ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Ο ΓΕΝΕΤΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ ΣΤΟ ΠΡΟΒΛΗΜΑ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

Διπλωματικη Εργασια

ΚΥΒΕΛΗ-ΧΡΙΣΤΙΝΑ ΤΣΙΩΛΗ

Επιβλέπων: Δρ. Φουσκάκης Δημήτρης
Αναπληρωτής Καθηγητής Ε.Μ.Π

Αθήνα, Μάρτιος 2020

(Υπογραφή)

.....

Κυβελη-Χριστινα Τσιωλη
Διπλωματούχος Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Ε.Μ.Π.

Ευχαριστίες

Θα ήθελα να εκφράσω τις ειλικρινείς ευχαριστίες μου στον επιβλέποντα καθηγητή μου, Δρ. Δημήτρη Φουσκάκη, για την δυνατότητα ανάληψης της παρούσας διπλωματικής εργασίας αλλά κυρίως για την υποστήριξη και την καθοδήγηση που μου παρείχε κατά την εκπόνησή της.

Στη συνέχεια, θα ήθελα να ευχαριστήσω από καρδιάς τους φίλους με τους οποίους μοιράστηκα κοινά όνειρα και στόχους, και ιδιαίτερα τη Γεωργία, τη Ρωξάνη και τον Κωνσταντίνο, που αποτέλεσαν αναπόσπαστο κομμάτι της φοιτητικής μου ζωής.

Η εργασία αφιερώνεται στην οικογένεια μου, για την απεριόριστη αγάπη, εμπιστοσύνη και υποστήριξη που δείχνουν στο πρόσωπό μου όλα αυτά τα χρόνια.

Περίληψη

Η ανάγκη για επίλυση προβλημάτων βελτιστοποίησης αναδείχθηκε με την αδυναμία των κλασσικών αναλυτικών μεθόδων να αντιμετωπίσουν πολύπλοκα συστήματα. Οι Ευριστικές μέθοδοι, αποτελούν ενεργό πεδίο της επιστημονικής έρευνας και αντιμετωπίζουν τέτοιου είδους προβλήματα με έναν εναλλακτικό τρόπο. Εκτελώντας επαναληπτικά μια περιορισμένη αναζήτηση στο χώρο των δυνατών λύσεων του προβλήματος, οι ευριστικές μέθοδοι είναι δυνατό να βρουν μια αποδεδειγμένα καλή λύση για το εκάστοτε πρόβλημα, εκμεταλλευόμενες την πληροφορία που τους παρέχεται σε κάθε βήμα. Μια τέτοια μέθοδος είναι και ο γενετικός αλγόριθμος.

Ο γενετικός αλγόριθμος, συγκεκριμένα, βασίζεται στην αρχή της εξέλιξης των ειδών και μιμείται τις διαδικασίες της αναπαραγωγής, της φυσικής επιλογής και της μετάλλαξης ώστε να αναζητήσει τη βέλτιστη λύση. Οι γενετικοί αλγόριθμοι επιχειρούν να προσομοιώσουν σε υπολογιστικό περιβάλλον τους μηχανισμούς της βιολογικής εξέλιξης, με τον ίδιο τρόπο που η τεχνητή νοημοσύνη επιχειρεί να προσομοιώσει τις νοητικές διεργασίες που συμβαίνουν στον ανθρώπινο εγκέφαλο.

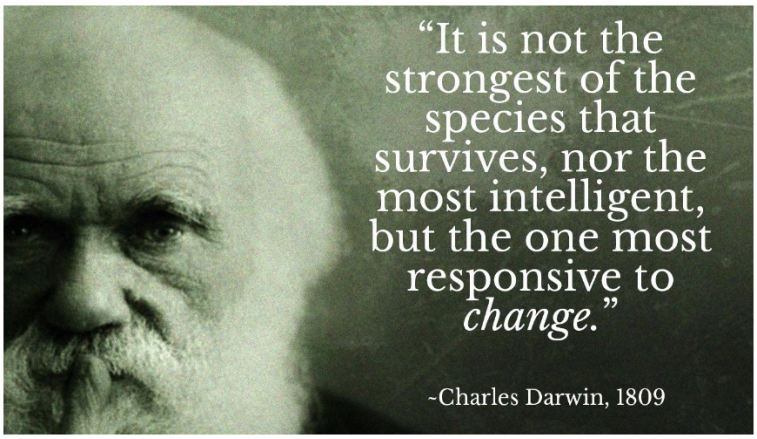
Η παρούσα διπλωματική εργασία διαρθρώνεται σε έξι κεφάλαια. Το πρώτο κεφάλαιο συνιστά μια εισαγωγή στους γενετικούς αλγορίθμους παρέχοντας στον αναγνώστη τις αρχές στις οποίες βασίζεται, και ταυτόχρονα παρουσιάζει το ερευνητικό πεδίο στο οποίο εφαρμόζονται. Στο δεύτερο κεφάλαιο, αναλύεται η δομή ενός γενετικού αλγορίθμου καθώς και τα πλεονεκτήματα και μειονεκτήματα που παρουσιάζει έναντι άλλων κλασσικών μεθόδων. Στη συνέχεια, το τρίτο κεφάλαιο εισάγει τον αναγνώστη στο πρόβλημα επιλογής μεταβλητών στην πολλαπλή γραμμική παλινδρόμηση, ενώ στο τέταρτο κεφάλαιο παρουσιάζεται διεξοδικά η μοντελοποίηση του γενετικού αλγορίθμου έτσι ώστε να επιλύει το συγκεκριμένο πρόβλημα. Το πέμπτο κεφάλαιο εξετάζει την απόδοση του γενετικού αλγορίθμου που σχεδιάστηκε σε προσομοιωμένα αλλά και πραγματικά δεδομένα και παραθέτει τα συμπεράσματα που προκύπτουν από την υλοποίηση. Τέλος, το έκτο κεφάλαιο αποτελεί μια πρόταση για μελλοντική επέκταση της διπλωματικής εργασίας ως συνεισφορά στο ευρύτερο επιστημονικό πεδίο της Στατιστικής και της Πληροφορικής. Η υλοποίηση του γενετικού αλγορίθμου έγινε με χρήση της γλώσσας προγραμματισμού R.

Περιεχόμενα

Κεφάλαιο 1

ΕΙΣΑΓΩΓΗ ΣΤΟΝ ΓΕΝΕΤΙΚΟ ΑΛΓΟΡΙΘΜΟ

1.1 Εισαγωγικά



Ο γενετικός αλγόριθμος είναι μια μέθοδος βελτιστοποίησης που βασίζεται στη θεωρία εξέλιξης των ειδών και της φυσικής επιλογής, όπως αυτή διατυπώθηκε και αναπτύχθηκε από τον Δαρβίνο κατά τον 19ο αιώνα. Σύμφωνα με τη θεωρία αυτή, οι οργανισμοί που είναι περισσότερο προσαρμοσμένοι στο περιβάλλον τους επιβιώνουν και αναπαράγονται περισσότερο από κάποιους άλλους. Τα κυριότερα σημεία αυτής της θεωρίας είναι τα εξής:

- Δεν υπάρχει αντικειμενική βάση διαχωρισμού των ζωντανών οργανισμών σε ανώτερους και κατώτερους. Σε κάθε βιολογικό είδος, μερικά άτομα αφήνουν περισσότερους απογόνους σε σύγκριση με τα υπόλοιπα και κατά συνέπεια τα χαρακτηριστικά τους κληροδοτούνται στην επόμενη γενιά. Οι δυσκολίες και οι αντιξοότητες της επιβίωσης των οργανισμών καθορίζουν ποιοι από αυτούς θα κατορθώσουν να ζήσουν και να πολλαπλασιαστούν. Με την αλλαγή του περιβάλλοντος, αλλάζουν και τα χαρακτηριστικά των οργανισμών ώστε να προσαρμοστούν στις νέες συνθήκες διαβίωσης και να μπορέσουν να εξασφαλίσουν την επιβίωσή τους.

- Αλλαγή στα χαρακτηριστικά των ατόμων, σημαίνει αλλαγή στα χρωμοσώματά τους (*chromosomes*), τα οποία συνιστούν μια οργανωμένη δομή που περιλαμβάνει τις γενετικές πληροφορίες που καθορίζουν τη βιολογική ανάπτυξη του οργανισμού. Τα χρωμοσώματα αποτελούνται με τη σειρά τους από γονίδια (*genes*) και το σύνολο της γενετικής πληροφορίας που είναι κωδικοποιημένη στα γονίδια ονομάζεται γονότυπος (*genotype*). Προκειμένου να δημιουργηθεί ένας νέος οργανισμός, χρειάζεται πρώτα να αποκωδικοποιηθούν τα χρωμοσώματα, έτσι ώστε να εκδηλωθούν όλα τα "ορατά" χαρακτηριστικά που συνιστούν τον φαινότυπο (*phenotype*).

- Η εξέλιξη συμπεριλαμβάνει τις λειτουργίες τόσο της αναπαραγωγής (*reproduction*), όσο και της μετάλλαξης (*mutation*). Η μετάλλαξη είναι η αλλαγή της δομής των χρωμοσωμάτων που συμβαίνει με τυχαίο τρόπο και επιφέρει αλλαγή σε κάποιο χαρακτηριστικό του οργανισμού. Με άλλα λόγια, μετάλλαξη είναι οποιαδήποτε μεταβολή μπορεί να συμβεί στο γενετικό υλικό ενός οργανισμού. Ειδικότερα, η μετάλλαξη είναι τυχαία, με την έννοια ότι η πιθανότητα να εμφανιστεί μια αλλαγή δε σχετίζεται με το βαθμό χρησιμότητάς της. Οι ευνοϊκές μεταλλάξεις στο γονότυπο ενός οργανισμού είναι σπανιότερες και ευνοούνται από τη φυσική επιλογή.

- Μέσω της αναπαραγωγής δημιουργείται ένας νέος οργανισμός με χρωμοσώματα που αποτελούνται από γονίδια προερχόμενα εξίσου από τον πατέρα όσο και από τη μητέρα. Κατά συνέπεια, για κάθε ένα από τα χαρακτηριστικά του, ο νέος οργανισμός έχει πάρει ένα γονίδιο από κάθε γονέα. Γονίδια που διεκδικούν την ίδια θέση σε ένα χρωμόσωμα λέγονται αλληλόμορφα (*allele*), ενώ το γονίδιο που τελικά επικρατεί και καθορίζει το χαρακτηριστικό και άρα τον φαινότυπο λέγεται κυρίαρχο ή επικρατές (*dominant*) και το άλλο υπολειπόμενο (*recessive*).

Ο γενετικός αλγόριθμος, βασιζόμενος στις παραπάνω παραδοχές, αποτελεί μια ειδική κατηγορία των γνωστών και ως "εξελικτικών" αλγορίθμων, οι οποίοι

χαρακτηρίζονται από την ύπαρξη πληθυσμού από άτομα που εκτίθενται στο περιβάλλον τους, μερικά εκ των οποίων ευνοούνται από τη φυσική επιλογή (τα πιο ικανά) και τελικά μεγιστοποιείται η ολική καταλληλότητα των ατόμων που επιβιώνουν. Ο μηχανισμός της φυσικής επιλογής παρακίνησε τον *John Holland*, πρωτοπόρο των γενετικών αλγορίθμων, στις αρχές της δεκαετίας του 1970 στη διερεύνηση της εφαρμογής των ιδεών αυτών για την επίλυση πολύπλοκων προβλημάτων με τη χρήση υπολογιστών. Αποτέλεσμα της ενασχόλησης αυτής του *Holland* ήταν η εδραίωση των γενετικών αλγορίθμων ως μια καινούργια τεχνική αναζήτησης και βελτιστοποίησης, βασισμένη στη μίμηση των μηχανισμών της φύσης.

