική προβολής των δεδομένων σε χαμηλότερη διάσταση, είναι με την μέθοδο ανάλυσης των κύριων συνιστωσών (Principal Component Analysis, PCA). Η PCA στοχεύει στην προβολή των δεδομένων σε ένα νέο σύστημα συντεταγμένων μετασχηματίζοντας γραμμικά τα αρχικά δεδομένα. Ενδέχεται το νέο σύστημα να έχει μικρότερη διάσταση από τα δεδομένα, γι΄ αυτό τον λόγο η PCA χρησιμοποιείται, εκτός από την προβολή δεδομένων σε χαμηλότερη διάσταση (2 ή 3), στην συμπίεση των δεδομένων.

Το νέο σύστημα συντεταγμένων είναι ορθογώνιο και οι νέες συντεταγμένες αποτελούν γραμμικό συνδυασμό των προηγούμενων μεταβλητών, με τέτοιο τρόπο ώστε οι συντεταγμένες της πρώτης νέας «μεταβλητής» (ή πρώτου κύριου «συστατικού») να περιέχουν την περισσότερη πληροφορία (ή την περισσότερη διακύμανση), το δεύτερο συστατικό να περιέχει πληροφορία που δεν υπάρχει στο πρώτο κ.ο.κ. Αυτό γίνεται εφικτό υπολογίζοντας τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα συνδιακύμανσης $C \in \mathbb{R}_{p\times p}$. Τότε εν λόγω πίνακας γράφεται ως:

$$C = \begin{bmatrix} Var\left(x_1\right) & \cdots & Cov\left(x_1, x_p\right) \\ \vdots & \ddots & \vdots \\ Cov\left(x_p, x_1\right) & \cdots & Var\left(x_p\right) \end{bmatrix} = \lambda_1 \boldsymbol{u_1} \boldsymbol{u_1}^{\mathrm{T}} + \ldots + \lambda_p \boldsymbol{u_p} \boldsymbol{u_p}^{\mathrm{T}}, \tag{4.2.8}$$

όπου λ_i και $\boldsymbol{u_i}$ οι αντίστοιχες ιδιοτιμές και τα ιδιοδιανύσματα, $i = 1, \ldots, p$. Παρατηρούμε ότι την μεγαλύτερη επίδραση έχουν τα ιδιοδιανύσματα στα οποία αντιστοιχεί μεγαλύτερη κατ΄ απόλυτο ιδιοτιμή λ_i. Επομένως σκοπός της PCA είναι να κατατάξει τα ιδιοδιανύσματα βάση της φθίνουσας κατ΄ απόλυτο ιδιοτιμής, και στην συνέχεια χρησιμοποιώντας τα πρώτα m ιδιοδιανύσματα προβάλλονται τα αρχικά δεδομένα στον επιθυμητό m-διάστατο χώρο.

Για τον υπολογισμό του παραπάνω πίνακα συνδιακύμανσης, αρχικά στον υπάρχον πίνακα των παρατηρήσεων $X = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_p}]$ θα αφαιρεθεί η αντίστοιχη δειγματική μέση τιμή $\overline{x_i}$, οποία αποτελεί αμερόληπτη εκτιμήτρια της πληθυσμιακής μέσης τιμής μ_i.

Αφαιρώντας την αντίστοιχη μέση τιμή λαμβάνουμε έναν νέο πίνακα $X^* = \left[\boldsymbol{x_1^*}, \ldots, \boldsymbol{x_p^*}\right]$ όπου κάθε μεταβλητή προέρχεται από κατανομή με μηδενική μέση τιμή. Τότε ο πίνακας:

$$X^* \left(X^*\right)^T = \begin{bmatrix} \sum_{m=1}^{n} \left(x^*_{m,1} \cdot x^*_{m,1}\right) & \cdots & \sum_{m=1}^{n} \left(x^*_{m,1} \cdot x^*_{m,p}\right) \\ \vdots & \ddots & \vdots \\ \sum_{m=1}^{n} \left(x^*_{m,p} \cdot x^*_{m,1}\right) & \cdots & \sum_{m=1}^{n} \left(x^*_{m,p} \cdot x^*_{m,p}\right) \end{bmatrix} \tag{4.2.9}$$

$$= \begin{bmatrix} \sum_{m=1}^{n} \left(x_{m,1} - \overline{x_1}\right)^2 & \cdots & \sum_{m=1}^{n} \left(x_{m,1} - \overline{x_1}\right)\left(x_{m,p} - \overline{x_p}\right) \\ \vdots & \ddots & \vdots \\ \sum_{m=1}^{n} \left(x_{m,p} - \overline{x_p}\right)\left(x_{m,1} - \overline{x_1}\right) & \cdots & \sum_{m=1}^{n} \left(x_{m,p} - \overline{x_p}\right)^2 \end{bmatrix} \tag{4.2.10}$$

είναι ανάλογος του πίνακα συνδιακύμανσης, αφού αποδεικνύεται (Κοκολάκης & Φουσκάκης, 2009) για τις αντίστοιχες ελεγχοσυναρτήσεις:

$$T_{i,i} = \sum_{m=1}^{n} \left(x_{m,i} - \overline{x_i}\right)^2 \rightarrow E\left[T_{i,i}\right] = (n-1)\,\sigma_i^2 \tag{4.2.11}$$

$$T_{i,j} = \sum_{m=1}^{n} \left(x_{m,i} - \overline{x_i}\right)\left(x_{m,j} - \overline{x_j}\right) \rightarrow E\left[T_{i,j}\right] = (n-1)\,Cov\left(X_i, X_j\right) \tag{4.2.12}$$

και τελικά καταλήγουμε για τον πίνακα συνδιακύμανσης ότι:

$$C = \frac{\left(X^*\right)^T X^*}{n-1}. \tag{4.2.13}$$

Το επόμενο βήμα για την ανάλυση των κύριων συνιστωσών, είναι η εύρεση των εν λόγω ιδιοτιμών και ιδιοδιανυσμάτων. Για να γίνει αυτό αρχικά ο πίνακας X^* με την βοήθεια της μεθόδου SVD, γράφεται στην ακόλουθη μορφή:

$$X^* = U\Sigma V^T \tag{4.2.14}$$

όπου οι $U, V \in \mathbb{R}_{n\times p}$ ημι-ορθογώνιοι πίνακες και $\Sigma \in \mathbb{R}_{p\times p}$ διαγώνιος πίνακας Άρα η σχέση 4.2.13 με την βοήθεια της 4.2.14 γράφεται ως:

$$C = \frac{\left(U\Sigma V^T\right)^T U\Sigma V^T}{n-1} \tag{4.2.15}$$

$$= \frac{V\Sigma^T U^T U\Sigma V^T}{n-1} \tag{4.2.16}$$

$$= V\frac{\Sigma^2}{n-1}V^T. \tag{4.2.17}$$

Επειδή όμως ο πίνακας συνδιακύμανσης είναι συμμετρικός, διαγωνοποιείται ως:

$$C = P\Lambda P^T \tag{4.2.18}$$

όπου P ο πίνακας με τα ιδιοδιανύσματα και Λ ο διαγώνιος πίνακας με τις ιδιοτιμές. Από τις 4.2.17 και 4.2.18 συμπεραίνουμε ότι $\Lambda = \Sigma^2/(n-1)$ και $P = V$. Άρα έχουμε τις ζητούμενες ιδιοτιμές και ιδιοδιανύσματα του πίνακα συνδιακύμανσης. Το τελευταίο βήμα στην PCA είναι να προβάλλουμε τις παρατηρήσεις στην διάσταση που μας ενδιαφέρει.

Για παράδειγμα αν μας ενδιαφέρει να προβάλλουμε τα δεδομένα σε 2 διαστάσεις, τότε βρίσκουμε τα ιδιοδιανύσματα $\boldsymbol{v_1}$, $\boldsymbol{v_2}$ του πίνακα V που αντιστοιχούν στις 2 μεγαλύτερες ιδιοτιμές του C ή ισοδύναμα στις 2 μεγαλύτερες κατ΄ απόλυτο ιδιάζουσες τιμές του πίνακα Σ. Στην συνέχεια λαμβάνουμε το πρώτο και το δεύτερο κύριο συστατικό κάνοντας τον ακόλουθο πολλαπλασιασμό:

$$\left[\boldsymbol{PCA_{(1)}}, \boldsymbol{PCA_{(2)}}\right] = \left[\boldsymbol{x_1}, \ldots, \boldsymbol{x_p}\right] \left[\boldsymbol{v_1}, \boldsymbol{v_2}\right] \tag{4.2.19}$$

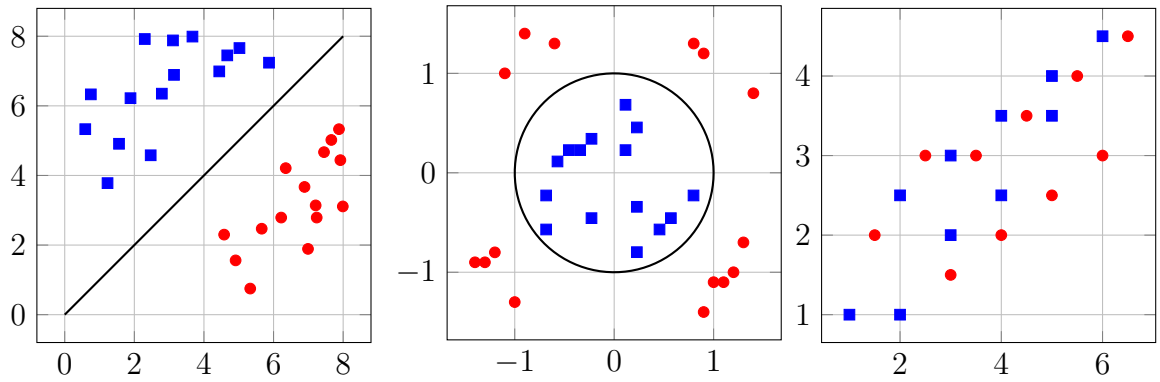
Τα πιθανά σενάρια για το διάγραμμα που θα λάβουμε (βλ. Διάγραμμα 4.10) προβάλλοντας τα δεδομένα σε 2 ή 3 διαστάσεις είναι τα εξής:

1. Τα δεδομένα ανάμεσα στις 2 κλάσεις του προβλήματος είναι γραμμικά διαχωρίσιμα. Τότε ακόμα και ένας απλός γραμμικός ταξινομητής μπορεί να επιλύσει το πρόβλημα με αρκετά καλή επίδοση.