1.2 Στοχαστική Βελτιστοποίηση

Η στοχαστική βελτιστοποίηση (Stochastic Optimisation) αναφέρεται σε ένα σύνολο μεθόδων που χρησιμοποιούνται για την ελαχιστοποίηση ή τη μεγιστοποίηση μιας αντικειμενικής συνάρτησης υπό την ύπαρξη τυχαιότητας, μέθοδοι δηλαδή στις οποίες παράγονται και χρησιμοποιούνται τυχαίες μεταβλητές. Τις τελευταίες δεκαετίες, αυτές οι μέθοδοι έχουν γίνει απαραίτητα εργαλεία για την επίλυση προβλημάτων στη μηχανική, στην επιστήμη των υπολογιστών, στις επιχειρήσεις αλλά και στη στατιστική. Η λέξη "στοχαστικός", δηλώνει την ενσωμάτωση της τύχης με την έννοια της πιθανότητας. Οι αλγόριθμοι στοχαστικής βελτιστοποίησης μπορούν να χειριστούν στοχαστικούς περιορισμούς ή να βελτιστοποιήσουν μια αντικειμενική συνάρτηση με τυχαίες μεταβλητές, επιδιώκοντας την εύρεση της βέλτιστης λύσης σε όσο το δυνατόν μικρότερο υπολογιστικό χρόνο. Λόγω της ευρείας χρήσης αλλά και ανάπτυξης των ηλεκτρονικών υπολογιστών, η αριθμητική επίλυση προβλημάτων βελτιστοποίησης έχει μελετηθεί σε βάθος, και η στοχαστική βελτιστοποίηση αποτελεί μια μέθοδο για την αντιμετώπισή τους.
Πιο αναλυτικά, στη στοχαστική βελτιστοποίηση η αναζήτηση για τη βέλτιστη λύση περιλαμβάνει την τυχαιότητα με έναν κατασκευαστικό, συστηματικό τρόπο. Συγκεκριμένα, αν με S συμβολίσουμε ένα πεπερασμένο σύνολο λύσεων, τότε επιδιώκουμε να μεγιστοποιήσουμε ή να ελαχιστοποιήσουμε την αντικειμενική συνάρτηση $f : S \to \mathbb{R}$. Στην περίπτωση της ελαχιστοποίησης με την οποία ασχολούμαστε στην παρούσα διπλωματική εργασία, το πρόβλημα είναι ισοδύναμο με την εύρεση ενός σχηματισμού (configuration) x_{opt} που ικανοποιεί τη σχέση : $f(x_{opt}) \leq f(x)$ για όλα τα x ϵ S.
Η μέθοδος βελτιστοποίησης που θα εξετάσουμε εδώ είναι διακριτή ως προς τον χρόνο, με την έννοια ότι ο αρχικός σχηματισμός x_0 του S που επιλέγεται αποτελεί την τρέχουσα τιμή x_t τη χρονική στιγμή $t = 0$, και ο αλγόριθ-

μος επαναληπτικά μεταβαίνει από την κατάσταση x_t στην x_{t+1}. Επίσης, πολλές από τις μεθόδους στοχαστικής βελτιστοποίησης, όπως και η μέθοδος που θα κατασκευάσουμε στην εν λόγω διπλωματική εργασία, θα βασίζονται στη διερεύνηση μιας "γειτονιάς" του χώρου αναζήτησης που θα αποφασίζει το πού θα κατευθυνθεί η αναζήτηση της λύσης στην επόμενη επανάληψη, και αυτό απαιτεί ένα μέτρο απόστασης που προσδιορίζει με μοναδικό τρόπο όλους τους "γείτονες" της τρέχουσας κατάστασης. Καθώς η διάσταση του χώρου του προβλήματος S αυξάνει, η επίλυση γίνεται πιο απαιτητική και χρειάζεται περισσότερος χρόνος για να εντοπιστεί η βέλτιστη λύση, ή ακόμη μια λύση που είναι κοντά στη βέλτιστη. Μια επιπρόσθετη και συχνή δυσκολία σε τέτοιου είδους προβλήματα υπάρχει όταν η αντικειμενική συνάρτηση, δηλαδή η συνάρτηση προς βελτιστοποίηση, εμφανίζει αρκετά το πλήθος τοπικά ακρότατα. Ενδεικτικά, ο γνωστός αλγόριθμος της τοπικής αναζήτησης (local search) ο οποίος σε κάθε επανάληψη εκτελεί αποκλειστικά και μόνο βήματα που αυξάνουν/ελαττώνουν την τιμή της αντικειμενικής συνάρτησης, ανάλογα με το αν επιλύουμε πρόβλημα μεγιστοποίησης ή ελαχιστοποίησης αντίστοιχα, δεν θα έχει καλεί απόδοση διότι είναι πιθανό η αναζήτηση να εγκλωβιστεί σε τοπικό ακρότατο και όχι ολικό. Επίσης, είναι φανερό ότι ο αλγόριθμος τοπικής αναζήτησης εμφανίζει ισχυρή εξάρτηση από την αρχική τιμή x_0 και επομένως το τοπικό ακρότατο στο οποίο θα καταλήξει θα εξαρτάται από την τιμή εκκίνησης του αλγορίθμου, παράλληλα όμως δεν υπάρχει άνω όριο που να καθορίζει τον υπολογιστικό χρόνο εκτέλεσης.

Προς αποφυγή των παραπάνω, μπορούν να χρησιμοποιηθούν οι στοχαστικοί αλγόριθμοι βελτιστοποίησης, βασιζόμενοι σε κάποιες αρχές που συνιστούν τροποποιήσεις στη διαδικασία εύρεσης του ολικού ακροτάτου συγκριτικά με τον αλγόριθμο τοπικής αναζήτησης. Πιο συγκεκριμένα, προβλήματα βελτιστοποίησης μπορούν να αντιμετωπιστούν με χρήση των αλγορίθμων στοχαστικής βελτιστοποίησης οι οποίοι εκτελούνται για ένα μεγάλο αριθμό αρχικών σχηματισμών, έστω M, με συνέπεια την αύξηση του υπολογιστικού χρόνου. Σε κάθε μία επανάληψη, γίνεται αξιοποίηση της ήδη υπάρχουσας πληροφορίας από την τρέχουσα επανάληψη του αλγορίθμου έτσι ώστε να γίνεται αποτελεσματικότερη επιλογή των σχηματισμών που θα δοκιμαστούν και ελεγχθούν στην αντικειμενική συνάρτηση. Επίσης, αλγόριθμοι που επεξεργάζονται σε κάθε βήμα ένα σύνολο σχηματισμών, δηλαδή υποψήφιων λύσεων, και όχι έναν μοναδικό σχηματισμό όπως στον αλγόριθμο τοπικής αναζήτησης, μπορούν να επιταχύνουν την εύρεση της βέλτιστης λύσης αλλά και να αποφύγουν τοπικά ακρότατα. Τέλος, οι στοχαστικοί αλγόριθμοι βελτιστοποίησης επιτρέπουν τη μετάβαση σε τιμές που οδηγούν σε αύξηση/μείωση της αντικειμενικής συνάρτησης, όταν από το πρόβλημα υπαγορεύεται το αντίθετο, δίνοντας τη δυνατότητα της αποφυγής των επικείμενων τοπικών ακροτάτων και με την ελπίδα ότι θα βρεθούν σε επόμενα βήματα καλύτερες λύσεις *(David, D. & Fouskakis, D. (2002). Stochastic Op-*

timization: a Review. International Statistical Review 70, 3, 315–49).
Στο επόμενο κεφάλαιο (Ενότητα 2.6) θα γίνει πιο εκτενής παρουσίαση των χαρακτηριστικών των στοχαστικών αλγορίθμων και θα εντοπιστούν οι διαφορές τους με τις κλασσικές (αναλυτικές) μεθόδους βελτιστοποίησης.

1.3 Ευριστικοί αλγόριθμοι

Οι ευριστικές μέθοδοι (heuristic methods) συνιστούν μια προσέγγιση των προβλημάτων βελτιστοποίησης χωρίς να εγγυώνται την εύρεση της βέλτιστης λύσης, αλλά μιας λύσης αποδεδειγμένα καλής σε εύλογο υπολογιστικό χρόνο. Χρησιμοποιούνται ως τεχνική σε προβλήματα στα οποία η εύρεση της βέλτιστης λύσης με εξαντλητική αναζήτηση στις υποψήφιες λύσεις είναι χρονοβόρα ή και ανέφικτη, επομένως η εφαρμογή ευριστικών μεθόδων επιταχύνει τη διαδικασία εύρεσης μιας ικανοποιητικής λύσης. Τέτοιου είδους προβλήματα είναι τα λεγόμενα NP-hard προβλήματα τα οποία λύνονται σε πολυωνυμικό χρόνο. Σε κάθε βήμα των ευριστικών μεθόδων, καθορίζεται το σύνολο των πιθανών λύσεων που θα διερευνηθούν σε επόμενο βήμα, με βάση το αποτέλεσμα του τρέχοντος βήματος. Η πρωταρχική και θεμελιώδης μέθοδος που ανήκει στην κατηγορία των ευριστικών είναι η "μέθοδος δοκιμής και σφάλματος" (trial and error) κατά την οποία δοκιμάζεται μια υποψήφια λύση και αν ο αλγόριθμος επιστρέψει μήνυμα λάθους, τότε δοκιμάζεται μια επόμενη υποψήφια. Οι ευριστικές μέθοδοι εξαρτώνται από το πρόβλημα που επιλύουν (problem-dependent), επομένως λαμβάνουν υπόψη τις ιδιαιτερότητες του προβλήματος. Ωστόσο, υπάρχει ο κίνδυνος να εγκλωβιστούν σε τοπικά ακρότατα, με συνέπεια να αποτυγχάνουν στην εύρεση της ολικής βέλτιστης λύσης.
Από την άλλη πλευρά, οι μεθευριστικές διαδικασίες (metaheuristics) είναι ανεξάρτητες από το πρόβλημα που επιλύουν (problem-independent) με την έννοια ότι δεν λαμβάνουν υπόψη ιδιαιτερότητες του προβλήματος. Σε αντίθεση με τις ευριστικές μεθόδους, έχουν την ικανότητα να αποφεύγουν τοπικά ακρότατα, και αυτό το επιτυγχάνουν με την προσωρινή αποδοχή χειρότερων λύσεων η οποία επιτρέπει την καλύτερη αναζήτηση του χώρου των λύσεων. Επιπλέον, καθώς ψάχνουν σε μεγάλες περιοχές του χώρου των εφικτών λύσεων, οι μεθευριστικοί αλγόριθμοι μπορούν συχνά να βρίσκουν αποδεδειγμένα καλές λύσεις, σε μικρότερο υπολογιστικό χρόνο συγκριτικά με τις κλασσικές τεχνικές βελτιστοποίησης και για το λόγο αυτό συνιστούν χρήσιμες προσεγγίσεις για τα προβλήματα αυτά.

1.4 Ορολογία-Βασικές Έννοιες

Η επιστήμη που ασχολείται με τους μηχανισμούς που είναι υπεύθυνοι για τις ομοιότητες και τις διαφορές των ειδών ονομάζεται Γενετική. Η λέξη "γενετική" προέρχεται από την ελληνική λέξη "γέννηση" και είναι η επιστήμη που μας βοηθά να διαφοροποιούμε την κληρονομικότητα από τις γενετικές τροποποιήσεις. Η ιδέα του γενετικού αλγορίθμου είναι εμπνευσμένη από τη φυσική εξέλιξη των ειδών, για το λόγο αυτό η ορολογία που εμπλέκεται στην κατασκευή και τη λειτουργία του γενετικού αλγορίθμου είναι δανεισμένη από την βιολογία και συγκεκριμένα τη γενετική.
Αρχικά, το **γονίδιο** αποτελεί τη βασική δομική μονάδα στη γενετική αλλά και στη μέθοδο βελτιστοποίησης την οποία εξετάζουμε. Στη βιολογία, τα γονίδια αποτελούν την κωδικοποιημένη παράσταση των ιδιοτήτων των ειδών, δηλαδή τα χαρακτηριστικά ενός ατόμου όπως λόγου χάρη το γονίδιο που καθορίζει το χρώμα των ματιών. Στον γενετικό αλγόριθμο, τα επιμέρους χαρακτηριστικά – μεταβλητές μιας μαθηματικής λύσης μπορούν να κωδικοποιηθούν με μια μορφή που να επιτρέπει την εφαρμογή διαδικασιών ανάλογων με τη βιολογική ανταλλαγή γενετικού υλικού. Όπως θα αναλυθεί και παρακάτω, η ευρύτερα χρησιμοποιούμενη αναπαράσταση είναι η δυαδική. Η ομάδα των δυαδικών ψηφίων (bits) που κωδικοποιούν μια συγκεκριμένη μεταβλητή είναι ανάλογη του γονιδίου στη βιολογία.
Όλη η γενετική πληροφορία αποθηκεύεται στα χρωμοσώματα. Κάθε χρωμόσωμα απαρτίζεται από τα γονίδια. Κατά αναλογία, ολόκληρη η αναπαράσταση μιας μαθηματικής λύσης είναι ένα **χρωμόσωμα** για το γενετικό αλγόριθμο. Συχνά στα πλαίσια του γενετικού αλγορίθμου, το χρωμόσωμα αναφέρεται και ως **άτομο**. Ένα σύνολο χρωμοσωμάτων στον γενετικό αλγόριθμο αποτελεί έναν **πληθυσμό**, δηλαδή ένα σύνολο από υποψήφιες λύσεις που πρέπει να ελεγχθούν ως προς την καταλληλότητά τους για το συγκεκριμένο πρόβλημα.
Το πόσο κατάλληλος για επιβίωση και αναπαραγωγή είναι ένας οργανισμός, εξαρτάται από την αλληλεπίδραση των χαρακτηριστικών που διαθέτει με το περιβάλλον. Στο πεδίο των γενετικών αλγορίθμων, το μαθηματικό ανάλογο των χαρακτηριστικών μπορεί να είναι οι μεταβλητές μιας συνάρτησης, ενώ το ρόλο του περιβάλλοντος διαδραματίζει η ίδια η συνάρτηση προς βελτιστοποίηση. Πιο απλά, η **συνάρτηση καταλληλότητας** ή **συνάρτηση αξιολόγησης** καθορίζει ποιά χρωμοσώματα είναι πιο "καλά" για το δεδομένο πρόβλημα και επομένως θα αναπαραχθούν για να δημιουργήσουν ακόμα καλύτερα χρωμοσώματα.
Η φυσική επιλογή είναι απόρροια της επιβίωσης των ισχυρότερων οργανισμών, και της συμμετοχής τους στη διαδικασία της αναπαραγωγής. Για τους γενετικούς αλγορίθμους, ο μηχανισμός αυτός υλοποιείται εν μέρει με την απόδοση καταλληλότητας σε άτομα-λύσεις, αλλά χρειάζεται και ένας αλγόριθμος

επιλογής των ατόμων με τις καλύτερες τιμές καταλληλότητας. Ακόμη πιστότερη μεταφορά του φυσικού μηχανισμού στους γενετικούς αλγορίθμους επιτυγχάνεται όταν υπεισέρχεται και ο παράγοντας της τυχαιότητας στην επιλογή. Η διαδικασία της επιλογής επιτρέπει στα καλύτερα προσαρμοσμένα άτομα να αναπαράγονται, και έτσι να μεταβιβάζουν τα χαρακτηριστικά τους στις επόμενες γενιές. Αυτό συμβαίνει και στον γενετικό αλγόριθμο, δηλαδή οι πιθανές λύσεις ανταγωνίζονται η μια με την άλλη με τέτοιο τρόπο, ώστε τα χρωμοσώματα που αντιπροσωπεύουν τις καλύτερες λύσεις για το δεδομένο πρόβλημα να έχουν περισσότερες πιθανότητες να επιλεγούν.
Μια βασική λειτουργία που συντελείται για την εξέλιξη των ειδών είναι η **αναπαραγωγή**. Η διαδικασία της αναπαραγωγής ανάγεται στη διασταύρωση των χρωμοσωμάτων που έχουν επιλεγεί και διαδραματίζουν το ρόλο των "γονέων". Με τον τρόπο αυτό προκύπτουν καινούρια χρωμοσώματα τα οποία συνδυάζουν τις ήδη υπάρχουσες αλληλουχίες των γονιδίων που αρχικά ανήκουν στους "γονείς". Πρόκειται για το ανάλογο της ανταλλαγής γενετικού υλικού που συμβαίνει κατά τη βιολογική αναπαραγωγή. Μια άλλη διαδικασία η οποία είναι εμπνευσμένη από τη βιολογία είναι η **μετάλλαξη**, η οποία θεωρείται μέρος της διαδικασίας της αναπαραγωγής και συνίσταται στην τυχαία μεταβολή του χρωμοσώματος. Για τους γενετικούς αλγορίθμους οι πράξεις της επιλογής, της διασταύρωσης και της μετάλλαξης ονομάζονται **γενετικοί τελεστές**. Ένα τελευταίο στοιχείο της βιολογικής εξέλιξης που μένει να προσομοιωθεί για να ολοκληρωθεί η σκιαγράφηση του γενετικού αλγορίθμου είναι η ανανέωση του πληθυσμού. Η παραγωγή των απογόνων από τον τρέχοντα πληθυσμό συνεπάγεται εν μέρει και την αντικατάσταση του από νέα άτομα. Ο πληθυσμός που προκύπτει σε κάθε βήμα ανανέωσης αναφέρεται ως **γενιά**.
Ο χώρος όλων των δυνατών λύσεων, δηλαδή ο χώρος στον οποίο βρίσκεται και η βέλτιστη λύση ονομάζεται χώρος αναζήτησης. Κάθε ένα σημείο του χώρου αυτού αντιπροσωπεύει μια δυνατή, εφικτή λύση. Συνεπώς, κάθε δυνατή λύση μπορεί να χαρακτηριστεί από την τιμή καταλληλότητάς της (fitness value), η οποία αποδίδεται από τη συνάρτηση καταλληλότητας που επιλέγουμε να χρησιμοποιήσουμε στο πρόβλημα. Σε κάθε περίπτωση, το επιθυμητό και αναμενόμενο είναι να αυξάνεται η μέση καταλληλότητα του πληθυσμού, βασιζόμενοι στην αρχή ότι επιβιώνουν και συνδυάζονται οι καλύτεροι. Κάθε άτομο που παράγεται είναι μια προσπάθεια από τον αλγόριθμο για εύρεση της βέλτιστης λύσης, μέγιστης ή ελάχιστης, η οποία βασίζεται σε κάποιο βαθμό στην τυχαιότητα. Κάθε συνδυασμός καλών ατόμων, ερμηνεύεται ως μετακίνηση προς την κατεύθυνση μιας καλύτερης λύσης, μέχρι ο αλγόριθμος τελικά να συγκλίνει.

Κεφάλαιο 2

Ο ΓΕΝΕΤΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ ΣΤΗ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

2.1 Είδη Γενετικών Αλγορίθμων

1. Generational Genetic Algorithm
Σύμφωνα με τον παραδοσιακό γενετικό αλγόριθμο, η εξέλιξη επιτυγχάνεται μέσα από μια αλληλουχία διακεκριμένων γενεών που η μια διαδέχεται την άλλη, χωρίς αυτές να αλληλεπικαλύπτονται. Ένα άτομο υπάρχει μόνο σε μια γενιά και μπορεί να επηρεάσει τις επόμενες μόνο διαμέσου της διασταύρωσης. Κατά συνέπεια, σε κάθε αναπαραγωγικό κύκλο ο αλγόριθμος παράγει μια εντελώς νέα γενιά από απογόνους η οποία αντικαθιστά εξ ολοκλήρου την προηγούμενη.

2. Steady-state Genetic Algorithm
Σε αυτό το είδος γενετικού αλγορίθμου μόνο ένα μέρος της τρέχουσας γενιάς αντικαθίσταται από παιδιά του αναπαραγωγικού κύκλου. Το γεγονός αυτό έχει ως αποτέλεσμα τη δημιουργία μιας γενιάς η οποία προσωρινά έχει μεγαλύτερο μέγεθος οπότε επιλέγονται και απομακρύνονται άτομα από τον προσωρινό αυτό πληθυσμό μέχρις ότου να μειωθεί στα κανονικά επίπεδα. Για διαγραφή προορίζονται τα άτομα που είναι πιο αδύναμα συγκριτικά με τα υπόλοιπα, δηλαδή παρουσιάζουν χαμηλή καταλληλότητα, όπως θα αναλυθεί παρακάτω. Έτσι διατηρείται σταθερό το μέγεθος του πληθυσμού σε κάθε γενιά. Το ποσοστό του πληθυσμού που θα αντικατασταθεί καθορίζεται από το χρήστη. Γενικά, οι steady-state γενετικοί αλγόριθμοι πετυχαίνουν γρηγορότερη σύγκλιση. Ο γενετικός αλγόριθμος που κατασκευάστηκε στην παρούσα διπώματική εργασία κατατάσσεται σε

αυτήν την κατηγορία.

2.2 Τύποι Κωδικοποίησης

Η κωδικοποίηση είναι η διαδικασία αναπαράστασης των γονιδίων ενός χρωμοσώματος. Η διαδικασία αυτή μπορεί να πραγματοποιηθεί χρησιμοποιώντας δυαδικά ψηφία (bits), αριθμούς, δένδρα, διανύσματα, λίστες ή άλλα αντικείμενα. Η κωδικοποίηση που θα χρησιμοποιηθεί εξαρτάται κυρίως από την επίλυση του προβλήματος.

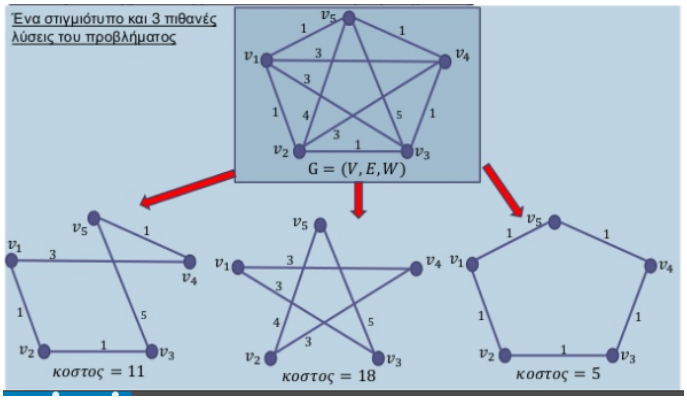
2.2.1 Δυαδική Κωδικοποίηση

Η δυαδική κωδικωποίηση (binary encoding) αποτελεί την πιο συνήθη μορφή κωδικοποίησης στους γενετικούς αλγορίθμους, διότι μέσω αυτής πραγματοποιούνται με μεγάλη ευκολία οι γενετικές διαδικασίες. Κάθε χρωμόσωμα κωδικοποιείται ως μια ακολουθία από δυαδικά ψηφία (bits). Κάθε δυαδικό ψηφίο στο διάνυσμα της λύσης αντιπροσωπεύει κάποιο χαρακτηριστικό της. Με αυτόν τον τρόπο, κάθε ακολουθία δυαδικών ψηφίων αναπαριστά οποιαδήποτε υποψήφια λύση του προβλήματος, όχι μόνο την καλύτερη λύση. Αυτός ο τύπος κωδικοποίησης χρησιμοποιήθηκε και στην παρούσα εργασία.

2.2.2 Κωδικοποίηση με φυσικούς αριθμούς

Υπάρχουν προβλήματα στα οποία τα χρωμοσώματα αναπαριστούν την αλληλουχία κάποιων φυσικών αριθμών. Το πρόβλημα του περιοδεύοντος πωλητή (travelling salesman problem) είναι χαρακτηριστικό αυτού του τύπου κωδικοποίησης. Θα παρουσιαστεί συνοπτικά το πρόβλημα και ο τρόπος με τον οποίο ο γενετικός αλγόριθμος εφαρμόζεται προκειμένου να το επιλύσει, έτσι ώστε ο αναγνώστης να αποκτήσει μια πληρέστερη εικόνα για τους γενετικούς αλγορίθμους, ωστόσο θα περιοριστούμε σε μια σύντομη παρουσίαση διότι ξεφεύγει των σκοπών της παρούσας διπλωματικής εργασίας.
Στο πρόβλημα αυτό ένας πωλητής πρέπει να διέλθει από ένα πεπερασμένο σύνολο πόλεων τα οποία είναι σημεία πώλησης, με τους εξής περιορισμούς: Πρέπει να περάσει ακριβώς μία φορά από κάθε πόλη ενώ ταυτόχρονα θα πρέπει

να ξεκινάει και να ολοκληρώνει την πορεία του στην ίδια πόλη. Ζητούμενο είναι να ελαχιστοποιηθεί η χιλιομετρική απόσταση που θα διανύσει ώστε να ελαχιστοποιηθεί και το κόστος για τη μεταφορά του. Η μοντελοποίηση του προβλήματος θα γίνει με χρήση ενός μη κατευθυνόμενου γράφου $G = (V, E, W)$ με βάρη στις ακμές όπου: V είναι το σύνολο των κορυφών (πόλεων), E είναι το σύνολο των ακμών (συνήθως συνδέεται κάθε ζεύγος διαφορετικών κορυφών) και W είναι η συνάρτηση βαρών ακμών, τέτοια ώστε $W : E \to R^+$, δηλαδή ανατίθενται θετικά βάρη στις ακμές. Αναζητούμε τον κύκλο Hamilton ελαχίστου βάρους, δηλαδή τον κύκλο που διέρχεται από όλες τις κορυφές. Στο ακόλουθο Σχήμα 2.1 φαίνονται μερικές ενδεικτικές λύσεις για το συγκεκριμένο πρόβλημα.