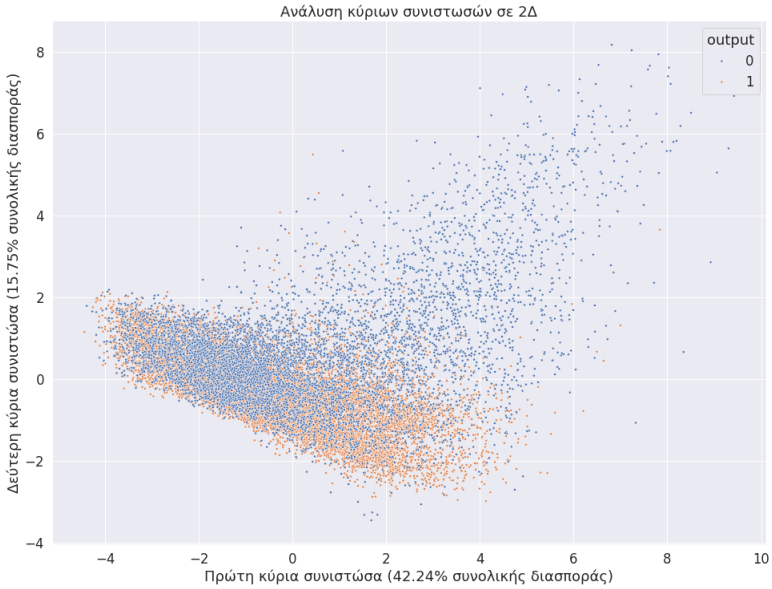
2. Οι παρατηρήσεις είναι διαχωρίσιμες όχι όμως γραμμικά. Σε αυτή την περίπτωση αναμένουμε ο γραμμικός ταξινομητής να ταξινομήσει στην λάθος κλάση ένα σημαντικό κομμάτι των παρατηρήσεων. Ενδεχομένως ένας μη γραμμικός ταξινομητής να μπορέσει να επιλύσει με μεγαλύτερη επιτυχία το πρόβλημα.

3. Οι παρατηρήσεις δεν διαχωρίζονται με κάποιον τρόπο, δηλαδή υπάρχει μεγάλη επικάλυψη μεταξύ των κλάσεων. Αυτό μπορεί να οφείλεται στο γεγονός ότι μεγάλο κομμάτι της διακύμανσης υπολείπεται στα εναπομείναντα κύρια συστατικά. Με άλλα λόγια ενδέχεται το πρόβλημα να παραμένει διαχωρίσιμο απλώς να μην μπορούμε να το δούμε σε λιγότερες διαστάσεις. Διαφορετικά μετασχηματίζοντας κάποιες μεταβλητές ή πραγματοποιώντας κάποιου είδους δειγματοληψία μέσα από το αρχικό δείγμα μπορεί να συμβάλλει στην βελτίωση της επίδοσης των ταξινομητών.

Διάγραμμα διασποράς ανάμεσα στις πρώτες 2 κύριες συνιστώσες



Διάγραμμα 4.10: Περιπτώσεις μετά την προβολή σε 2 διαστάσεις. Με μπλε και κόκκινο εναλλάσσονται τα δεδομένα της εκάστοτε κλάσης. Αριστερό διάγραμμα: Γραμμικά διαχωρίσιμα σημεία. Μεσαίο διάγραμμα: Μη-γραμμικώς διαχωρίσιμα. Δεξί διάγραμμα: Μη-διαχωρίσιμα.



Διάγραμμα 4.11: Προβολή σε 2 Διαστάσεις των δεδομένων Magic Telescope μέσω ανάλυσης PCA.

Στο Διάγραμμα 4.11 παρατηρούμε το διάγραμμα διασποράς των πρώτων δύο κύριων συστατικών που προκύπτουν προβάλλοντας τα υπό μελέτη δεδομένα σε 2 διαστάσεις. Οι παρατηρήσεις της κλάσης 0 φαίνεται να είναι συγκεντρωμένες στην κάτω αριστερή περιοχή του διαγράμματος, ενώ σε ένα κομμάτι επικαλύπτονται με αυτές της κλάσης 1. Εκ πρώτης όψεως δεν φαίνονται διαχωρίσιμα, παρόλα αυτά οι πρώτες δύο κύριες συνιστώσες περιλαμβάνουν λιγότερο από το 60% της συνολικής διακύμανσης. Επομένως σημαντικό μέρος της πληροφορίας δεν είναι εμφανές στο εν λόγω διάγραμμα.

4.3 Ισοστάθμιση Δεδομένων

Όπως φάνηκε και στο ιστόγραμμα του Διαγράμματος 4.5, οι παρατηρήσεις της κλάσης 1 αποτελούν σημαντικά μεγαλύτερο ποσοστό των δεδομένων, έναντι αυτών της κλάσης 0. Η ανισορροπία μεταξύ των δύο κλάσεων αποτελεί πρόβλημα σε έναν ταξινομητή, καθώς ενδέχεται να αναπτύξει χαρακτηριστικά που αγνοούν τις παρατηρήσεις της μειοψηφικής κλάσης. Αυτό θα σήμαινε για το πρόβλημα της δυαδικής ταξινόμησης ότι ο ταξινομητής θα χαρακτήριζε κάθε παρατήρηση ως «κλάση 1», και θα είχε ποσοστό επιτυχίας 65%. Το παραπάνω φαινόμενο γίνεται ακόμα πιο έντονο σε προβλήματα ταξινόμησης με παραπάνω από δύο κλάσεις, όπου μπορεί να υπάρχουν περισσότερες κατηγορίες της μεταβλητής απόκρισης με ελάχιστες παρατηρήσεις.

Πιθανές λύσεις στο παραπάνω πρόβλημα είναι:

- Η αξιολόγηση του ταξινομητή με μεγαλύτερη ποινή στην λανθασμένη ταξινόμηση των παρατηρήσεων της μειοψηφικής κλάσης.
- Η επαύξηση των δεδομένων της μειοψηφικής κλάσης (over-sampling).
- Η επιλογή μόνο ενός μέρους από τις παρατηρήσεις της πλειοψηφικής κλάσης (under-sampling).

Στην πρώτη πιθανή λύση, εάν η επιλογή της εν λόγω ποινής δεν γίνει με προσοχή, στην πραγματικότητα μπορεί να επιφέρει το αντίθετο αποτέλεσμα απ΄ ότι εάν χρησιμοποιούσαμε τα αρχικά δεδομένα όπως ήταν. Δηλαδή ο ταξινομητής θα τείνει να ταξινομεί όλες τις παρατηρήσεις στην μειοψηφική κλάση (κλάση 0), καταλήγοντας σε ένα τετριμμένο μοντέλο.

Η δεύτερη λύση, της επαύξησης των δεδομένων της μειοψηφικής κλάσης, περιλαμβάνει τεχνικές οι οποίες επαναλαμβάνουν τα αρχικά δεδομένα ή δημιουργούν νέες παρατηρήσεις παρεμβάλλοντας τα υπάρχοντα δεδομένα της εν λόγω κλάσης. Η μεγαλύτερη πρόκληση σε αυτή την μέθοδο, αποτελεί η διατήρηση της ποιότητας των δεδομένων, καθώς μέσα από το over-sampling ενδέχεται να επαναληφθούν «σπάνιες» παρατηρήσεις ή γενικότερα να αλλοιωθεί η κατανομή των χαρακτηριστικών του πληθυσμού. Με άλλα λόγια υπάρχει ο κίνδυνος το σύνολο δεδομένων που θα προκύψει να μην είναι αντιπροσωπευτικό αφού θα υπάρχει μεροληψία (bias).

Η τρίτη μέθοδος ισοστάθμισης των δεδομένων, με χρήση ενός τμήματος των παρατηρήσεων της πλειοψηφικής κλάσης, αποτελεί την τελική επιλογή για τα υπό μελέτη δεδομένα αυτής της ενότητας. Το μειονέκτημα αυτής της επιλογής είναι ότι κατά την αφαίρεση παρατηρήσεων από το σύνολο δεδομένων, χάνεται πληροφορία από το μοντέλο και ενδέχεται να χαθούν μοτίβα συμπεριφοράς των δεδομένων τα οποία θα διευκόλυναν τον ταξινομητή στην επίλυση του προβλήματος. Επιπλέον θα πρέπει να ληφθεί υπόψιν ότι η υπο-δειγματοληψία του πληθυσμού μπορεί να οδηγήσει σε ένα αρκετό μικρό σύνολο δεδομένων μη επαρκές.

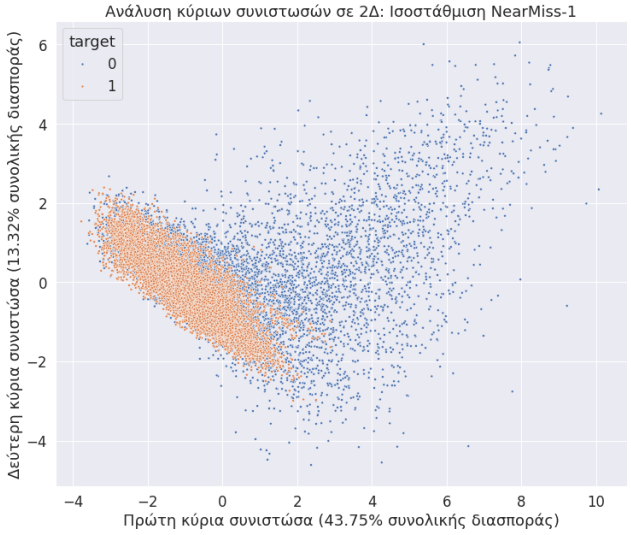
Το σύνολο δεδομένων MagicTelescope, αποτελείται από 19,020 παρατηρήσεις, το οποίο διαμερίζεται ως εξής:

- Κλάση 0: 6,688 παρατηρήσεις (35.2%).
- Κλάση 1: 12,332 παρατηρήσεις (64.8%).

Για να έρθουν σε μια ισορροπία τα δεδομένα θα πρέπει να αφαιρεθούν παρατηρήσεις από την πλειοψηφική κλάση 1. Ο απλούστερος τρόπος για να γίνει είναι με τυχαία αφαίρεση παρατηρήσεων από το σύνολο δεδομένων. Αυτό όμως δεν θα αναιρούσε τα μειονεκτήματα που προαναφέρθηκαν, επομένως αξίζει να αναζητήσει κανείς εναλλακτική μεθοδολογία (Zhang & Mani, 2003):

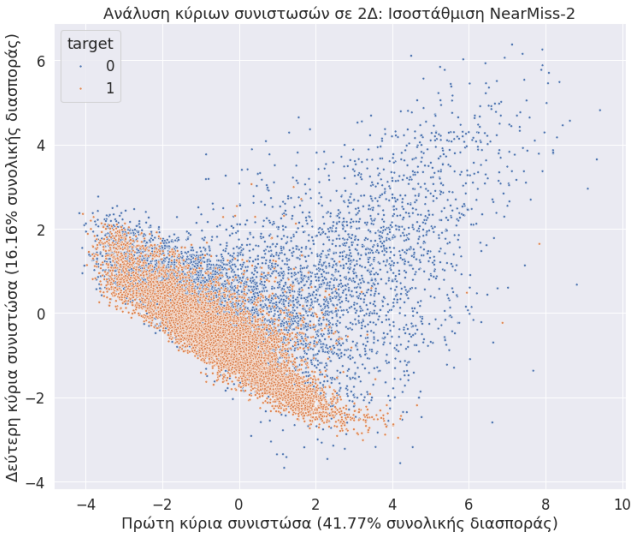
- NearMiss-1: Επιλέγει τα στοιχεία της πλειοψηφικής κλάσης τα οποία έχουν την μικρότερη μέση απόσταση από N στοιχεία της μειοψηφικής κλάσης.

Διάγραμμα 4.12: PCA ανάλυση 2 διαστάσεων των δεδομένων με εφαρμογή της NearMiss-1.



- NearMiss-2: Επιλέγει τα στοιχεία της πλειοψηφικής κλάσης τα οποία έχουν την μεγαλύτερη μέση απόσταση από N στοιχεία της μειοψηφικής κλάσης.
- NearMiss-3: Για κάθε στοιχείο της μειοψηφικής κλάσης επιλέγει πρώτα τους M κοντινότερους γείτονες της πλειοψηφικής κλάσης. Εν συνεχεία για κάθε ένα από τα εναπομείναντα στοιχεία της πλειοψηφικής κλάσης, παραμένουν στο τελικό σύνολο δεδομένων εκείνα που έχουν την μεγαλύτερη μέση απόσταση από τους N γείτονες της μειοψηφικής κλάσης.

Στην πραγματικότητα οι παραπάνω μεθοδολογίες επιχειρούν να απορρίψουν σημεία τα οποία είτε απέχουν πολύ μακριά το ένα από το άλλο, σαν να είναι έκτροπα, είτε είναι πολύ κοντά το ένα με το άλλο, σαν να επικαλύπτονται. Οι παράμετροι M, N επιλέγονται από τον χρήστη, ενώ ως απόσταση μπορεί να χρησιμοποιηθεί η Ευκλείδεια απόσταση:



Διάγραμμα 4.13: PCA ανάλυση 2 διαστάσεων των δεδομένων με εφαρμογή της NearMiss-2.

$$d\left(\boldsymbol{x_1}, \boldsymbol{x_2}\right) = \sqrt{\left(x_{1,1} - x_{2,1}\right)^2 + \ldots + \left(x_{1,p} - x_{2,p}\right)^2}, \qquad \forall \boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^p. \tag{4.3.1}$$

Τα Διαγράμματα 4.12, 4.13 & 4.14 προβάλλουν τα δεδομένα σε 2 διαστάσεις με την μέθοδο ανάλυσης των κύριων συνιστωσών, όπως αυτή αναπτύχθηκε στην προηγούμενη υποενότητα 4.2.3. Σε κάθε διάγραμμα οι παρατηρήσεις της πλειοψηφικής κλάσης 1 υποβλήθηκαν σε κάθε μια από τις μεθόδους NearMiss-1, 2 & 3 αντίστοιχα, με N=M=3 μέχρι οι δύο κλάσεις να έχουν παρόμοια πληθικότητα.

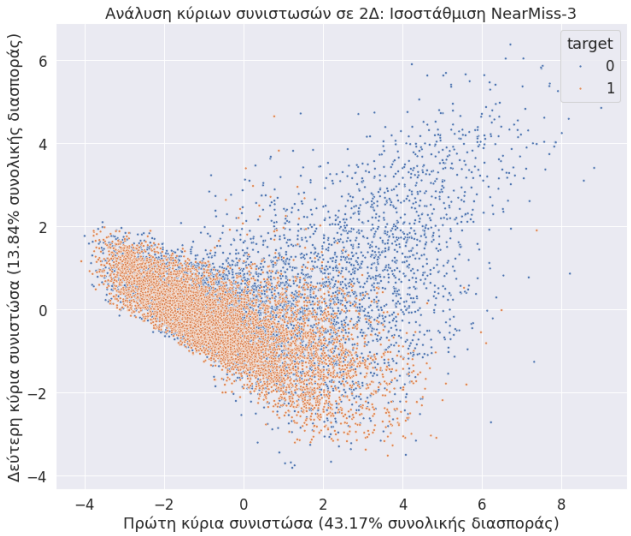
Και στα 3 διαγράμματα παρατηρούμε ότι τα σημεία έχουν λιγότερη επικάλυψη σε σχέση με το αρχικό διάγραμμα 4.11. Επιπλέον, δεν υπάρχει σημαντική μεταβολή στο ποσοστό της συνολικής διασποράς που περιλαμβάνει το πρώτο και το δεύτερο κύριο συστατικό. Στο Διάγραμμα 4.13 της NearMiss-2, παρατηρούμε ότι τα αντίστοιχα σημεία των κλάσεων 0 & 1, φαίνεται να μπορούν να διαχωριστούν με μία γραμμή, σε αντίθεση με τα υπόλοιπα διαγράμματα όπου τα αντίστοιχα σημεία είναι περισσότερο ανακατεμένα και οι παρατηρήσεις της κλάσης 0 περιβάλλουν αυτά της κλάσης 1.

4.4 Αξιολόγηση Ταξινομητών

4.4.1 Χωρισμός Δεδομένων

Η αξιολόγηση της επίδοσης ενός ταξινομητή αποτελεί ένα από τα σημαντικότερα χαρακτηριστικά σε ένα πρόβλημα μηχανικής μάθησης. Η επιλογή κατάλληλης μεθόδου αξιολόγησης διαφέρει από την φύση του προβλήματος, και πρέπει να γίνει προσεκτικά καθώς επηρεάζει άμεσα την επίδοση του τελικού μοντέλου που θα προκύψει.

Η αξιολόγηση στους περισσότερους ταξινομητές αποτελεί κομμάτι στην διαδικασία της



Διάγραμμα 4.14: PCA ανάλυση 2 διαστάσεων των δεδομένων με εφαρμογή της NearMiss-3.

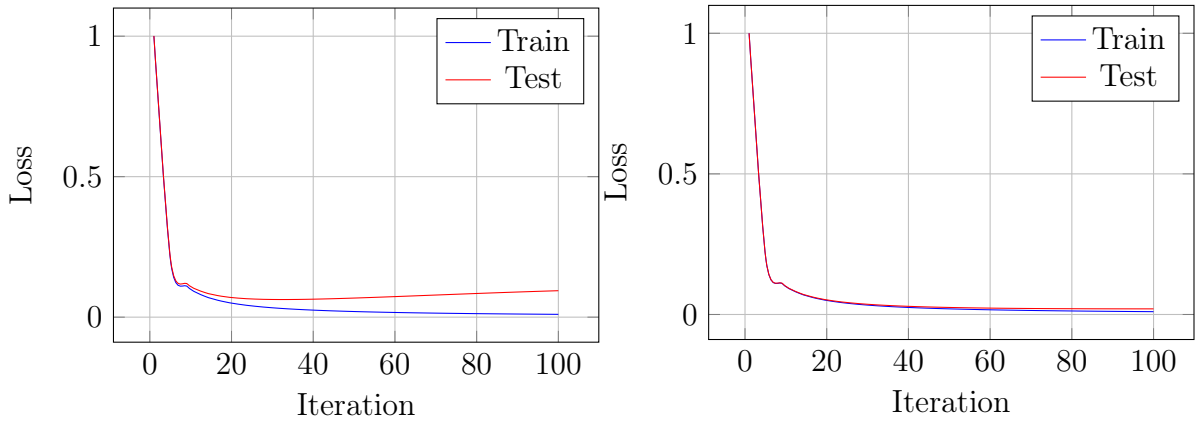
μάθησης, και χρησιμεύει επίσης στην επικύρωση του μοντέλου. Πιο συγκεκριμένα, κατά την διάρκεια της μάθησης η δομή του εκάστοτε ταξινομητή διαμορφώνεται κατάλληλα σε κάθε επανάληψη, προκειμένου να βελτιωθεί κατά το δυνατόν περισσότερο η επίδοση βάση της μετρικής που έχει επιλεχθεί. Το πρόβλημα που προκύπτει είναι ότι ορισμένοι ταξινομήτες όπως τα Νευρωνικά Δίκτυα, έχουν την τάση να αποστηθίζουν το σύνολο δεδομένων, και να εμφανίζουν κακή επίδοση σε παρατηρήσεις εκτός του αρχικού συνόλου δεδομένων (πρόβλημα γνωστό ως υπερπροσαρμογή). Για τον σκοπό αυτό το αρχικό σύνολο δεδομένων, συχνά χωρίζεται σε δύο επιμέρους σύνολα:

- Σύνολο εκπαίδευσης: Το σύνολο εκπαίδευσης αποτελείται από τις παρατηρήσεις με βάση τις οποίες εκπαιδεύεται ο ταξινομητής και μαθαίνει να λύνει το πρόβλημα. Συνήθως είναι το 50% με 80% των αρχικών δεδομένων.

- Σύνολο δοκιμής ή αξιολόγησης: Το σύνολο δοκιμής περιέχει τις εναπομείναντες παρατηρήσεις από το αρχικό σύνολο, οι οποίες δεν χρησιμοποιήθηκαν κατά την διάρκεια της εκπαίδευσης. Η αξιολόγηση του ταξινομητή στον εν λόγω σύνολο δείχνει πόσο καλά γενικεύεται το μοντέλο, και σε σύγκριση με την επίδοση στο σύνολο εκπαίδευσης μπορεί να υποδείξει μη επιθυμητές συμπεριφορές, όπως εκείνη του Διαγράμματος 4.15 όπου έχει γίνει υπερπροσαρμογή (overfitting).

Συχνά η συμπεριφορά των ταξινομητών μπορεί να παραμετροποιηθεί από την χρήστη, επιλέγοντας κατάλληλες «υπέρ-παραμέτρους» (hyper-parameters). Για παράδειγμα, με βάση όσα αναφέρθηκαν και σε προηγούμενες ενότητες, σε ένα γενετικό αλγόριθμο οι αντίστοιχοι υπέρ-παράμετροι είναι:

- Το μέγεθος το πληθυσμού N_p.



Διάγραμμα 4.15: Υπερπροσαρμογή: Στο αριστερό διάγραμμα το σφάλμα στο σύνολο εκπαίδευσης μειώνεται ενώ στο σύνολο δοκιμής αυξάνεται. Το δεξί διάγραμμα παρουσιάζει την ιδανική περίπτωση.