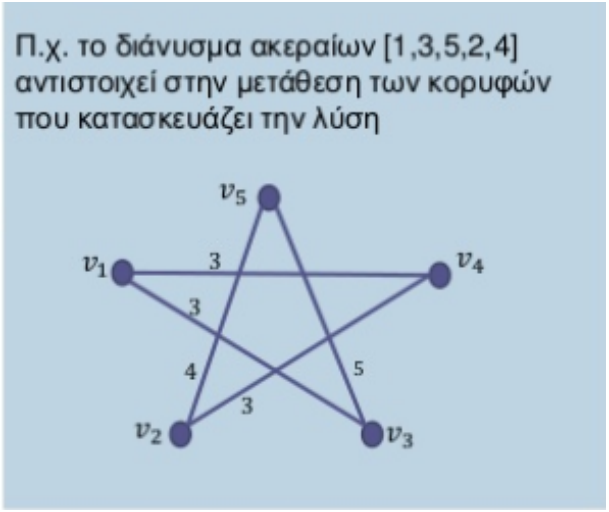
Σχήμα 2.1: Πιθανές λύσεις στο πρόβλημα του περιοδεύοντος πωλητή

Μια λύση, δηλαδή ένα χρωμόσωμα, θα αναπαρίσταται ως ένα διάνυσμα φυσικών αριθμών που απεικονίζει τη σειρά επίσκεψης των κόμβων στην τρέχουσα λύση του προβλήματος. Οι λύσεις που έχει ο χώρος αναζήτησης είναι εκθετικά πολλές, αποδεικνύεται μάλιστα ότι είναι n!, όσες δηλαδή και οι τοποθετήσεις των n πόλεων σε μία σειρά. Ο γενετικός αλγόριθμος, κατάλληλα τροποποιημένος ώστε να δουλεύει με διάνυσμα φυσικών, παρέχει έναν αποδοτικό τρόπο εξερεύνησης του χώρου αναζήτησης του προβλήματος. Η αντικειμενική συνάρτηση για το πρόβλημα του περιοδεύοντος πωλητή ορίζεται ως το άθροισμα των βαρών των ακμών που χρησιμοποιεί η τρέχουσα λύση, δηλαδή

$$f(\pi) = \sum_{i=1}^{n} (w(\pi(i), \pi(i+1))$$

όπου π μια μετάθεση των κορυφών του γραφήματος και θεωρούμε ότι $\pi(n+1) = \pi(1)$, δηλαδή ο κόμβος αρχής ταυτίζεται με τον κόμβο τέλους. Επειδή πρόκειται

για συνάρτηση ελαχιστοποίησης, πρέπει να την τροποποιήσουμε κατάλληλα:
A τρόπος: F=-f+C , όπου C κατάλληλη σταθερά τέτοια ώστε
C : (πόλεις) x (Μέγιστη απόσταση δύο πόλεων)
B τρόπος: F=1/f.
Επιλέγουμε να δουλέψουμε με τον B τρόπο.
Για παράδειγμα, στο γράφο 5 πόλεων η πρώτη υποψήφια λύση αναπαρίσταται με το διάνυσμα $[v_1, v_2, v_3, v_5, v_4]$, η δεύτερη λύση με το διάνυσμα $[v_1, v_3, v_5, v_2, v_4]$, η τρίτη λύση με το διάνυσμα $[v_1, v_2, v_3, v_4, v_5]$. Αυτό σημαίνει ότι η δεύτερη μετάθεση των κορυφών περιγράφεται με μαθηματικό τρόπο από τη μετάθεση $\pi = [1, 3, 5, 2, 4]$. Άρα, η αντικειμενική συνάρτηση είναι $f(\pi) = w(\pi(1), \pi(2)) + w(\pi(2), \pi(3)) + w(\pi(3), \pi(4)) + w(\pi(4), \pi(5)) + w(\pi(5), \pi(1)) = w(1, 3) + w(3, 5) + w(5, 2) + w(2, 4) + w(4, 1) = 3 + 3 + 4 + 5 + 3 = 18$. Επιλέξαμε τον B τρόπο, άρα η τελική τιμή της αντικειμενικής συνάρτησης θα έχει την τιμή $F = 1/18 = 0.056$. Το Σχήμα 2.2 απεικονίζει τη μετάθεση των κορυφών που αντιστοιχεί στη δεύτερη προτεινόμενη λύση και τα ανάλογα βάρη.



Σχήμα 2.2: Πιθανή λύση στο πρόβλημα του περιοδεύοντος πωλητή που αντιστοιχεί στη μετάθεση π=[1,3,5,2,4]

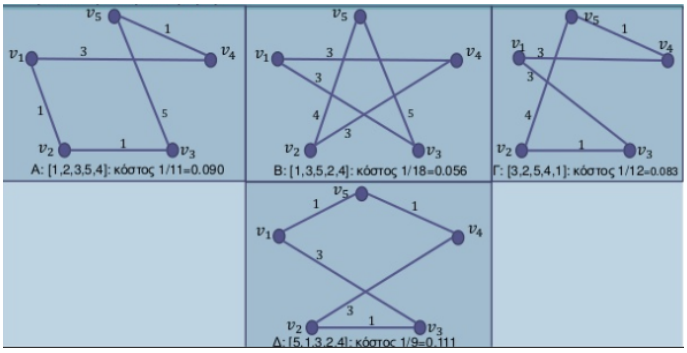
Αφού παρουσιαστούν οι γενετικοί τελεστές που αποτελούν αναπόσπαστο κομμάτι για την εφαρμογή του γενετικού αλγορίθμου, θα γίνει αναφορά και στον τρόπο που αυτοί εφαρμόζονται στο πρόβλημα του περιοδεύοντος πωλητή.

2.3 Παρουσίαση Γενετικών Τελεστών

2.3.1 Αρχικοποίηση

Η αρχικοποίηση του πληθυσμού (Initialisation) αποτελεί το πρώτο βήμα του γενετικού αλγορίθμου. Ο αρχικός πληθυσμός αποτελεί και την πρώτη γενιά του αλγορίθμου, η οποία είναι ένα υποσύνολο του χώρου των δυνατών λύσεων. Πιο συγκεκριμένα, ο αρχικός πληθυσμός αποτελείται από τα χρωμοσώματα, τα οποία μπορούν να παραχθούν είτε με τυχαίο τρόπο (random initialisation) είτε με χρήση κάποιου άλλου ευριστικού αλγορίθμου (heuristic initialisation). Στην περίπτωσή μας, τα χρωμοσώματα αναπαρίστανται ως δυαδικά διανύσματα, δηλαδή ως μια αλληλουχία από 0 και 1 με καθορισμένο μήκος. Με αυτήν την κωδικοποίηση, επιλέχθηκε η τυχαία δειγματοληψία με επανάθεση από μηδενικά και άσσους έτσι ώστε κάθε διάνυσμα (ισοδύναμα κάθε χρωμόσωμα) να αποτελεί δείγμα με μέγεθος ίσο με αυτό του διανύσματος. Αυτό σημαίνει ότι ο αρχικός πληθυσμός θα αποτελείται από εντελώς τυχαίες λύσεις που όμως είναι εφικτές, διότι ανήκουν στο χώρο λύσεων του προβλήματος. Γενικότερα, η συνήθης κωδικοποίηση που πραγματοποιείται σε γενετικούς αλγορίθμους για τα χρωμοσώματα είναι η δυαδική, δηλαδή τα χρωμοσώματα αποτελούνται από μηδενικά και άσσους. Με αυτόν τον τρόπο διευκολύνονται όλες οι γενετικές διαδικασίες όπως θα δούμε παρακάτω.
Όπως αναφέρθηκε παραπάνω, η επιλογή των αρχικών χρωμοσωμάτων μπορεί να γίνει και με ευριστικές μεθόδους όπως η βελτιστοποίηση σμήνους σωματιδίων (particle swarm optimization) δίνοντας εξαρχής ένα πλεονέκτημα στον γενετικό αλγόριθμο ως προς την αναζήτηση της βέλτιστης λύσης. Ωστόσο, αυτό κρύβει τον κίνδυνο ο αρχικός πληθυσμός να αποτελείται από παρόμοια χρωμοσώματα και άρα να αναπαριστά λύσεις με μικρές διαφοροποιήσεις, που σημαίνει ότι ο αλγόριθμος περιορίζεται σε ένα μικρό μέρος του χώρου των δυνατών λύσεων. Επειδή με αυτόν τόν τρόπο ενδέχεται ο αλγόριθμος να οδηγηθεί σε πρόωρη σύγκλιση (βλέπε Ενότητα 2.5.1), συστήνεται η αρχικοποίηση να γίνεται ως εξής: να επιλέγεται ένα σημείο στον πληθυσμό μέχρι το οποίο τα χρωμοσώματα έχουν παραχθεί από ευριστική μέθοδο, ενώ το υπόλοιπο μέρος του πληθυσμού να αποτελείται από τυχαία χρωμοσώματα. Έτσι συνδυάζονται αποτελεσματικά οι δύο προαναφερθείσες μέθοδοι αρχικοποίησης του πληθυσμού.
Στο πρόβλημα του περιοδεύοντος πωλητή όπου η αναπαράσταση γίνεται με τη χρήση φυσικών αριθμών, στο βήμα της αρχικοποίησης δημιουργούμε έναν τυχαίο πληθυσμό από δυνατές λύσεις με καθορισμένο πλήθος. Η αρχικοποίηση μπορεί να γίνει είτε με χρήση κάποιου άπληστου αλγορίθμου (greedy algorithm), δηλαδή ενός αλγορίθμου που επιλέγει την πόλη που βρίσκεται πιο κοντά στην τρέχουσα θέση του πωλητή κάθε φορά, είτε κατασκευάζοντας μια τυχαία μετάθεση των πόλεων. Παραδείγματος χάρη, έστω ότι για το συγκεκριμένο

πρόβλημα ο πληθυσμός έχει μέγεθος 4, αυτό σημαίνει ότι θα παράξουμε 4 διανύσματα φυσικών αριθμών με τυχαίο τρόπο. Ενδεικτικά:
Α=[1,2,3,5,4]
Β=[1,3,5,2,4]
Γ=[3,2,5,4,1]
Δ=[5,1,2,3,4]
Ως επόμενο βήμα, πρέπει να υπολογίσουμε την τιμή της αντικειμενικής συνάρτησης για τις εν λόγω αρχικές λύσεις. Το αποτέλεσμα που δίνει κάθε μία από τις λύσεις Α, Β, Γ και Δ φαίνεται στο Σχήμα 2.3.



Σχήμα 2.3: Υπολογισμός αντικειμενικής συνάρτησης πιθανών λύσεων στο πρόβλημα του περιοδεύοντος πωλητή

2.3.2 Επιλογή των γονέων

Για να διαπιστώσουμε πόσο αποδοτικό είναι ένα χρωμόσωμα, δηλαδή πόσο "κοντά" βρίσκεται στη βέλτιστη λύση του προβλήματος προς επίλυση, υπολογίζουμε μια συνάρτηση η οποία ονομάζεται **συνάρτηση καταλληλότητας** (fitness function) ή ισοδύναμα **συνάρτηση αξιολόγησης** (evaluation function). Όπως έχει ήδη αναφερθεί, ο γενετικός αλγόριθμος σε τελικό στάδιο πρέπει να δώσει ως αποτέλεσμα τις βέλτιστες λύσεις για το πρόβλημα που επιλύει. Κατά συνέπεια, χρειάζεται να αποδώσουμε μια τιμή σε κάθε χρωμόσωμα που να είναι ενδεικτική της ποιότητας της λύσης που αυτό αντιπροσωπεύει, ένα μέτρο δηλαδή που να περιγράφει το πόσο ικανοποιητική είναι η λύση.

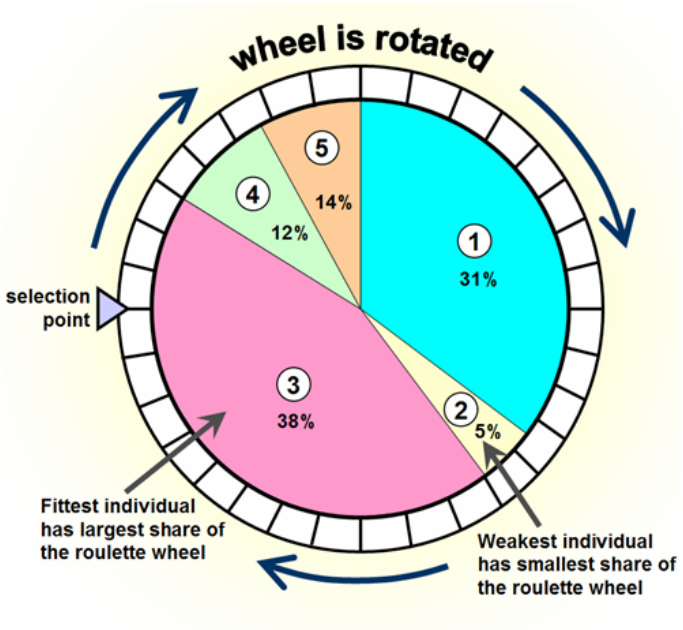
Με την επιλογή (Selection) εφαρμόζεται στα πλαίσια του αλγορίθμου, ο νόμος επιβίωσης του ικανότερου. Είναι η διαδικασία η οποία προηγείται της αναπαραγωγής, και καθορίζει ποιά άτομα από τον υπάρχοντα πληθυσμό θα έχουν την ευκαιρία να συμμετέχουν στη διαδικασία της αναπαραγωγής και επομένως να κληροδοτήσουν στην επόμενη γενιά τα χαρακτηριστικά τους. Στόχος, λοιπόν,