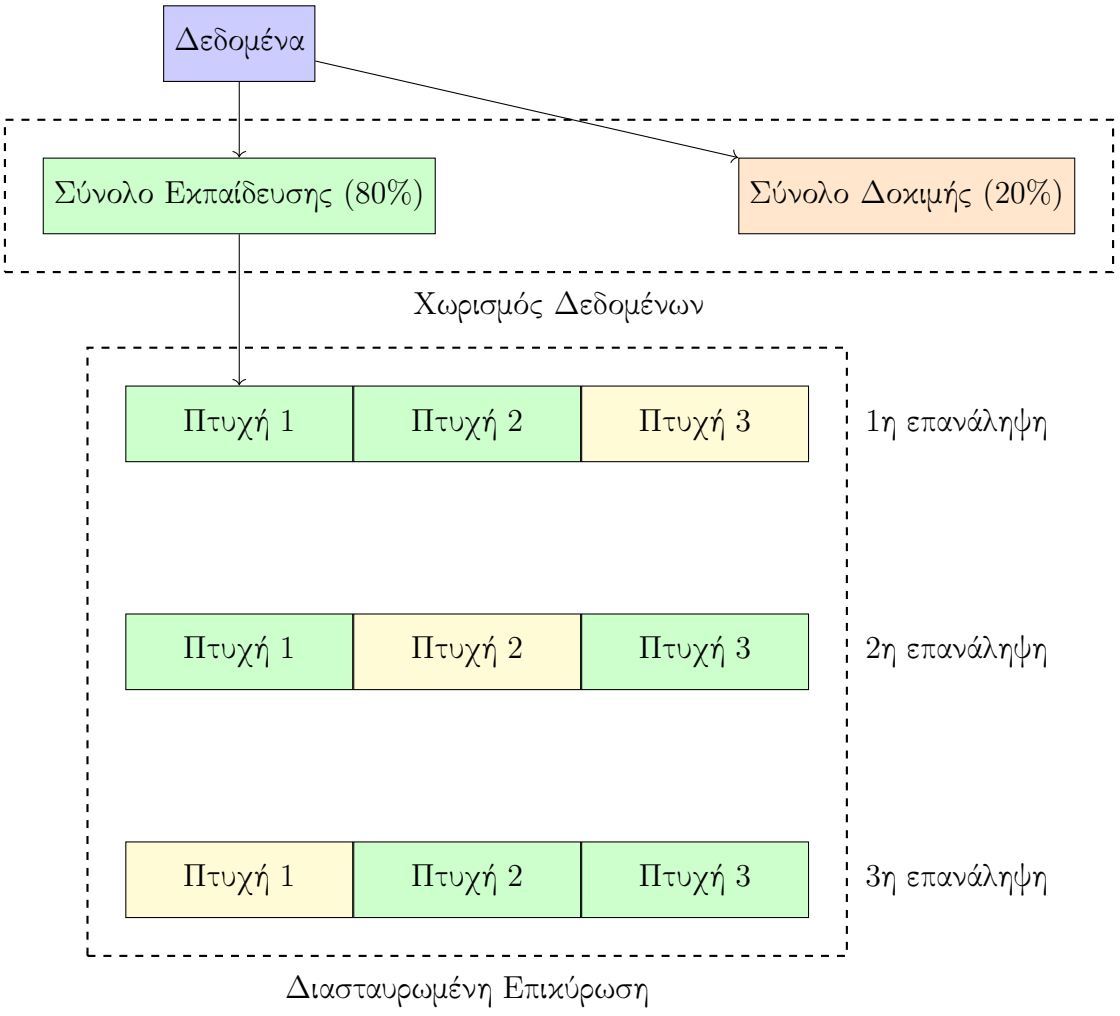
- Η πιθανότητα επιλογής P_s.
- Η πιθανότητα διασταύρωσης P_c.
- Η πιθανότητα μετάλλαξης P_m.
- Κάποιο κριτήριο τερματισμού, που σηματοδοτεί την ολοκλήρωση της διαδικασίας εκμάθησης.

Οι εν λόγω παράμετροι προφανώς εξαρτώνται από το πρόβλημα που καλούμαστε να επιλύσουμε, αφετέρου δεν είναι γνωστοί εκ των προτέρων. Στην περίπτωση που οι υπερπαράμετροι καθορίζονταν από την επίδοση του ταξινομητή στο σύνολο δοκιμής, τότε θα χειραγωγούσαμε την διαδικασία της εκπαίδευσης για να έχει τα επιθυμητά αποτελέσματα στο σύνολο δοκιμής. Αυτό αναιρεί τον λόγο για τον οποίο τα αρχικά δεδομένα χωρίστηκαν σε δύο επιμέρους σύνολα, για την εκπαίδευση και την αξιολόγηση, αφού υπάρχει διαρροή πληροφορίας από το σύνολο αξιολόγησης σε αυτό της εκπαίδευσης, δηλαδή εμφανίζεται η υπερεκπαίδευση.

Η λύση στο πρόβλημα της επιλογής κατάλληλων παραμέτρων δίνεται με έναν από τους εξής τρόπους:

- Δημιουργία συνόλου επικύρωσης (validation set): Το σύνολο εκπαίδευσης χωρίζεται σε δύο επιμέρους σύνολα. Το πρώτο σύνολο χρησιμοποιείται για την εκπαίδευση (∼80% των αρχικών δεδομένων εκπαίδευσης), ενώ το δεύτερο για την επικύρωση, δηλαδή την αξιολόγηση του ταξινομητή με τις επιλεγμένες παραμέτρους.
- Διασταυρωμένη Επικύρωση (cross-validation): Η διασταυρωμένη επικύρωση αποδίδει ακόμα και όταν τα δεδομένα περιέχουν λίγες παρατηρήσεις. Αυτό οφείλεται στο γεγονός ότι τα σύνολα εκπαίδευσης και επικύρωσης, εναλλάσσονται διαδοχικά σε γύρους,

με σκοπό κάθε παρατήρηση να βρεθεί στο σύνολο επικύρωσης τουλάχιστον μία φορά. Στην συνέχεια η επίδοση του ταξινομητή, για τον επιλεγμένο συνδυασμό υπερπαραμέτρων, υπολογίζεται ως η μέση τιμή της επίδοσης στον σύνολο επικύρωσης του εκάστοτε γύρου. Αρκετά διαδεδομένη τεχνική διασταυρωμένης επικύρωσης, είναι σε κ-πτυχές (k-fold cross validation), όπου δημιουργούνται κ-πτυχές παρόμοιας πληθικότητας, εκ των οποίων σε κάθε γύρο οι κ-1 έχουν το ρόλο του συνόλου εκπαίδευσης, ενώ αυτή που απομένει του συνόλου επικύρωσης. Το Διάγραμμα 4.16 παρουσιάζει την εν λόγω διαδικασία. Τέλος τονίζεται ότι ο αριθμός των πτυχών δεν έχει νόημα να είναι αρκετά μεγάλος (πάνω από 5), διότι τότε το σύνολο εκπαίδευσης θα αποτελείται από σχεδόν όλα τα δεδομένα, ενώ το σύνολο επικύρωσης θα τείνει να γίνει κενό.



Διάγραμμα 4.16: Διασταυρωμένη επικύρωση με 3 πτυχές: Το σύνολο εκπαίδευσης χωρίζεται σε επιμέρους υποσύνολα για την εύρεση των κατάλληλων υπερ-παραμέτρων. Σε κάθε επανάληψη οι πράσινες πτυχές αποτελούν το σύνολο εκπαίδευσης, ενώ η κίτρινη το σύνολο επικύρωσης.

4.4.2 Συναρτήσεις αξιολόγησης στην Δυαδική ταξινόμηση

Στο πρόβλημα της δυαδικής ταξινόμησης, εν γένει θέλουμε να ελαχιστοποιήσουμε το πλήθος των λάθος ταξινομημένων παρατηρήσεων. Συχνά τα αποτελέσματα παρουσιάζονται με την μορφή του ακόλουθου πίνακα, ο οποίος καλείται πίνακας σύγχυσης, και συμβολίζεται ως:

- TP (true positives): Το πλήθος των σωστά ταξινομημένων παρατηρήσεων στην κλάση 1.
- FP (false positives): Το πλήθος των εσφαλμένα ταξινομημένων παρατηρήσεων στην κλάση 1.
- TN (true negatives): Το πλήθος των σωστά ταξινομημένων παρατηρήσεων στην κλάση 0 .
- FN (false negatives): Το πλήθος των εσφαλμένα ταξινομημένων παρατηρήσεων στην κλάση 0.

		Προβλεπόμενη κλάση	
		Κλάση 0	Κλάση 1
Πραγματική κλάση	Κλάση 0	TN	FN
	Κλάση 1	FP	TP

Πίνακας 3: Πίνακας σύγχυσης

Παρατηρούμε ότι οι γραμμές του Πίνακα 3 αθροίζουν στο πλήθος των παρατηρήσεων που ανήκουν στην κάθε κλάση αντίστοιχα, σύμφωνα με τα δεδομένα, ενώ οι στήλες αθροίζουν στο πλήθος των παρατηρήσεων που ταξινομήθηκαν στην κλάση 0 ή 1 αντίστοιχα.

Ένας ιδανικός ταξινομητής θα αντιστοίχιζε τις παρατηρήσεις στην πραγματική κλάση την οποία ανήκουν, επομένως θα είχε καλή ορθότητα (accuracy). Η ορθότητα σε ένα πρόβλημα ταξινόμησης υπολογίζεται ως:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{4.4.1}$$

Η παραπάνω ποσότητα είναι φραγμένη στο $[0, 1]$, και τείνει στο 1, όταν οι λάθος ταξινομήσεις FP & FN τείνουν στο μηδέν.

Αρκετές φορές δεν είναι εφικτό να γίνεται ταυτόχρονη μείωση των εσφαλμένα ταξινομημένων παρατηρήσεων στη κλάση 0 ή 1. Επιπλέον, το σκορ της ορθότητας 4.4.1 δεν προσφέρει

κάποια παραπάνω πληροφορία σχετικά με το αν ο ταξινομητής πραγματοποιεί περισσότερες εσφαλμένες ταξινομήσεις FP ή FN, ενώ ενδέχεται για το πρόβλημα ταξινόμησης που καλούμαστε να επιλύσουμε να ενδιαφερόμαστε στο να μεγιστοποιήσουμε τις σωστές ταξινομήσεις μόνο για την κλάση 0 ή την κλάση 1. Για τους παραπάνω λόγους συχνά αξιοποιούνται ως συναρτήσεις αξιολόγησης η ακρίβεια (precission) και η ανάκληση (recall) αντίστοιχα:

$$Precision = \frac{TP}{TP + FP} \tag{4.4.2}$$

$$Recall = \frac{TP}{TP + FN}. \tag{4.4.3}$$

Οι σχέσεις 4.4.2 & 4.4.3 για τα σκορ της ακρίβειας και την ανάκλησης ενός ταξινομητή, είναι φραγμένες στο $[0, 1]$.

Η ακρίβεια στοχεύει στην ελάττωση των λανθασμένων ταξινομήσεων στην κλάση 1. Αυτό σημαίνει ότι οι παρατηρήσεις που ταξινομούνται στην εν λόγω κλάση, ανήκουν πράγματι σε αυτή. Η ακρίβεια θα μπορούσε να είναι χρήσιμη, για παράδειγμα, στην αναγνώριση των ανεπιθύμητων μηνυμάτων. Εστιάζοντας στην ακρίβεια των ταξινομήσεων, μειώνεται το ενδεχόμενο να χαρακτηριστεί λανθασμένα η αλληλογραφία ως ανεπιθύμητη και να παραμείνει μη αναγνωσμένη από τον χρήστη.

Από την άλλη, η ανάκληση στοχεύει στην ελάττωση των λανθασμένων ταξινομήσεων στην κλάση 0. Δηλαδή μειώνει το ενδεχόμενο να υπάρχουν παρατηρήσεις οι οποίες ανήκουν στην κλάση 1, όμως δεν ανιχνεύθηκαν. Η ανάκληση συνήθως αποτελεί πολύτιμο εργαλείο σε ιατρικούς ελέγχους, για την ανίχνευση των ασθενών που πάσχουν από κάποια ασθένεια και χρήζουν περίθαλψης.

Οι παραπάνω συναρτήσεις αξιολόγησης βαθμολογούν την επίδοση του ταξινομητή εκ του αποτελέσματος και λαμβάνοντας υπόψιν μόνο την τρέχουσα κατανομή που ακολουθεί η κάθε κλάση. Η χαρακτηριστική καμπύλη ROC (receiver operating characteristic) περιγράφει την διακριτική ικανότητα του ταξινομητή για τις διάφορες οριακές τιμές t (threshold values). Ως οριακή τιμή χαρακτηρίζεται το κατώτερο όριο, πάνω από το οποίο μία παρατήρηση ταξινομείται στην κλάση 1. Για την κατασκευή της καμπύλης ROC θα πρέπει για τις διάφορες τιμές του $t \in [0, 1]$ να υπολογιστεί ο ρυθμός των εσφαλμένα θετικών ταξινομήσεων (false positive rate ή fpr) και ο ρυθμός των ορθώς θετικών ταξινομήσεων (true positive rate ή tpr):

$$tpr = \mathbb{P}\left[s > t | \boldsymbol{Y} = 1\right] \tag{4.4.4}$$

$$fpr = \mathbb{P}\left[s > t | \boldsymbol{Y} = 0\right]. \tag{4.4.5}$$

Για την εκτίμηση των παραπάνω, αρχικά θα πρέπει να εκτιμηθεί η πιθανότητα με την οποία ο ταξινομητής, ταξινομεί την κάθε παρατήρηση στην κλάση 1. Στην δυαδική ταξινόμηση, αυτό

γίνεται συνήθως με την βοήθεια της σιγμοειδούς συνάρτησης:

$$\hat{\mathbb{P}}\left[\boldsymbol{Y}=1\right]=s_i=\frac{1}{1+e^{-z_i}}, \tag{4.4.6}$$

όπου z_i η έξοδος του ταξινομητή, πριν την εφαρμογή της τελικής συνάρτησης ενεργοποίησης.

Εν συνεχεία, η παρατήρηση x_i ταξινομείται:

- Στην κλάση 1 εάν $s_i > t$.
- Στην κλάση 0 εάν $s_i \leq t$.

Με την βοήθεια του πίνακα σύγχυσης 3, για κάθε οριακή τιμή t εκτιμώνται οι 4.4.4 & 4.4.5 ως:

$$\widehat{tpr}=\frac{TP}{TP+FN} \tag{4.4.7}$$

$$\widehat{fpr}=\frac{FP}{FP+TN}. \tag{4.4.8}$$

Γενικότερα, η επιλογή κατάλληλης οριακής τιμής t που να διαχωρίζει ικανοποιητικά τις δύο κλάσεις δεν είναι εύκολη ούτε προφανής καθώς εξαρτάται και από την φύση του προβλήματος.

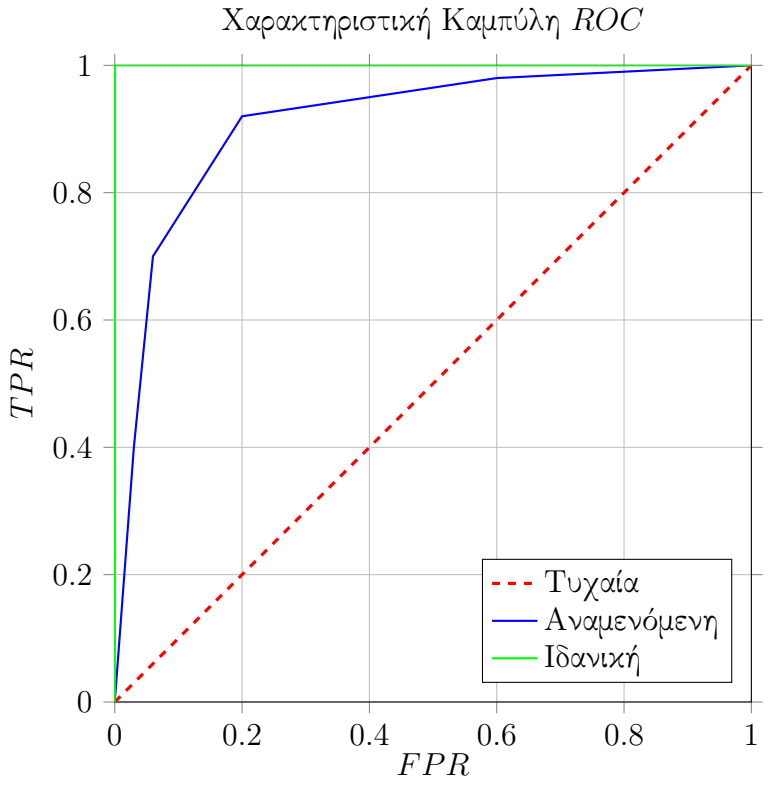
Στο Διάγραμμα 4.17 απεικονίζεται με μπλε η υποθετική καμπύλη που θα μπορούσε να έχει κάποιος ταξινομητής. Ιδανικά θα θέλαμε η εν λόγω καμπύλη να πλησιάζει την ιδανική, δηλαδή να είναι όσο το δυνατόν πιο μακρυά από την κόκκινη. Στην περίπτωση που λαμβάνουμε διάγραμμα όμοιο με αυτό της κόκκινης, τότε ο ταξινομητής στην πραγματικότητα πραγματοποιεί τις ταξινομήσεις με εντελώς τυχαίο τρόπο, το οποίο δεν επιθυμητό καθώς δεν αναμένουμε να έχει καλή προβλεπτική ικανότητα. Αντίθετα, όταν η καμπύλη πλησιάζει την ιδανική, αυτό σημαίνει ότι ο ταξινομητής, ανεξάρτητα της οριακής τιμής t, εκτιμά σωστά με μεγάλη πιθανότητα την κλάση από την οποία προέρχεται η κάθε παρατήρηση.

Ένας τρόπος να ποσοτικοποιηθούν οι παραπάνω παρατηρήσεις, είναι αξιοποιώντας το εμβαδόν της καμπύλης ROC (AUC). Το εν λόγω εμβαδόν ορίζεται ως:

$$AUC=\int_0^1 y\left(x\right)dx, \tag{4.4.9}$$

όπου y ο ρυθμός των ορθώς θετικών ταξινομήσεων (tpr) συναρτήσει του ρυθμού των εσφαλμένα θετικών ταξινομήσεων (fpr). Η παραπάνω έκφραση αποδεικνύεται ότι ισοδυναμεί με την πιθανότητα το σκορ που αποδίδει ο ταξινομητής στις TP παρατηρήσεις, να είναι μεγαλύτερο από αυτό των FP. Επιπλέον οι πιθανές ερμηνείες που μπορούν να αποδοθούν στο εμβαδόν και την αντίστοιχη καμπύλη είναι οι εξής (Krzanowski & Hand, 2009):

Διάγραμμα 4.17: Ενδεικτική καμπύλη ROC. Η κόκκινη καμπύλη προκύπτει από τυχαίες ταξινομήσεις, η πράσινη από πλήρως ορθές, ενώ η μπλε αντιπροσωπεύει μία ρεαλιστική κατάσταση.



1. Παρομοιάζοντας την καμπύλη ROC με αυτή της καμπύλης Lorenz που χρησιμοποιείται στα χρηματοοικονομικά. Πιο συγκεκριμένα, η καμπύλη Lorenz παριστάνει την αθροιστική κατανομή του εισοδήματος ανάμεσα σε δύο χαρακτηριστικά. Η τυχαία καμπύλη του Διαγράμματος 4.17 θα σήμαινε δίκαιο μοίρασμα ανάμεσα στα δύο χαρακτηριστικά, ενώ η μπλε καμπύλη υποδεικνύει προοδευτικά μεγαλύτερες ανισότητες ανάμεσα στα δύο μεγέθη.

2. Σε συνέχεια της καμπύλης Lorenz, ο Ιταλός στατιστικός Corrado Gini το 1912, πρότεινε ως μέτρο της ανισότητας ανάμεσα στα δύο μεγέθη την σχετική διαφορά ανάμεσα στο εμβαδόν της καμπύλης Lorenz και της ευθείας που παριστάνει το δίκαιο μοίρασμα του εισοδήματος ανάμεσα σε δύο χαρακτηριστικά. Το παραπάνω είναι γνωστό ως συντελεστής gini, και στην περίπτωση της δυαδικής ταξινόμησης υπολογίζεται ως:

$$g = \frac{AUC - 0.5}{0.5}. \tag{4.4.10}$$

4.5 Εκπαίδευση

Για λόγους μείωσης του υπολογιστικού κόστους και βελτίωσης της ποιότητας των δεδομένων, σύμφωνα με όσα προαναφέρθηκαν για την ισοστάθμιση των δεδομένων στην αντίστοιχη υποενότητα 4.3, θα χρησιμοποιηθεί για την εκπαίδευση και την αξιολόγηση των ταξινομητών το σύνολο δεδομένων που προκύπτει μετά από την εφαρμογή της μεθόδου NearMiss-2. Εν συνεχεία, τα δεδομένα χωρίζονται σε δύο υποσύνολα:

- Υποσύνολο 1: Σύνολο εκπαίδευσης (75% των αρχικών παρατηρήσεων).
- Υποσύνολο 2: Σύνολο αξιολόγησης (25% των αρχικών παρατηρήσεων).