της λειτουργίας της επιλογής είναι να επιτρέπει στα "ικανότερα" άτομα να αυξάνονται στον πληθυσμό, έτσι ώστε να παράγουν "παιδιά" με υψηλότερη τιμή στη συνάρτηση καταλληλότητας. Ο ρόλος της επιλογής ως προκαταρκτικό στάδιο της αναπαραγωγής είναι καθοριστικός, διότι χωρίς αυτήν ο γενετικός αλγόριθμος εκτελεί τυχαίο ψάξιμο (random search) στο χώρο των λύσεων.
Αυτό που πρέπει να καθοριστεί από τον αλγόριθμο, είναι ο τρόπος με τον οποίο θα επιλεχθούν αυτοί οι δύο "γονείς" και ο αριθμός των "παιδιών" που θα δημιουργήσουν. Υπάρχουν διάφοροι τρόποι επιλογής των γονέων, οι οποίοι αναλύονται παρακάτω. Αυτό που πρέπει να τονιστεί όμως, είναι το γεγονός ότι "αυστηρή επιλογή" με την έννοια της αυξημένης απαίτησης ως προς την τιμή της συνάρτησης αξιολόγησης οδηγεί σε επικράτηση περιορισμένων χρωμοσωμάτων στον πληθυσμό, σε βάρος της ποικιλότητας που απαιτείται ώστε να διευκολυνθεί η καλύτερη εξερεύνηση του χώρου των λύσεων. Σε αντίθεση, "ασθενής επιλογή" συνεπάγεται και αργή εξέλιξη του πληθυσμού, άρα και αργή σύγκλιση. Για το λόγο αυτό, θα πρέπει να επιλέγεται προσεχτικά η συνάρτηση αξιολόγησης που χρησιμοποιείται.
Οι πιο γνωστοί τρόποι επιλογής είναι οι ακόλουθοι:

- **Μέθοδος του τροχού της τύχης**

Η μέθοδος του τροχού της τύχης (Roulette Wheel Method) εναλλακτικά, καλείται "επιλογή ανάλογη της καταλληλότητας" ή "ποσοστιαία επιλογή".
Αυτή η μέθοδος αποτελεί μια από τις επικρατέστερες μεθόδους επιλογής και βασίζεται στην τιμή καταλληλότητας που εκχωρείται σε κάθε μέλος που ανήκει στον πληθυσμό. Η τιμή καταλληλότητας μας παρέχει ένα μέτρο ποιότητας του εκάστοτε χρωμοσώματος σχετικά με το ενδεχόμενο της επιλογής του ως "γονέα" της επόμενης γενιάς. Αν υποθέσουμε ότι f_i είναι η τιμή της συνάρτησης αξιολόγησης για το χρωμόσωμα i του πληθυσμού και N το πλήθος των χρωμοσωμάτων, τότε η πιθανότητα του i-οστού χρωμοσώματος να επιλεγεί ως γονέας για αναπαραγωγή είναι ίση με το λόγο:
$p_i = \frac{f_i}{\sum_{j=1}^{N} f_j}$.
Αν φανταστούμε το συνολικό πληθυσμό να σχηματίζει το δίσκο της ρουλέτας (τροχός της τύχης) και κάθε άτομο του πληθυσμού να αντιπροσωπεύεται από ένα χώρο ανάλογο με την κατάλληλότητά του, τότε προκύπτει εύλογα η αναλογία με τη ρουλέτα. Είναι φανερό ότι όσο μεγαλύτερη είναι η τιμή καταλληλότητας για κάποιο μέλος του πληθυσμού, τόσο μεγαλύτερο χώρο θα καταλαμβάνει στη ρουλέτα και επομένως τόσο μεγαλύτερη είναι η πιθανότητα να επιλεγεί ως γονέας συγκριτικά με τα υπόλοιπα μέλη που η τιμή καταλληλοτητάς τους είναι μικρότερη. Πρόκειται για μια "δίκαιη" διαδικασία επιλογής, καθώς όλα τα μέλη του πληθυσμού έχουν τη δυνατότητα να επιλεγούν και να συνεχίσει η

διαδικασία της εξέλιξής τους μέσω των υπόλοιπων γενετικών τελεστών.
Η διαδικασία με την οποία γίνεται η επιλογή μέσω της μεθόδου του τροχού της τύχης είναι η ακόλουθη: Ο τροχός γυρίζει επαναλαμβανόμενα N φορές, όπου N ο αριθμός των ατόμων του πληθυσμού. Τα μέλη επιλέγονται όταν ο τροχός σταματήσει στη θέση που υποδεικνύεται από τον δείκτη. Πρόκειται για μια στοχαστική διαδικασία η οποία υλοποιείται μέσω τυχαίας δειγματοληψίας με αντικατάσταση. Στο Σχήμα 2.4 φαίνεται και γραφικά η διαδικασία επιλογής των γονέων με τη μέθοδο του τροχού της τύχης.



Σχήμα 2.4: Μέθοδος του τροχού της τύχης

Ο αλγόριθμος που υλοποιεί τη μέθοδο του τροχού της τύχης συνοψίζεται ακολούθως:
Βήμα 1: [Υπολογισμός της συνάρτησης καταλληλότητας]
Για κάθε χρωμόσωμα του πληθυσμού υπολογίζεται η τιμή της συνάρτησης καταλληλότητας.
Βήμα 2: [Άθροισμα]
Αθροίζουμε τις τιμές καταλληλότητας των χρωμοσωμάτων του πληθυσμού και το άθροισμα αυτό το συμβολίζουμε με S. Δηλαδή

$$S = \sum_{j=1}^{N} f_j \, .$$

Βήμα 3: [Επιλογή]
Παράγουμε έναν τυχαίο αριθμό στο διάστημα [0,S] από την ομοιόμορφη κατανομή τον οποίο συμβολίζουμε με r.
Βήμα 4: [Επαναλήψεις]
Αθροίζουμε τις τιμές καταλληλότητας των χρωμοσωμάτων μέχρι το άθροισμα να γίνει μεγαλύτερο ή ίσο από την τιμή του δείκτη r. Το χρωμόσωμα που θα επιλεγεί είναι το τελευταίο στοιχείο που εισάγαμε στην άθροιση. Η διαδικασία αυτή πραγματοποιείται συνολικά N φορές, όσο και το μέγεθος του πληθυσμού.

- **Επιλογή με βάση την κατάταξη**

Σύμφωνα με την επιλογή κατάταξης (Rank Selection), τα άτομα του πληθυσμού ταξινομούνται ανάλογα με την καταλληλότητά τους και κατόπιν λαμβάνουν μια τιμή καταλληλότητας που προκύπτει από αυτήν την ταξινόμηση. Συγκεκριμένα, το άτομο με τη μικρότερη τιμή στη συνάρτηση καταλληλότητας θα λάβει την τιμή 1, το δεύτερο χειρότερο την τιμή 2 κ.ο.κ, ενώ το καλύτερο θα λάβει τιμή ίση με το μέγεθος του πληθυσμού. Με τη διαδικασία αυτή, όλα τα άτομα έχουν τη δυνατότητα να επιλεχθούν ως "γονείς". Με άλλα λόγια, αντί να λαμβάνεται υπόψη η ίδια η καταλληλότητα του ατόμου για την επιλογή, χρησιμοποιείται ο βαθμός κατάταξης αυτής, με συνέπεια να αποφεύγεται η εκτεταμένη επανεπιλογή κάποιων συγκεκριμένων αποδεδειγμένα καλών ατόμων. Η μέθοδος αυτή αποτρέπει την πρόωρη σύγκλιση του γενετικού αλγορίθμου, διότι μειώνει έμμεσα την κυριαρχία των πολύ καλών ατόμων, ωστόσο μπορεί να αποδειχθεί υπολογιστικά χρονοβόρα, διότι συρρικνώνει τις διαφορές στην τιμή καταλληλότητας μεταξύ των ατόμων με αποτέλεσμα να οδηγεί σε αργή σύγκλιση. Οι έννοιες αυτές για τα σύγκλιση θα μελετηθούν εκτενέστερα στην Ενότητα 2.5.
Η πιθανότητα επιλογής του εκάστοτε χρωμοσώματος ως "γονέα" σύμφωνα με τη μέθοδο αυτή ορίζεται ως:
$p_i = \frac{rank(i)}{n \cdot (n-1)}$
όπου n το μέγεθος του πληθυσμού και $rank(i)$ η κατάταξη του εκάστοτε ατόμου.

- **Τυχαία επιλογή**

Υπάρχουν αρκετές εκδοχές αυτής της μεθόδου, η πιο γενική από αυτές είναι η κ-τυχαία επιλογή (k-tournament selection). Σύμφωνα με τη μέθοδο αυτή, επιλέγονται κ άτομα από τον πληθυσμό και τελικά κρατείται αυτό με τη μεγαλύτερη τιμή στη συνάρτηση καταλληλότητας. Τα υπόλοιπα κ-1 άτομα που δεν επιλέχθηκαν, επιστρέφουν στον πληθυσμό και η διαδικασία αυτή επαναλαμβάνεται τόσες φορές, όσες και ο αριθμός των "γονέων" που θέλουμε. Τελικά,

τα άτομα που προκύπτουν από τη διαδικασια αυτή δεν επιλέγονται με κάποιο ντετερμινιστικό τρόπο, αλλά στοχαστικά, μέσω μιας τυχαίας δειγματοληψίας με επανάθεση στον πληθυσμό.
Στη συγκεκριμένη εργασία, υλοποιήθηκε αυτή η μέθοδος επιλογής "γονέων" για κ=50, όσο και το μέγεθος του πληθυσμού που ορίστηκε στον συγκεκριμένο γενετικό αλγόριθο. Πιο συγκεκριμένα, από όλο τον πληθυσμό επιλέχθηκε ένα άτομο μέσω τυχαίας δειγματοληψίας με επανάθεση και η διαδικασία αυτή πραγματοποιήθηκε 50 φορές συνολικά, όσες και το μέγεθος του πληθυσμού. Αυτό έγινε επειδή επιλέξαμε το πλήθος των "γονέων" να ισούται με το μέγεθος του πληθυσμού σε κάθε επανάληψη. Οι πιθανότητες επιλογής των γονέων είναι ανάλογες των τιμών καταλληλότητας, με την έννοια ότι τα άτομα με μεγαλύτερη τιμή καταλληλότητας είναι πιο πιθανό να επιλεγούν. Επειδή η δειγματοληψία γίνεται με επανάθεση, αυτό σημαίνει ότι ενδέχεται κάποια χρωμοσώματα να επιλεγούν παραπάνω από μία φορά ως "γονείς", ενώ κάποια άλλα με μικρή τιμή καταλληλότητας να μην επιλεγούν καθόλου.

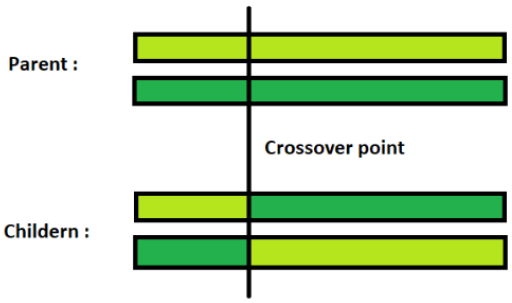
2.3.3 Διασταύρωση

Αφού έχει πραγματοποιηθεί η επιλογή των γονέων, ακολουθεί ίσως η σημαντικότερη λειτουργία του γενετικού αλγορίθμου, η αναπαραγωγή ή διασταύρωση (Crossover).
Η διαδικασία της διασταύρωσης ταυτίζεται με τη διαδικασία της φυσικής αναπαραγωγής. Στην αναπαραγωγή, όπως γίνεται στη φύση, τελείται αμοιβαία ανταλλαγή μέρους του DNA μεταξύ μη συγγενών ζευγαριών χρωμοσωμάτων. Παρόμοια, στους γενετικούς αλγορίθμους κατά τη διάρκεια της εφαρμογής των τελεστών της διασταύρωσης, εκτελείται μια διαδικασία αντιγραφής μέρους του γενετικού υλικού μεταξύ των εμπλεκόμενων μελών. Έτσι έχουμε τη διάδοση χαρακτηριστικών και ιδιοτήτων των προγενέστερων προς τους απογόνους οι οποίοι κληρονομούν ορισμένα από τα χαρακτηριστικά της προηγούμενης γενιάς, όπως συμβαίνει και στα βιολογικά συστήματα.
Πιο συγκεκριμένα, ο πληθυσμός που προέκυψε από τη διαδικασία της επιλογής πρέπει να περάσει από τη φάση του "ζευγαρώματος" για να πραγματοποιηθεί ένα είδος αναπαραγωγής, σύμφωνα με τα παραπάνω. Η νέα, λοιπόν, ομάδα ατόμων που προέκυψε σχηματίζει με τυχαίο τρόπο ομάδες των δύο. Το ποιος θα ζευγαρώσει με ποιον ενδέχεται να επηρεάζει την ταχύτητα σύγκλισης του αλγορίθμου, προς το παρόν όμως αυτό αποτελεί αντικείμενο μελέτης, και στην παρούσα εργασία θα περιοριστούμε στο ζευγάρωμα με τυχαίο τρόπο.
Σε κάθε ομάδα, τα δύο χρωμοσώματα που λαμβάνουν μέρος στη διαδικασία ανταλλαγής γενετικού υλικού ονομάζονται "γονείς". Η διασταύρωση είναι απαραίτητη λειτουργία που συμβάλλει ουσιαστικά στην επίδοση ενός γενετικού

αλγορίθμου, και για το λόγο αυτό έχουν επινοηθεί αρκετοί διαφορετικοί τρόποι υλοποίησής της. Μερικοί μπορούν να εφαρμοστούν σε κάθε τύπο προβλήματος, ενώ άλλοι είναι πιο εξειδικευμένοι για να επιλύσουν ορισμένα προβλήματα βελτιστοποίησης, και επομένως αφήνεται στη διακριτική ευχέρεια του σχεδιαστή του προγράμματος να επιλέξει ποιόν τρόπο διασταύρωσης θα αξιοποιήσει για το εκάστοτε πρόβλημα. Στόχος της διασταύρωσης είναι η νέα γενιά που θα προκύψει από το συνδυασμό των "γονέων" να περιλαμβάνει άτομα που θα φέρουν τα καλύτερα χαρακτηριστικά από αυτά που είχε η προηγούμενη γενιά.
Η διαδικασία της διασταύρωσης, εκτός των άλλων, είναι ιδιαίτερα χρήσιμη επειδή ανακατευθύνει το ψάξιμο σε ανεξερεύνητες περιοχές του χώρου αναζήτησης και άρα αυξάνονται οι πιθανότητες επιτυχίας του γενετικού αλγορίθμου. Τα νέα άτομα που προκύπτουν είναι στην πλειοψηφία τους επιτυχημένοι συνδυασμοί υψηλότερης ικανότητας από τους "γονείς" τους, που ανήκουν στην προηγούμενη γενιά. Στην περίπτωση που από τη διασταύρωση προκύψουν χρωμοσώματα με χαμηλότερη καταλληλότητα, αυτά δε θα έχουν μεγάλη πιθανότητα να επιλεγούν ως "γονείς" στον επόμενο αναπαραγωγικό κύκλο και επομένως δε θα επιβιώσουν.
Στην πράξη, η διασταύρωση λαμβάνει χώρα με κάποια πιθανότητα, τη λεγόμενη "πιθανότητα διασταύρωσης" (crossover probability). Η πιθανότητα αυτή καθορίζεται από το χρήστη και δίνεται ως παράμετρος για να ξεκινήσει ο αλγόριθμος. Ωστόσο, η τιμή της μπορεί να αλλάξει με το πέρας των επαναλήψεων, επηρεάζοντας έτσι τη σύγκλιση του γενετικού αλγορίθμου. Πιο συγκεκριμένα, συνήθως επιλέγεται μεγάλη πιθανότητα διασταύρωσης κατά τα πρώτα στάδια του αλγορίθμου, δηλαδή στην παραγωγή των αρχικών γενεών που σηματοδοτούν την αρχή του ψαξίματος στο χώρο αναζήτησης, ενώ η πιθανότητα αυτή μειώνεται όταν ο αλγόριθμος προσεγγίζει την τιμή του βέλτιστου *(Deepa, S.N. & Sivanandam, S.N. (2008). Introduction to Genetic Algorithms. Springer-Verlag Berlin Heidelberg*).
Στα πλαίσια της εφαρμογής της διασταύρωσης σε δυαδικές συμβολοσειρές, όπως ισχύει και στην περίπτωση μας, παράγονται ακριβώς δύο "απόγονοι" από τον συνδυασμό δύο "γονέων". Ο συνδυασμός των τμημάτων των αρχικών δυαδικών ακολουθιών-χρωμοσωμάτων γίνεται με τρόπο ώστε το i-οστό bit του "απογόνου" να είναι το i-οστό bit ενός από τους δύο "γονείς".
Υπάρχουν αρκετές τεχνικές διασταύρωσης. Η ιδέα πίσω από αυτές είναι ότι τα τμήματα ενός χρωμοσώματος που συμβάλλουν περισσότερο στην απόδοσή του ενδέχεται να μην είναι γειτονικά. Για το λόγο αυτό, ανταλλάσσονται τμήματα των υπαρχόντων χρωμοσωμάτων με την προσδοκία βελτίωσης της απόδοσης των χρωμοσωμάτων που θα προκύψουν.
Οι πιο γνωστές μέθοδοι διασταύρωσης είναι οι παρακάτω:

- **Διασταύρωση ενός σημείου**

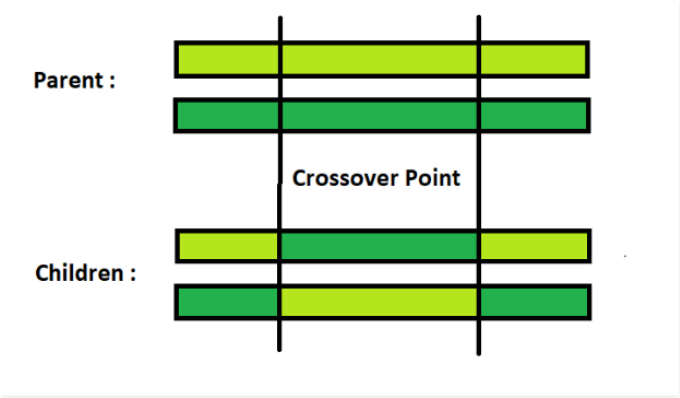
Στη διασταύρωση ενός σημείου (One-point crossover) πραγματοποιείται τυχαία επιλογή μιας θέσης (γονίδιο) εντός των χρωμοσωμάτων που εμπλέκονται ("γονείς"). Κάθε χρωμόσωμα έχει καθορισμένο μήκος και αυτό είναι κοινό για όλα τα χρωμοσώματα που παράγονται κατά τη διάρκεια των επαναλήψεων. Το σημείο που επιλέχθηκε ονομάζεται "σημείο διασταύρωσης". Το πρώτο χρωμόσωμα-"απόγονος" προκύπτει από τη συνένωση του πρώτου τμήματος του ενός "γονέα", μέχρι το σημείο διασταύρωσης, και του δεύτερου τμήματος του δεύτερου "γονέα". Αντίστροφα για το δεύτερο χρωμόσωμα-"απόγονο". Με άλλα λόγια, γίνεται μια τομή σε τυχαίο σημείο του χρωμοσώματος και εναλλάσσονται τα δύο τμήματα των "γονέων" πριν και μετά την τομή. Στο Σχήμα 2.5 απεικονίζεται ο τρόπος υλοποίησης της διασταύρωσης ενός σημείου.



Σχήμα 2.5: Διασταύρωση ενός σημείου

- **Διασταύρωση δύο σημείων**

Στη διασταύρωση δύο σημείων (Two-point crossover) αυτή τα σημεία που επιλέγονται με τυχαίο τρόπο είναι δύο. Φυσικά, θα πρέπει να είναι κοινά και για τους δύο "γονείς" και να βρίσκονται στο εσωτερικό των εμπλεκόμενων χρωμοσωμάτων. Στη διασταύρωση δύο σημείων γίνονται εναλλαγές μεταξύ των τμημάτων των χρωμοσωμάτων-"γονέων" που προκύπτουν μετά την τομή. Έτσι προκύπτουν και πάλι δύο νέα χρωμοσώματα. Αυτός ο τύπος διασταύρωσης επεκτείνεται και για n-σημεία, οπότε προκύπτει η "διασταύρωση n-σημείων" ή "διασταύρωση πολλαπλών σημείων". Στο Σχήμα 2.6 απεικονίζεται ο τρόπος υλοποίησης της διασταύρωσης δύο σημείων.



Σχήμα 2.6: Διασταύρωση δύο σημείων

Στα Σχήματα 2.7 και 2.8 φαίνεται η υλοποίηση της διασταύρωσης ενός και δύο σημείων με χρήση δυαδικών διανυσμάτων.

Chromosome1	11011|00100110110
Chromosome2	11011|11000011110
Offspring1	11011|11000011110
Offspring2	11011|00100110110

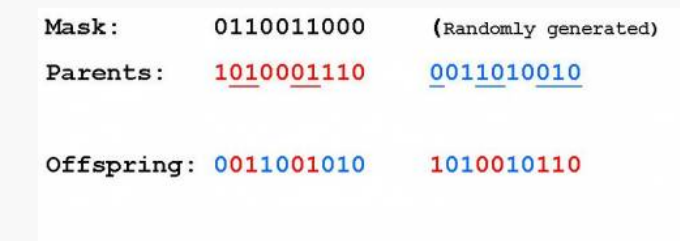
Σχήμα 2.7: Διασταύρωση ενός σημείου με χρήση δυαδικών διανυσμάτων-χρωμοσωμάτων

Chromosome1	11011|00100|110110
Chromosome2	10101|11000|011110
Offspring1	11011|11000|110110
Offspring2	10101|00100|011110

Σχήμα 2.8: Διασταύρωση δύο σημείων με χρήση δυαδικών διανυσμάτων-χρωμοσωμάτων

- **Ομοιόμορφη διασταύρωση**

Υπενθυμίζεται ότι η διασταύρωση πολλαπλών σημείων καθορίζει τα σημεία διασταύρωσης ως τις θέσεις όπου το χρωμόσωμα μπορεί να διαιρείται. Η ομοιόμορφη διασταύρωση (Uniform crossover) γενικεύει την ιδέα αυτή και θεωρεί κάθε θέση μέσα στο χρωμόσωμα ως σημείο διασταύρωσης. Αυτό επιτυγχάνεται με τη χρήση μιας "μάσκας διασταύρωσης" (mask crossover) η οποία δημιουργείται τυχαία και είναι ίδιου μήκους με το μήκος των "γονέων". Ουσιαστικά, κάθε γονίδιο στους "απογόνους" δημιουργείται από την αντιγραφή του αντίστοιχου γονιδίου από έναν από τους "γονείς" και η επιλογή του "γονέα" γίνεται μέσω της μάσκας. Στην περίπτωση που έχει επιλεχθεί δυαδική κωδικοποίηση για τον γενετικό αλγόριθμο, η μάσκα είναι επίσης μια δυαδική συμβολοσειρά τα ψηφία της οποίας καθορίζουν τον γονέα ο οποίος θα παρέχει στον απόγονο το εκάστοτε ψηφίο. Συγκεκριμένα, όταν το ψηφίο της μάσκας είναι 1, τότε ο πρώτος "απόγονος" θα κληροδοτήσει το αντίστοιχο ψηφίο από τον πρώτο "γονέα", ενώ αντίθετα όταν το ψηφίο της μάσκας είναι 0 τότε ο πρώτος "απόγονος" θα κληροδοτήσει το αντίστοιχο ψηφίο από τον δεύτερο "γονέα". Ο δεύτερος "απόγονος" προκύπτει με την ίδια λογική αν πάρουμε αντίθετη μάσκα ή ισοδύναμα αν ανταλλάξουμε τους δύο "γονείς". Με αυτόν τον τρόπο, οι απόγονοι περιέχουν ένα μείγμα από γονίδια από τον εκάστοτε "γονέα".
Στο Σχήμα 2.9 φαίνεται μέσω ενός παραδείγματος η υλοποίηση της ομοιόμορφης διασταύρωσης.



Σχήμα 2.9: Ομοιόμορφη διασταύρωση

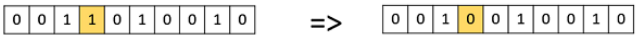
Είναι φανερό από τα παραπάνω ότι η διασταύρωση ενός σημείου είναι η απλούστερη μορφή διασταύρωσης. Ωστόσο, η χρήση της συνεπάγεται το εξής μειονέκτημα: εφόσον υπάρχει ένα μόνο σημείο τομής, η κεφαλή και η ουρά ενός χρωμοσώματος-"γονέα" (οι πρώτες και οι τελευταίες θέσεις του χρωμοσώματος αντίστοιχα) δεν μπορούν να συνυπάρξουν και να "περάσουν" αναλλοίωτες στα χρωμοσώματα-"απογόνους", σε περίπτωση που περιέχουν καλό γενετικό υλικό, δηλαδή χρήσιμη πληροφορία. Αυτό μπορεί να αποφευχθεί με τη χρήση της διασταύρωσης δύο ή και παραπάνω σημείων. Συστήνεται να δοκιμαστούν αρκετοί διαφορετικοί τύποι διασταύρωσης προκειμένου να διερευνηθεί ποιός είναι ο κατάλληλος για το εκάστοτε πρόβλημα βελτιστοποίησης.
Αξίζει να σημειωθεί ότι η επιλογή του κατάλληλου τελεστή διασταύρωσης εξαρτάται εν πολλοίς από την αναπαράσταση του χώρου αναζήτησης λύσης του προβλήματος. Τα ακολουθιακά προβλήματα, όπως τα προβλήματα αναζήτησης βέλτιστης πορείας (πρόβλημα περιοδεύοντος πωλητή) συχνά απαιτούν τη χρήση διαφορετικών τελεστών από αυτούς που περιγράφηκαν παραπάνω, διότι ενδέχεται οι απόγονοι που θα προκύψουν να είναι εκτός του συνόλου των επιτρεπόμενων λύσεων. Θα εξετάσουμε τη δυνατότητα διασταύρωσης στο πρόβλημα του περιοδεύοντος πωλητή στο οποίο η κωδικοποίηση που απαιτείται για τον γενετικό αλγόριθμο δεν είναι η δυαδική, αλλά γίνεται χρήση της κωδικοποίησης με φυσικούς αριθμούς.
Επιστρέφοντας στο πρόβλημα του περιοδεύοντος πωλητή, επιβάλλεται να κάνουμε κάποιες τροποποιήσεις στη διαδικασία της διασταύρωσης προκειμένου να πραγματοποιηθεί σεβόμενη τους περιορισμούς που επιβάλλονται από το πρόβλημα. Πιο συγκεκριμένα, έστω δύο χρωμοσώματα-"γονείς":
Γονέας 1: 7 5 2 6 3 1 9 4 8
Γονέας 2: 9 1 2 3 8 6 7 5 4
Έστω ότι επιλέγουμε διασταύρωση ενός σημείου, και έστω ότι το σημείο της διασταύρωσης είναι η θέση μεταξύ τρίτου και τέταρτου ψηφίου. Τότε, οι απόγονοι που θα προέκυπταν μέσω της διαδικασίας της διασταύρωσης είναι οι εξής:
7 5 2—3 8 6 7 5 4 και 9 1 2—6 3 1 9 4 8. Τότε όμως προκύπτουν λύσεις στις οποίες υπάρχουν επαναλαμβανόμενες πόλεις οι οποίες αντιστοιχούν στις