Τα υποσύνολα χωρίζονται κατά συστάδες, δηλαδή διατηρούν παρόμοιο ποσοστό παρατηρήσεων από κάθε κλάση, με αυτό του συνόλου δεδομένων.

Σε κάθε περίπτωση οι υπερπαράμετροι του κάθε ταξινομητή που χρησιμοποιηθήκαν, εκτιμήθηκαν πραγματοποιώντας στο σύνολο εκπαίδευσης διασταυρωμένη επικύρωση σε 5-πτυχές. Επίσης, τα δεδομένα εκπαίδευσης τυποποιήθηκαν προκειμένου να προέρχονται από κατανομή με μέση τιμή 0 και διασπορά 1:

$$\boldsymbol{x}^*_{\boldsymbol{i,train}} = \frac{\boldsymbol{x_{i,train}} - \overline{x_{i,train}}}{s_{i,train}}, \tag{4.5.1}$$

όπου $\boldsymbol{x_{i,train}}$ η τιμή της i-οστής επεξηγηματικής μεταβλητής, $\overline{x_{i,train}}$ η δειγματική της μέση τιμή και $s_{i,train}$ η δειγματική της τυπική απόκλιση αντίστοιχα.

Για την αξιολόγηση των ταξινομητών, εφαρμόζεται ο ίδιος μετασχηματισμός στο σύνολο αξιολόγησης, δηλαδή:

$$\boldsymbol{x}^*_{\boldsymbol{i,test}} = \frac{\boldsymbol{x_{i,test}} - \overline{x_{i,train}}}{s_{i,train}}. \tag{4.5.2}$$

4.5.1 Προετοιμασία αλγορίθμου συμβολικής παλινδρόμησης

Στον αλγόριθμο της συμβολικής παλινδρόμησης, ορίζεται ως αντικειμενική συνάρτηση η ορθότητα 4.4.1. Το αριθμητικό αποτέλεσμα της έκφρασης z που παράγουν τα συμβολικά δέντρα του γενετικού αλγορίθμου πρέπει να μεταφραστούν στο αν η παρατήρηση είναι ακτίνα γάμμα (κλάση 1) ή αδρόνιο (κλάση 0). Αυτό πραγματοποιείται με την βοήθεια της σιγμοειδούς συνάρτησης:

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{4.5.3}$$

Τιμές μεγαλύτερες του 0.5 ταξινομούνται ως κλάση 1, διαφορετικά ως κλάση 0. Στην συνέχεια δίνεται η δυνατότητα στον αλγόριθμο να αξιοποιήσει για την επίλυση του προβλήματος οι ακόλουθες συναρτήσεις:

- Διανυσματικής Πρόσθεσης.

- Βαθμωτού Πολλαπλασιασμού.
- Βαθμωτής Διαίρεσης.
- Ύψωσης στο τετράγωνο.
- Απόλυτης τιμής.
- Τόξου εφαπτομένης.

Τέλος, ορίζεται το μέγεθος του πληθυσμού να είναι 400, η πιθανότητα μετάλλαξης 0.1, η πιθανότητα διασταύρωσης 0.8 και ως κριτήριο τερματισμού, η μεταβολή της τιμής της αντικειμενικής ανάμεσα στους καλύτερους υποψηφίους δύο διαδοχικών γενεών να είναι μικρότερη από 10^{-4}.

4.5.2 Αποτελέσματα συμβολικής παλινδρόμησης

Η έκφραση z της σιγμοειδούς συνάρτησης 4.5.3, εκτιμήθηκε ότι είναι η ακόλουθη:

$$z = -1.48 + 0.77 \times \text{fM3Long} - 3.46 \times \text{fAlpha} - 2.53 \times \text{fLength}. \tag{4.5.4}$$

Παρατηρούμε ότι το τελικό μοντέλο αποτελείται μόνο από 3 εκ των 10 επεξηγηματικών μεταβλητών που δόθηκαν. Από τις μεταβλητές fLength, fWidth, fSize, fConc & fConc1, που είχαν υψηλή συσχέτιση μεταξύ τους (άνω του 0.7), μπήκε στο τελικό μοντέλο μόνο η fLength, η οποία παρουσίαζε την μεγαλύτερη γραμμική συσχέτιση με την μεταβλητή απόκρισης σε σχέση με τις υπόλοιπες 4. Επιπροσθέτως, τα αντίστοιχα θηκοδιαγράμματα και ιστογράμματα των επεξηγηματικών μεταβλητών που απαρτίζουν το τελικό μοντέλο, ήταν εκείνα στα οποία οι διαφορές ανάμεσα στους πληθυσμούς των δύο κλάσεων ήταν περισσότερο εμφανείς.

Το μοντέλο της συμβολικής ταξινόμησης 4.5.4, παρατηρούμε ότι είναι γραμμικό, και ισχύει ότι:

$$\frac{\sigma\left(z_{(i)}\right)}{1-\sigma\left(z_{(i)}\right)} = e^{z}e^{\beta_i} \tag{4.5.5}$$

$$= \frac{\sigma\left(z\right)}{1-\sigma\left(z\right)}e^{\beta_i} \tag{4.5.6}$$

$$= \rho \cdot e^{\beta_i}, \tag{4.5.7}$$

όπου $z_{(i)}$ η τιμή της Εξίσωσης 4.5.4, αν η μεταβλητή x_i αυξηθεί κατά μία μονάδα, β_i ο αντίστοιχος συντελεστής της και ρ ο λόγος σχετικών πιθανοτήτων (odds ratio) . Η παραπάνω σχέση, συμβάλει στην ερμηνεία του μοντέλου της συμβολικής παλινδρόμησης ως εξής:

- Αυξάνοντας κατά μία μονάδα της τιμή της κυβικής ρίζας της ροπής τρίτης τάξεως κατά μήκος του κύριου άξονα της έλλειψης (fM3Long), ο αντίστοιχος λόγος των σχετικών πιθανοτήτων, αυξάνεται κατά:

$$e^{0.77} - 1 = 2.16 - 1 = 116\,\%,$$

διατηρώντας σταθερές τις υπόλοιπες μεταβλήτες.

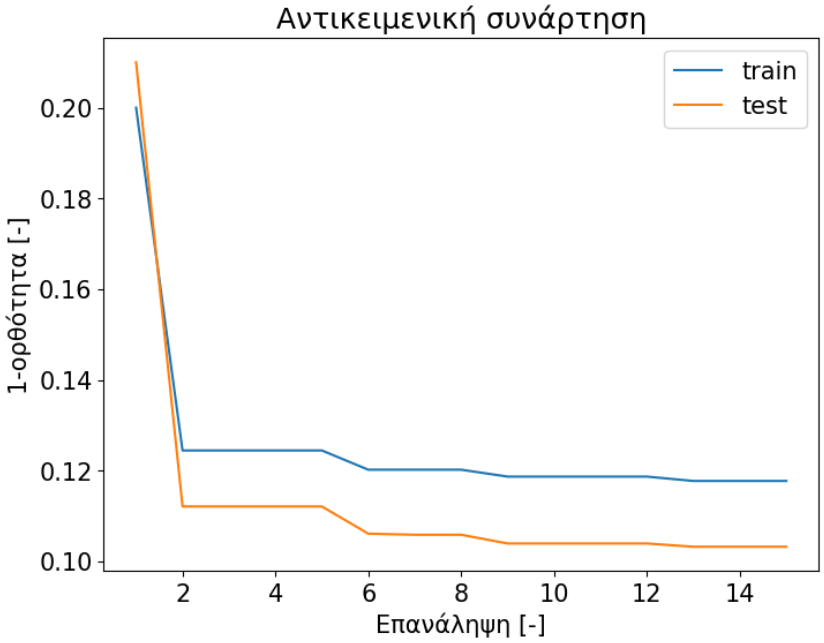
- Διατηρώντας σταθερή την τιμή των υπόλοιπων μεταβλητών και αυξάνοντας κατά μία μονάδα, την γωνία (fAlpha) που σχηματίζεται μεταξύ του κύριου άξονα της έλλειψης και του διανύσματος με αρχή το κέντρο της εικόνας και πέρας το κέντρο της έλλειψης, έχει ως αποτέλεσμα ο λόγος των σχετικών πιθανοτήτων, να είναι μειωμένος κατά:

$$e^{-0.36} - 1 = 0.03 - 1 = -97\,\%.$$

- Τέλος, η αύξηση του μήκους του κύριου άξονα της έλλειψης (fLength) κατά μία μονάδα, έχει ως αποτέλεσμα ο λόγος των σχετικών πιθανοτήτων να γίνει μικρότερος κατά:

$$1 - e^{-2.53} = 1 - 0.08 = 92\,\%,$$

διατηρώντας σταθερή την τιμή των υπόλοιπων μεταβλητών του μοντέλου.

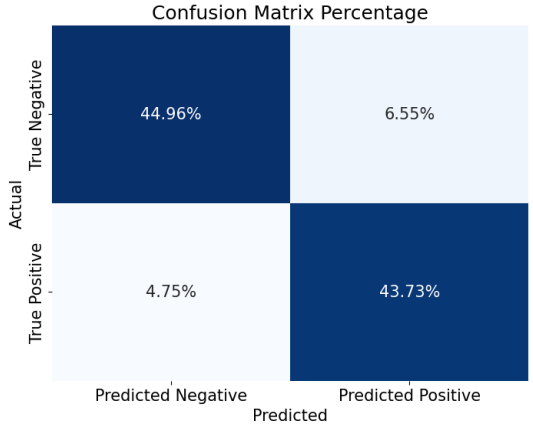


Διάγραμμα 4.18: Καμπύλη εκπαίδευσης του ταξινομητή Συμβολικής Παλινδρόμησης.

Στο Διάγραμμα 4.18 βλέπουμε ότι χρειάστηκαν μόλις 15 επαναλήψεις μέχρι να συγκλίνει

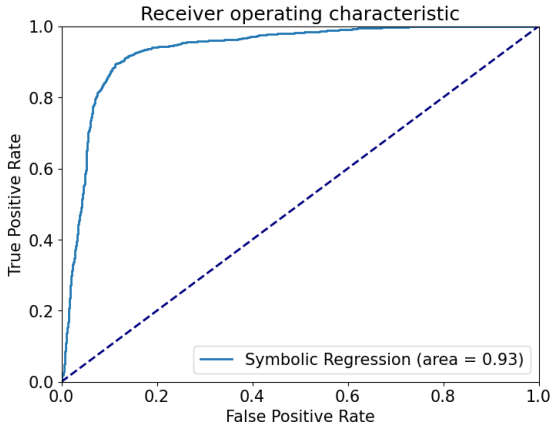
ο αλγόριθμος σε κάποια λύση. Η τιμή της αντικειμενικής συνάρτησης βλέπουμε ότι ακολουθεί παρόμοια τάση με το πέρας των επαναλήψεων, τόσο στο σύνολο εκπαίδευσης, όσο και στο σύνολο αξιολόγησης. Η τιμή της ορθότητας είναι παρόμοια ανάμεσα στα δύο σύνολα, επομένως το μοντέλο γενικεύεται με καλά αποτελέσματα σε καινούριες παρατηρήσεις.