υπογραμμισμένες ακολουθίες στα χρωμοσώματα 7 5 2 3 8 6 7 5 4 και 9 1 2 6 3 1 9 4 8. Αυτές οι λύσεις δεν είναι αποδεκτές διότι αντιβαίνουν στον κανόνα της μοναδικότητας ύπαρξης της εκάστοτε πόλης από την οποία διέρχεται κάθε φορά ο πωλητής. Για να εκτελεστεί η διασταύρωση σωστά σε τέτοιου είδους προβλήματα, επιλέγονται κάποιες συγκεκριμένες θέσεις πόλεων, παραδείγματος χάρη η τέταρτη, η έκτη και η έβδομη θέση του χρωμοσώματος. Αυτές αποτελούν και τις θέσεις του πρώτου χρωμοσώματος-"γονέα" που θα γίνει αλλαγή με ψηφία από το δεύτερο χρωμόσωμα-"γονέα", ενώ κάποιες άλλες τρεις θέσεις του δεύτερου χρωμοσώματος-"γονέα" θα αλλάξουν σύμφωνα με τα ψηφία του πρώτου χρωμοσώματος-"γονέα". Συγκεκριμένα, για τον πρώτο γονέα και τις θέσεις που επιλέξαμε προκύπτει 7 5 2 **6** 3 **1 9** 4 8. Τα ψηφία αυτά (6,1,9) θα πάρουν τη θέση των ίδιων στοιχείων στο δεύτερο χρωμόσωμα με τη σειρά που υπάρχουν στο πρώτο. Πιο αναλυτικά, αυτό σημαίνει ότι για το δεύτερο χρωμόσωμα θα έχουμε την τροποποίηση από **9 1** 2 3 8 **6** 7 5 4 σε **6 1** 2 3 8 **9** 7 5 4 το οποίο αποτελεί και τον πρώτο απόγονο. Αντίστοιχα, για την τέταρτη, την έκτη και την έβδομη θέση του δεύτερου χρωμοσώματος έχουμε τα στοιχεία 9 1 2 **3** 8 **6 7** 5 4. Κατά τον ίδιο τρόπο, τα ψηφία αυτά (3,6,7) θα πάρουν τη θέση των ίδιων στοιχείων στο πρώτο χρωμόσωμα με τη σειρά που εντοπίζονται στο δεύτερο. Δηλαδή για το πρώτο χρωμόσωμα θα έχουμε την τροποποίηση **7** 5 2 **6 3** 1 9 4 8 σε **3** 5 2 **6 7** 1 9 4 8, το οποίο θα είναι και ο δεύτερος απόγονος. Με τη μέθοδο αυτή επιτυγχάνουμε μια αναδιάταξη των στοιχείων των δύο αρχικών χρωμοσωμάτων κατά τέτοιο τρόπο ώστε να μην υπάρχουν επαναλήψεις των ίδιων στοιχείων. Βέβαια, αυτός είναι ένας ενδεικτικός τρόπος εφαρμογής της διασταύρωσης στο πρόβλημα του περιοδεύοντος πωλητή και υπάρχουν και άλλοι δυνατοί τρόποι, η παρουσίαση των οποίων ξεφεύγει από τα πλαίσια της παρούσας διπλωματικής εργασίας.

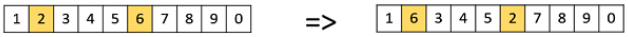
2.3.4 Μετάλλαξη

Κατόπιν της επιλογής και της διασταύρωσης, σειρά έχει ο τελεστής της μετάλλαξης (Mutation). Ως διαδικασία είναι λιγότερο σημαντική από τη διασταύρωση, αλλά σίγουρα χρήσιμη διότι με τη χρήση του μπορεί να ανακτηθεί τυχόν χαμένη πληροφορία σε κάποιο γονίδιο.
Στη γενετική, όταν αναφερόμαστε στον όρο της μετάλλαξης, εννοούμε κάποια γενετική ανωμαλία. Η μετάλλαξη στη φύση μπορεί να συμβεί κατά τη διάρκεια της φάσης της αντιγραφής του γενετικού υλικού. Γενικά, οι γενετικές ανωμαλίες οφείλονται στην μικρή τροποποίηση που μπορεί να υποστεί το γενετικό υλικό. Αυτή είναι και η λειτουργία της μετάλλαξης στους γενετικούς αλγορίθμους, με την έννοια ότι ο τελεστής αυτός προκαλεί συνήθως μικρές αλλαγές στην τιμή ενός ή περισσότερων γονιδίων στα χρωμοσώματα του πληθυσμού.

Στην περίπτωση των γενετικών αλγορίθμων η μετάλλαξη γίνεται με κάποια τυχαιότητα, όσον αφορά τόσο το χρωμόσωμα, όσο και το γονίδιο στο οποίο θα εφαρμοστεί. Συγκεκριμένα, με βάση την "πιθανότητα μετάλλαξης" (mutation probability) επιλέγονται τα χρωμοσώματα τα οποία θα μεταλλαχτούν. Η πιθανότητα αυτή δίνεται ως παράμετρος στον αλγόριθμο από τον χρήστη, ενώ δύναται να πάρει άλλη τιμή με το πέρας των επαναλήψεων. Στην απλούστερη μορφή του, ο τελεστής μετάλλαξης αλλάζει την τιμή ενός τυχαίου bit (γονιδίων) κάποιου "απογόνου". Μια άλλη περίπτωση είναι η αντιστροφή περισσότερων τυχαίων bit, ενώ είναι δυνατή και η πλήρης αντικατάσταση του χρωμοσώματος από κάποιο άλλο τυχαίο. Ανεξαρτήτως μορφής, συνήθως εφαρμόζεται μετά τη διασταύρωση, με μικρή πιθανότητα, συνήθως κυμαίνεται μεταξύ του 5-10%.
Η μετάλλαξη "προλαμβάνει" τον αλγόριθμο ώστε να μην παγιδευτεί σε τοπικό ακρότατο. Αν η διασταύρωση είναι υπεύθυνη για τη σωστή εκμετάλλευση των έως τώρα λύσεων, με την έννοια ότι κατευθύνει τον αλγόριθμο σε καλύτερες περιοχές αναζήτησης λύσεων, τότε η μετάλλαξη εξασφαλίζει ότι κανένα σημείο του χώρου δεν αποκλείεται από την αναζήτηση. Γενικά, η μετάλλαξη θεωρείται ως ο τελεστής που διατηρεί την γενετική ποικιλομορφία στον πληθυσμό, αφού εισάγει νέα γενετική δομή στον πληθυσμό τροποποιώντας με τυχαίο τρόπο τις ήδη υπάρχουσες γενετικές δομές. Επίσης εξασφαλίζει την εργοδικότητα του χώρου των λύσεων, με την έννοια ότι αποδίδει μη μηδενική πιθανότητα στην παραγωγή οποιασδήποτε λύσης κατά την δημιουργία του εκάστοτε πληθυσμού-γενιάς.
Συμπερασματικά, λοιπόν, ο τελεστής της μετάλλαξης είναι σημαντικός στην αναπαραγωγική διαδικασία καθώς φροντίζει ώστε να εισαχθούν νέα άτομα στον πληθυσμό, άρα και νέα χαρακτηριστικά. Όταν συμβαίνει, επιφέρει ποικιλία στον πληθυσμό και ανακατευθύνει την αναζήτηση σε νέες ανεξερεύνητες έως τώρα περιοχές του χώρου των λύσεων προκειμένου να αποφευχθεί ο εγκλωβισμός σε τοπικά ακρότατα. Λειτουργεί ως "ασφαλιστική δικλείδα" για τις περιπτώσεις κατά τις οποίες η επιλογή και η διασταύρωση χάσουν κάποιες πολύτιμες γενετικές πληροφορίες.
Υπάρχουν αρκετοί διαφορετικοί τρόποι μετάλλαξης για όλους τους διαφορετικούς τύπους κωδικοποίησης σε έναν γενετικό αλγόριθμο. Για τη δυαδική κωδικοποίηση συγκεκριμένα, η πιο συνήθης μέθοδος μετάλλαξης πραγματοποιείται με την αντιστροφή ενός γονιδίου σε ένα χρωμόσωμα με κάποια μικρή πιθανότητα (bit flip). Αυτή η πιθανότητα συνήθως ορίζεται ως 1/L, όπου L το μήκος του χρωμοσώματος. Ένας άλλος τρόπος μετάλλαξης είναι η επιλογή δύο τυχαίων θέσεων εντός του χρωμοσώματος και η εναλλαγή των ψηφίων τους (swap).
Στα Σχήματα 2.10 και 2.11 που ακολουθούν απεικονίζονται οι δύο προαναφερθείσες μέθοδοι μετάλλαξης.



Σχήμα 2.10: Μετάλλαξη αντιστροφής ψηφίου



Σχήμα 2.11: Μετάλλαξη εναλλαγής ψηφίων

Στο πρόβλημα του περιοδεύοντος πωλητή η μετάλλαξη γίνεται ανταλλάσσοντας δύο γονίδια ενός χρωμοσώματος. Για παράδειγμα, αν η λύση που εξετάζεται είναι η B=[4 5 2 1 8 7 6 9 3] και επιλεχθούν οι θέσεις 4 και 7 με τυχαίο τρόπο, τότε η διαδικασία της μετάλλαξης λειτουργεί ως εξής: Από την B=[4 5 2 **1** 8 7 **6** 9 3] παράγεται η λύση B'=[4 5 2 **6** 8 7 **1** 9 3]. Με αυτόν τον τρόπο αποφεύγεται η επανάληψη διπλών πόλεων που απαγορεύεται να συμβεί από την οδηγία του προβλήματος.

2.3.5 Ελιτισμός

Το πόσο γρήγορα εντοπίζει ο γενετικός αλγόριθμος τη βέλτιστη λύση, ή μια αποδεδειγμένα καλή λύση, είναι ενδεικτικό και της καλής απόδοσής του. Ο Ελιτισμός (Elitism), είναι ένα μέτρο που διευκολύνει τη σύγκλιση και η εφαρμογή του είναι δυνατή σε οποιοδήποτε από τα σχήματα επιλογής που έχουν παρουσιαστεί. Η ελιτίστικη στρατηγική αποτελεί μέρος της διαδικασίας της αναπαραγωγής στα πλαίσια του γενετικού αλγορίθμου, δεν ακολουθεί όμως τα πρότυπα της φύσης, με την έννοια ότι κατά κάποιον τρόπο επιβάλλει στο μηχανισμό επιλογής του γενετικού αλγορίθμου να επιλέξει κάποιο στοιχείο του πληθυσμού που ίσως να μην επιλεγόταν. Συνήθως η ελιτίστικη στρατηγική χρησιμοποιείται σαν προσθήκη στις μεθόδους επιλογής που ήδη αναφέρθηκαν. Επί της ουσίας, ο ελιτισμός είναι μια διαφορετική επιλογή για την ανανέωση του πληθυσμού. Στη βιβλιογραφία συχνά αναφέρεται και ως "πολιτική μερικής ανανέωσης". Σε όσα είδαμε μέχρι στιγμής, τα άτομα κάποιας γενιάς αντικαθίστανται εξ'ολοκλήρου από την επόμενη με το πέρας των επαναλήψεων. Ο ελιτισμός συνεπάγεται ότι τα καταλληλότερα άτομα μιας γενιάς θα αντιγραφούν στην επόμενη, αλλά και θα συμμετέχουν κανονικά στη διαδικασία της δια-

σταύρωσης. Πιο συγκεκριμένα, ο ελιτισμός επιτρέπει στα καλύτερα χρωμοσώματα από τον τρέχοντα πληθυσμό να μεταφερθούν αναλλοίωτα στον επόμενο πληθυσμό-γενιά που θα προκύψει. Πρόκειται δηλαδή για έναν τελεστή που χρησιμοποιείται προκειμένου να διατηρηθεί η ποικιλομορφία του πληθυσμού. Κατά τον ελιτισμό, τα άτομα του πληθυσμού που θεωρούνται πιο "αδύναμα" στην τρέχουσα "γενιά", δηλαδή έχουν χαμηλή τιμή καταλληλότητας σε σχέση με τα υπόλοιπα άτομα, αντικαθίστανται από τα πιο "ισχυρά" άτομα της αμέσως προηγούμενης "γενιάς". Η ελιτίστικη στρατηγική έχει παρόμοιο στόχο με τη μετάλλαξη, δηλαδή φροντίζει ώστε να εισαχθεί νέο γενετικό υλικό στον τρέχοντα πληθυσμό.
Ο ελιτισμός επιλέγεται αν θα εφαρμοστεί από το χρήστη μέσω μιας λογικής παραμέτρου κατά την έναρξη του αλγορίθμου που δέχεται τιμές αληθής (TRUE) ή ψευδής (FALSE). Το πλεονέκτημα που προκύπτει από την εφαρμογή της μεθόδου αυτής είναι ότι η σύγκλιση είναι εγγυημένη, σε περίπτωση που ανακαλυφθεί το ολικό μέγιστο ή ελάχιστο κατά περίπτωση. Ισοδύναμα αυτό σημαίνει ότι δεν τίθεται περίπτωση να χαθεί η καλύτερη λύση σε κάποιο από τα ενδιάμεσα βήματα του αλγορίθμου κατά την παραγωγή και ανανέωση των πληθυσμών.