Από τον πίνακα σύγχυσης (βλ. Διάγραμμα 4.19), βλέπουμε ότι η πλειοψηφία των παρατηρήσεων ταξινομείται ορθώς. Οι λανθασμένες ταξινομήσεις παριστάνουν περίπου το 10%. Στο εν λόγω ποσοστό, μεγαλύτερο μερίδιο κατέχουν οι λανθασμένα ταξινομημένες τιμές στην κλάση 1 (False Positives). Επομένως αναμένουμε το μοντέλο να έχει καλύτερο σκορ ανάκλησης (4.4.3), έναντι ακρίβειας (4.4.2).



Διάγραμμα 4.19: Πίνακας σύγχυσης του ταξινομητή Συμβολικής Παλινδρόμησης.

Στο Διάγραμμα 4.20 της χαρακτηριστικής καμπύλης ROC, φαίνεται ότι οι ταξινομήσεις γίνονται με τρόπο που απέχει σε μεγάλο βαθμό από την τυχαία ταξινόμηση κάθε παρατήρησης με πιθανότητα 0.5 σε κάθε κλάση. Επιπλέον, η τιμή του εμβαδού κάτω από την καμπύλη είναι 0.93, αρκετά κοντά στο ιδανικό που είναι το 1. Με άλλα λόγια η ταξινόμηση των παρατηρήσεων γίνεται με μεγάλη πιθανότητα προς την σωστή κλάση.



Διάγραμμα 4.20: Χαρακτηριστική καμπύλη ROC Συμβολικής Παλινδρόμησης.

4.5.3 Αποτελέσματα υπόλοιπων ταξινομητών

Για λόγους σύγκρισης, η επίδοση του ταξινομητή της συμβολικής παλινδρόμησης συγκρίνεται με αυτή άλλων ταξινομητών που χρησιμοποιούνται συχνά σε προβλήματα μηχανικής μάθησης. Πιο συγκεκριμένα, οι ταξινομητές που θα εξεταστούν σε αυτή την υποενότητα είναι (Kuncheva, 2014):

- Γραμμικός ταξινομητής που προκύπτει έπειτα από ανάλυση κύριων συνιστωσών.
- Ενδυναμωμένο δέντρο απόφασης μέγιστου ύψους 2.
- Νευρωνικό δίκτυο Perceptron με 1 & 2 κρυφά στρώματα.

Τα αποτελέσματα στο σύνολο αξιολόγησης παρουσιάζονται στον ακόλουθο πίνακα:

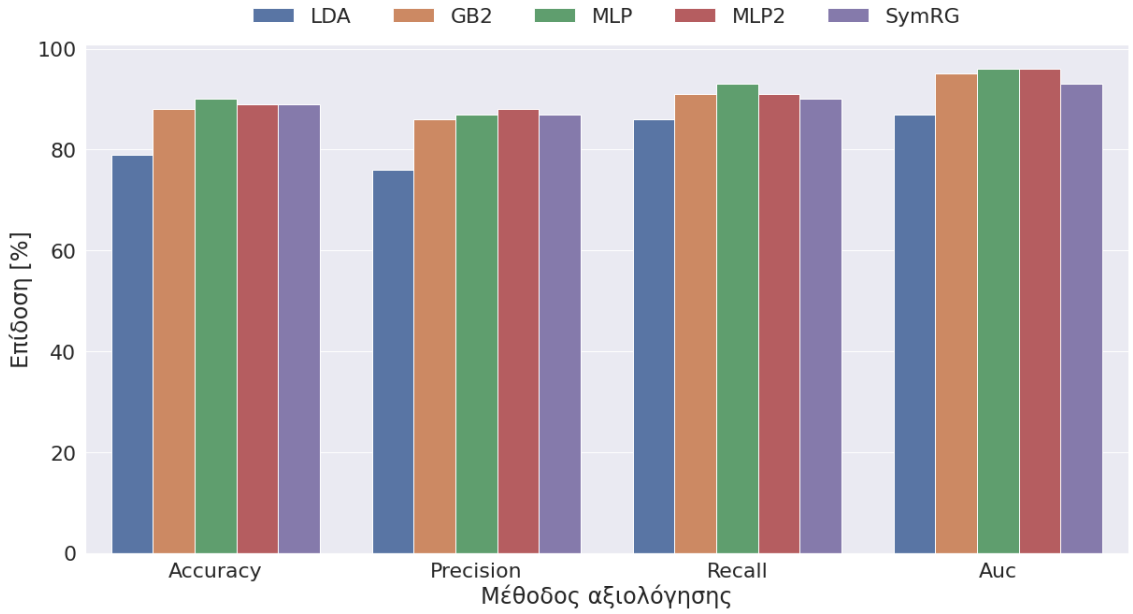
	Ορθότητα	Ακρίβεια	Ανάκληση	Εμβαδόν καμπύλης ROC
1. Γραμμικός Ταξινομητής	0.79	0.76	0.86	0.87
2. Ενδυναμωμένο Δέντρο Απόφασης	0.88	0.86	0.91	0.95
3. Νευρωνικό Δίκτυο Perceptron (1 κρυφό στρώμα)	0.90	0.87	0.93	0.96
4. Νευρωνικό Δίκτυο Perceptron (2 κρυφά στρώματα)	0.89	0.88	0.91	0.96
5. Συμβολική παλινδρόμηση	0.89	0.87	0.90	0.93

Πίνακας 4: Συγκεντρωτικός πίνακας αξιολόγησης ταξινομητών στο σύνολο εκπαίδευσης.

Ο γραμμικός ταξινομητής έχει συνολικά την χειρότερη επίδοση. Προφανώς τα μοτίβα που υπάρχουν πίσω από τα δεδομένα δεν μπορούν να αναγνωριστούν με γραμμικές μετασχηματίσεις των αρχικών δεδομένων. Παρόλα αυτά ο γραμμικός ταξινομητής αποτελεί την απλούστερη επιλογή σε σχέση με όλους τους υπόλοιπους ως προς τον τρόπο λειτουργίας του, και έχει καταφέρει ως ένα βαθμό να ταξινομήσει ορθά ένα σημαντικό ποσοστό των παρατηρήσεων (79%). Η ανάκληση είναι αρκετά μεγαλύτερη από την ακρίβεια, επομένως ο εν λόγω ταξινομητής πραγματοποιεί περισσότερες FN ταξινομήσεις από FP.

Την καλύτερη επίδοση φαίνεται να έχουν τα νευρωνικά δίκτυα. Πιο συγκεκριμένα, το νευρωνικό δίκτυο με ένα κρυφό στρώμα έχει καλύτερο σκορ αξιολόγησης από αυτό με τα δύο κρυφά στρώματα, αν και το τελευταίο λόγω αυξημένης πολυπλοκότητας έχει την δυνατότητα να περιγράφει πιο περιπλοκές σχέσεις ανάμεσα στα δεδομένα. Αυτό αποτελεί ένδειξη ότι ίσως αρχίζει να γίνεται γρηγορότερα υπερεκπαίδευση με αποτέλεσμα να πέφτει η επίδοση στο σύνολο εκπαίδευσης. Επιπλέον παρατηρούμε και εδώ ότι το σκορ της ανάκλησης είναι μεγαλύτερο από τη ακρίβεια, άρα οι περισσότερες λάθος ταξινομήσεις είναι False Negatives.

Ο Πίνακας 4 περιλαμβάνει και τα αποτελέσματα από την Συμβολική Παλινδρόμηση. Πιο συγκεκριμένα, σε σχέση με τα Νευρωνικά Δίκτυα και το Ενδυναμωμένο Δέντρο απόφασης, τα σκορ αξιολόγησης της ορθότητα, ακρίβειας, είναι σχεδόν ίδια ή αποκλίνουν το πολύ κατά 1%. Η ικανότητα της ανάκλησης φαίνεται να είναι λίγο χειρότερη, επηρεάζοντας ταυτόχρονα και το εμβαδόν της καμπύλης ROC. Παρόλα αυτά συνολικά η επίδοση είναι πολύ παρόμοια και έχουμε το πλεονέκτημα ότι γνωρίζουμε τον κλειστό τύπο του μοντέλου, ο οποίος αποτελείται από μόλις 3 επεξηγηματικές μεταβλητές, έναντι των 10 που απαιτούνται από τους υπόλοιπους ταξινομητές.



Διάγραμμα 4.21: Ραβδόγραμμα επίδοσης ταξινομητών με διάφορες μεθόδους αξιολόγησης.

Το Διάγραμμα 4.21 παρουσιάζει τα αποτελέσματα του Πίνακα 4 σε μορφή ραβδογράμματος, και συμβολίζεται ως:

- LDA: Ο γραμμικός ταξινομητής.
- MLP1: Νευρωνικό δίκτυο Perceptron με 1 κρυφό στρώμα.
- MLP2: Νευρωνικό δίκτυο Perceptron με 2 κρυφά στρώματα.
- GB2: Ενδυναμωμένο δέντρο απόφασης μέγιστου βάθους 2.
- SymRG: Ταξινομητής συμβολικής παλινδρόμησης (βλ υποενότητα 4.5.2).

4.6 Επίλογος

Στο παρόν Κεφάλαιο, η μέθοδος της συμβολικής παλινδρόμησης καλείται να επιλύσει ένα πραγματικό πρόβλημα ταξινόμησης, στο οποίο η λύση δεν είναι γνωστή. Αρχικά τα δεδομένα

του προβλήματος αναλύονται με διαγράμματα, και υπολογίζονται οι μεταξύ τους συσχετίσεις. Εν συνεχεία γίνεται προβολή των δεδομένων σε μικρότερη διάσταση με την μέθοδο PCA. Το πρόβλημα που εντοπίζεται είναι η διαφορά στην πληθικότητα ανάμεσα στις δύο κλάσεις, το οποίο διορθώνεται εφαρμόζοντας την μέθοδο NearMiss-2. Αυτό έχει ως αποτέλεσμα να βελτιώνεται, τουλάχιστον γραφικά, η διαχωρισιμότητα μεταξύ των δύο κλάσεων. Στο τέλος προσαρμόζεται το τελικό μοντέλο της Συμβολικής παλινδρόμησης, το οποίο όχι μόνο παρουσιάζει καλή επίδοση, συγκρίσιμη με αυτή άλλων ταξινομητών, αλλά το μοντέλο που προκύπτει είναι αρκετά απλό με εμφανείς την συναρτησιακή εξάρτηση που περιγράφει την σχέση ανάμεσα στην μεταβλητή απόκρισης και τις επεξηγηματικές μεταβλητές.

5 Συμπεράσματα

Σκοπός της διπλωματικής εργασίας ήταν να επιλυθεί το πρόβλημα της συμβολικής παλινδρόμησης με χρήση γενετικών αλγορίθμων. Η υλοποίηση του αλγορίθμου στο τρίτο κεφαλαίου, αποδείχθηκε αποτελεσματική, καθώς στο πρόβλημα παλινδρόμησης του ίδιου κεφαλαίου, ανακτήθηκε το αρχικό μοντέλο με τους συντελεστές του. Επιπλέον, η εκπαίδευση του μοντέλου ταξινόμησης στο τελευταίο κεφάλαιο, επαλήθευσε την ικανότητα προσαρμογής του γενετικού αλγορίθμου σε ένα πραγματικό πρόβλημα ταξινόμησης.

Τα θεωρητικά πλεονεκτήματα της συμβολικής παλινδρόμησης, επιβεβαιώνονται στο Κεφάλαιο 4. Συγκεκριμένα, ο αλγόριθμος κατασκεύασε ένα εξηγήσιμο μοντέλο για την ταξινόμηση των παρατηρήσεων. Από την εφαρμογή του γενετικού αλγορίθμου στο εν λόγω πρόβλημα, αναδεικνύεται μία ακόμα ικανότητα του αλγορίθμου, αυτή της επιλογής μεταβλητών. Στο τελικό μοντέλο εισέρχονται μόνο οι μεταβλητές που προσφέρουν ουσιαστική βελτίωση στην τιμή της αντικειμενικής συνάρτησης. Ταυτόχρονα, η επίδοση υπολογισμένη με τις μεθόδους αξιολόγησης που παρουσιάστηκαν, παραμένει συγκρίσιμη με αυτή του νευρωνικού δικτύου και των ενδυναμωμένων δέντρων. Φυσικά οι τελευταίοι ταξινομητές υστερούν στο κομμάτι της απλότητας και της επεξηγηματικότητας, έναντι του μοντέλου της συμβολικής παλινδρόμησης.

Η παραπάνω μέθοδος παρουσιάζει και ορισμένα στοιχεία τα οποία χρήζουν βελτίωσης. Πρώτα απ΄ όλα αν δεν σταματήσει η εκπαίδευση στην κατάλληλη χρονική στιγμή, το μοντέλο γίνεται εξαιρετικά περίπλοκο με μικρές βελτιώσεις στην τιμή της αντικειμενικής συνάρτησης. Αυτό οδηγεί τον αλγόριθμο στην αποστήθιση των δεδομένων εκπαίδευσης, παρουσιάζοντας πτώση στην επίδοση του συνόλου αξιολόγησης, με συνέπεια το τελικό μοντέλο να χάνει το πλεονέκτημα της επεξηγισιμότητας. Τέλος, απαιτείται από τον χρήστη να καθορίσει ορισμένες υπερπαραμέτρους, οι οποίες συνήθως επιλέγονται εμπειρικά.

Η σημαντικότερη διαφοροποίηση της υλοποίησης που προτείνει το Κεφάλαιο 3, σε σχέση με τους αλγορίθμους που παρουσιάστηκαν στο Κεφάλαιο 2, αποτελεί η απουσία κάποιας μεθόδου απλοποίησης των συμβολικών δέντρων. Αυτό οδηγεί σε ορισμένες περιπτώσεις να επαναλαμβάνονται οι ίδιοι όροι στο μοντέλο, ή το δέντρο να έχει μεγαλύτερη πολυπλοκότητα σε σχέση με την έκφραση που επιχειρεί να κωδικοποιήσει, όπως φάνηκε και στην αντίστοιχη εφαρμογή. Βέβαια η απλοποίηση των εν λόγω εκφράσεων δεν αποτελεί εύκολη διαδικασία, ειδικά όταν οι διαθέσιμοι τελεστές καθορίζονται από τον χρήστη.

Οι στοχαστικοί αλγόριθμοι βελτιστοποίησης, χαρακτηρίζονται από το αυξημένο υπολογιστικό κόστος. Ο γενετικός αλγόριθμος μπορεί να τροποποιηθεί ώστε να τρέχει παράλληλα σε όλους τους πυρήνες, είτε χωρίζοντας τον πληθυσμό σε υποπληθυσμούς, είτε αναθέτοντας σε κάθε πυρήνα επεξεργαστή την διασταύρωση, μετάλλαξη και βελτιστοποίηση των σταθερών, ορισμένων από τους επιλεχθέντες υποψηφίους για διασταύρωση. Ο αλγόριθμος συμβολικής παλινδρόμησης της εργασίας, αντί αυτού εστιάζει στην χρήση διανυσματοποιημένων συναρτήσεων (vectorization), οι οποίες επιταχύνουν έως και 1000 φορές τους υπο-

λογισμούς. Αυτό οφείλεται στην αρχιτεκτονική των σύγχρονων επεξεργαστών, οι οποίοι μπορούν να εκτελέσουν στοιχειώδης πράξεις ταυτόχρονα σε στοιχεία του διανύσματος.

Ως βελτίωση στον χρόνο εκτέλεσης, μπορεί να αποτελέσει η αξιοποίηση των πυρήνων που διαθέτουν οι κάρτες γραφικών. Πιο συγκεκριμένα, οι περισσότεροι πυρήνες της κάρτας γραφικών, έναντι του επεξεργαστή, αν και έχουν μικρότερη συχνότητα, μπορούν να επιταχύνουν έως και 200 φορές την ταχύτητα εκτέλεσης, όταν τα δεδομένα είναι πολύ μεγάλα. Γι΄ αυτό άλλωστε τα περισσότερα πακέτα μηχανικής μάθησης, όπως είναι το PyTorch[3], cuPy[4] και Tensorflow[5] στην Python, προσφέρουν την παραπάνω δυνατότητα.

Κλείνοντας, αξίζει να διερευνηθεί η χρήση πιο περίπλοκων αντικειμενικών συναρτήσεων. Συγκεκριμένα στον γενετικό αλγόριθμο, φαίνεται να ταιριάζει περισσότερο η χρήση της ύστερης κατανομής 2.3.3 στον γενετικό αλγόριθμο, για τους εξής λόγους:

1. Ο γενετικός αλγόριθμος στηρίζεται ήδη σε πληθυσμό, επομένως οι υπερπαράμετροι α_o και β_o της πρότερης κατανομής 2.3.11 μπορούν να εκτιμηθούν σε κάθε επανάληψη του αλγορίθμου.

2. Η συχνότητα εμφάνισης του κάθε τελεστή n_o υπολογίζεται ήδη προκειμένου να υπολογισθεί η τιμή της εντροπίας 3.2.6.

Επομένως μπορεί να μεριμνήσει η ύστερη κατανομή, ως αντικειμενική συνάρτηση, στην διατήρηση απλών συμβολικών εκφράσεων, χωρίς να έχει ληφθεί κάποιο άλλο μέτρο ποινής.

[3] https://pytorch.org/
[4] https://cupy.dev/
[5] https://www.tensorflow.org/

Βιβλιογραφικές αναφορές

Ελληνικές

Καρώνη, Χ. & Οικονόμου, Π. (2020). *Στατιστικά Μοντέλα Παλινδρόμησης (2η έκδοση)*. Εκδόσεις Συμεών: Αθήνα.

Κοκολάκης, Γ. & Φουσκάκης, Δ. (2009). *Στατιστική Θεωρία & Εφαρμογές*. Εκδόσεις Συμεών: Αθήνα.

Λουλάκης, Μ. (2019). *Στοχαστικές Διαδικασίες*, (ππ. 84–86). Κάλλιπος: Αθήνα.

Φουσκάκης, Δ. (2013). *Ανάλυση Δεδομένων με Χρήση της R*. Εκδόσεις Τσότρας: Αθήνα.

Διεθνής

Bock, R. (2007). MAGIC Gamma Telescope. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52C8B.

Cranmer, M. (2023). Interpretable machine learning for science with pysr and symbolicregression.jl.

ESA (2018). Why do we observe gamma rays? https://www.esa.int/Science_Exploration/Space_Science/Integral/Why_do_we_observe_gamma_rays. Accessed: 2024-05-05.

Esposito, W. R. & Floudas, C. A. (2009). *Gauss–Newton method: Least squares, relation to Newton's method*, (pp. 1129–1134). Springer US: Boston, MA.

Goodrich, M. T., Tamassia, R., & Goldwasser, M. H. (2016). *Data Structures and Algorithms in Java, 6th Edition*, (pp. 308–350). WILEY: Hoboken, NJ.

Goswami, R. D., Chakraborty, S., & Misra, B. (2023). *Variants of Genetic Algorithms and Their Applications*, (pp. 1–20). Springer Nature Singapore: Singapore.

Guimerà, R., Reichardt, I., Aguilar-Mogas, A., Massucci, F. A., Miranda, M., Pallarès, J., & Sales-Pardo, M. (2020). A bayesian machine scientist to aid in the solution of challenging scientific problems. *Science Advances*, 6(5).

Hornby, G. (2006). Alps: The age-layered population structure for reducing the problem of premature convergence. volume 1.

Häggström, O. (2002). *Finite Markov Chains and Algorithmic Applications*. London Mathematical Society Student Texts. Cambridge University Press.

Krzanowski, W. J. & Hand, D. J. (2009). *ROC Curves for Continuous Data*. Chapman and Hall/CRC.

Kuncheva, L. (2014). *Base Classifiers*, chapter 2, (pp. 49–93). John Wiley and Sons, Ltd.

Moré, J. J. (1978). The levenberg-marquardt algorithm: Implementation and theory. In G. A. Watson (Ed.), *Numerical Analysis* (pp. 105–116). Berlin, Heidelberg: Springer Berlin Heidelberg.

Schmidt, M. & Lipson, H. (2011). *Age-Fitness Pareto Optimization*, (pp. 129–146). Springer New York: New York, NY.

Udrescu, S.-M. & Tegmark, M. (2020). Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16), eaay2631.

Uy, N. Q., Hoai, N. X., O'Neill, M., McKay, R. I., & Galván-López, E. (2011). Semantically-based crossover in genetic programming: application to real-valued symbolic regression. *Genetic Programming and Evolvable Machines*, 12(2), 91–119.

Zhang, J. & Mani, I. (2003). KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets*.