2.3.6 Τεχνικές αντικατάστασης πληθυσμού

Αφού έχει δημιουργηθεί η νέα γενιά των "απογόνων" από τη διασταύρωση και τη μετάλλαξη, τίθεται το ερώτημα ποια από τις νέες υποψήφιες λύσεις-χρωμοσώματα πρέπει να γίνουν μέλη της επόμενης γενιάς. Η απόφαση αυτή επηρεάζει ουσιαστικά την ταχύτητα σύγκλισης του αλγορίθμου και για το λόγο αυτό έχουν προταθεί αρκετές τεχνικές αντικατάστασης-ανανέωσης πληθυσμού στους γενετικούς αλγορίθμους.

- **Αντικατάσταση γενιάς**

Ολόκληρος ο πληθυσμός αντικαθίσταται από τους "απογόνους" του. Με αυτόν τον τρόπο όλος ο πληθυσμός της προηγούμενης γενιάς χάνεται, με αποτέλεσμα υποψήφιες καλές λύσεις να μην διατηρούνται στην επόμενη γενιά με τον κίνδυνο ο αλγόριθμος να μην συγκλίνει σε μια αποδεδειγμένα καλή λύση. Αξίζει να τονιστεί ότι με το πέρασμα από γενιά σε γενιά, η λύση που εξετάζεται από τον αλγόριθμο ενδέχεται να είναι χειρότερη από την προηγούμενη, και αυτό οφείλεται στην τυχαιότητα που υπάρχει στους στοχαστικούς αλγορίθμους. Επίσης, ενδέχεται μια λύση που ήταν η καλύτερη στη γενιά της, να χαθεί στην επόμενη γενιά και να επανέλθει σε μεταγενέστερη γενιά, οπότε και η ταχύτητα σύγκλισης του αλγορίθμου μειώνεται.

- **Ελιτισμός**

Όταν ο γενετικός αλγόριθμος εφαρμόζει ελιτισμό, η καλύτερη λύση που έχει εντοπιστεί (ή οι λ καλύτερες) δεν χάνεται κατά την εξέλιξη του πληθυσμού και το πέρασμα από τη μια γενιά στην επόμενη, παρά μόνο αν βρεθεί άλλη λύση καλύτερη. Συγκεκριμένα, συγκρίνονται οι απόγονοι με τους γονείς από τους οποίους δημιουργήθηκαν, και τελικά στον νέο πληθυσμό περνά το καλύτερο ζευγάρι αυτών.

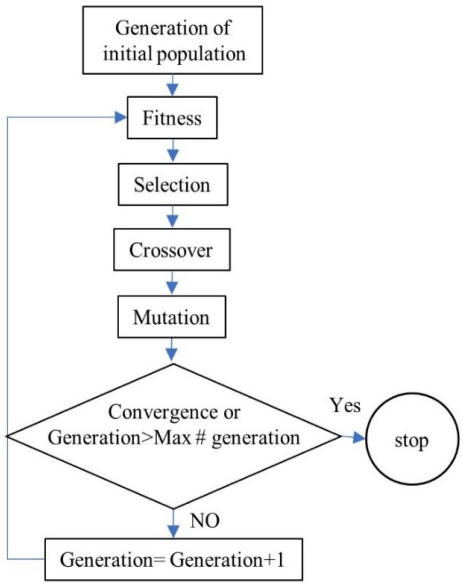
- **Διαγραφή m τελευταίων**

Τα m πιο αδύναμα χρωμοσώματα αντικαθίστανται από m απογόνους οι οποίοι παράγονται με τυχαίο τρόπο, δηλαδή όπως παράχθηκε ο αρχικός πληθυσμός. Με αυτόν τον τρόπο, γίνεται ανανέωση του πληθυσμού χωρίς να χάνονται τα καλύτερα μέλη αυτού και συγχρόνως τα νέα άτομα που εισέρχονται δεν βασίζονται στα προϋπάρχοντα, οπότε επιτυγχάνεται καλύτερη αναζήτηση στον χώρο των λύσεων. Φυσικά, για να έχει νόημα η διαδικασία αυτή θα πρέπει να ισχύει $m < N$, δηλαδή ο αριθμός των ατόμων που θα διαγραφούν να είναι μικρότερος από το μέγεθος του πληθυσμού.

- **Διαγραφή m ατόμων**

Σε αντίθεση με την τεχνική διαγραφής των m πιο αδύναμων χρωμοσωμάτων, σε αυτήν την τεχνική αντικαθίστανται m αυθαίρετα επιλεγμένα χρωμοσώματα της παλιάς γενιάς. Αυτό, αν και επιβραδύνει την ταχύτητα σύγκλισης του αλγορίθμου, εντούτοις αποτρέπει το ενδεχόμενο της πρόωρης σύγκλισης (βλέπε Ενότητα 2.5.1) του αλγορίθμου σε τοπικά ακρότατα. Ο περιορισμός $m < N$ ισχύει και σε αυτήν την τεχνική.

Με βάση όλα τα παραπάνω, τα στάδια του γενετικού αλγορίθμου συνοψίζονται στο Σχήμα 2.12.



Σχήμα 2.12: Διάγραμμα ροής γενετικού αλγορίθμου

2.4 Παράμετροι Γενετικού Αλγορίθμου

Η χρήση γενετικού αλγορίθμου, και γενικότερα εξελικτικών αλγορίθμων, συνεπάγεται και τον προσδιορισμό των παραμέτρων που χρησιμοποιούνται για να πραγματοποιηθούν οι γενετικές διαδικασίες. Οι σημαντικότερες παράμετροι που χαρακτηρίζουν τον γενετικό αλγόριθμο είναι οι εξής:

- **Μέγεθος πληθυσμού**

Το μέγεθος του πληθυσμού (population size) καθορίζει το πλήθος των χρωμοσωμάτων που απαρτίζουν την εκάστοτε γενιά ατόμων. Ισοδύναμα, ισούται με τον αριθμό των υποψήφιων λύσεων που επεξεργάζεται ο γενετικός αλγόριθμος σε κάθε του επανάληψη, και στα πλαίσια αυτής της εργασίας συμβολίζεται με N. Το μέγεθος του πληθυσμού θα πρέπει να είναι τέτοιο ώστε ο αλγόριθμος να είναι αφενός γρήγορος και αφετέρου να εξερευνά επαρκές τμήμα του χώρου των λύσεων. Εάν ο πληθυσμός απαρτίζεται από λίγα χρωμοσώματα, ο γενετικός αλγόριθμος έχει μικρές πιθανότητες να εκτελέσει διασταύρωση και κατά συνέπεια διερευνάται ένα μικρό μόνο μέρος του χώρου αναζήτησης. Αντίθετα, εάν υπάρχουν υπερβολικά πολλά χρωμοσώματα στον πληθυσμό, δηλαδή το μέγεθος του πληθυσμού είναι πολύ μεγάλο, ο αλγόριθμος επιβραδύνεται σημαντικά και γίνεται χρονοβόρος. Η μέθοδος δοκιμής-σφάλματος (trial and error) που διέπει όλες τις ευριστικές μεθόδους εφαρμόζεται και κατά την προσπάθεια προσδιορισμού των παραμέτρων, οπότε προτείνονται δοκιμές έτσι ώστε να βρεθεί το

κατάλληλο μέγεθος πληθυσμού. Έρευνες έχουν δείξει ότι από κάποιο όριο και έπειτα, δεν είναι χρήσιμο να αυξάνεται το μέγεθος του πληθυσμού διότι δεν βοηθά στην γρηγορότερη επίλυση του προβλήματος. Γενικά, η παράμετρος αυτή εξαρτάται από τον τύπο της κωδικοποίησης που χρησιμοποιήθηκε αλλά και το πρόβλημα αυτό καθ'αυτό.

- **Αριθμός γενεών**

Ο αριθμός των γενεών (generation size) ταυτίζεται με τον αριθμό των επαναλήψεων που εκτελούνται από τον γενετικό αλγόριθμο. Σε κάθε κύκλο επανάληψης, δίνεται η ευκαιρία στον αλγόριθμο να εκτελέσει την επιλογή των "γονέων" και τις γενετικές διαδικασίες της αναπαραγωγής και μετάλλαξης ώστε να οδηγηθεί στη διερεύνηση καινούριων υποψήφιων λύσεων. Κατά αντιστοιχία με το μέγεθος πληθυσμού, έτσι και ο αριθμός γενεών χρειάζεται να λάβει μια τιμή τέτοια ώστε να εξερευνά ικανοποιητικά το χώρο αλλά ταυτόχρονα να μην επιβραδύνει σημαντικά τη σύγκλιση.

- **Πιθανότητα διασταύρωσης**

Η πιθανότητα διασταύρωσης (crossover probability) είναι η παράμετρος που ελέγχει το πόσο συχνά θα εκτελείται η διασταύρωση. Αν η τιμή της είναι ίση με μηδέν, τότε δεν εκτελείται ποτέ διασταύρωση και οι "απόγονοι" είναι ακριβή αντίτυπα των "γονέων" τους. Αν η τιμή της πάρει οποιαδήποτε άλλη τιμή στο διάστημα (0,1) τότε οι απόγονοι αποτελούνται από κομμάτια και των δύο "γονέων" σύμφωνα με τα σχήματα διασταύρωσης που περιγράφηκαν πιο πάνω. Αν η τιμή της είναι ίση με ένα, τότε όλοι οι απόγονοι δημιουργούνται από διασταύρωση "γονέων" και τότε δε μεταφέρονται αυτούσιες λύσεις από γενιά σε γενιά. Όσο πιο υψηλή είναι αυτή η πιθανότητα, τόσο πιο πολλοί "απόγονοι" δημιουργούνται και επομένως γίνεται εκτενής αναζήτηση στο χώρο των λύσεων. Αντίθετα, όσο πιο χαμηλή είναι, τόσο πιο πολλές λύσεις συντηρούνται κατά το πέρας των γενεών και η αναζήτηση γίνεται πιο συντηρητική.

- **Πιθανότητα μετάλλαξης**

Η πιθανότητα μετάλλαξης (mutation probability) είναι η παράμετρος που καθορίζει τη συχνότητα με την οποία μεταλλάσσονται τα γονίδια των χρωμοσωμάτων. Αν η τιμή της ορισθεί να είναι ίση με το μηδέν, τότε οι "απόγονοι" δημιουργούνται ακριβώς μετά το πέρας της διασταύρωσης ή αντιγράφονται κατευθείαν από τους "γονείς" τους χωρίς να υπόβάλλονται σε καμία αλλαγή. Αν η τιμή της πάρει κάποια τιμή μέσα στο διάστημα (0,1), ένα ή περισσότερα γονίδια στο χρωμόσωμα αντιστρέφονται (για δυαδική κωδικοποίηση). Αν η τιμή της πιθανότητας μετάλλαξης λάβει την τιμή ένα, τότε όλο το χρωμόσωμα

τροποποιείται. Η μετάλλαξη, όπως έχει αναλυθεί και παραπάνω, γενικά προλαμβάνει τον γενετικό αλγόριθμος από τον εγκλωβισμό σε τοπικά ακρότατα, ωστόσο δεν πρέπει να πραγματοποιείται αρκετά συχνά διότι τότε ο αλγόριθμος εκφυλίζεται σε τυχαίο ψάξιμο.

Στα πρώτα στάδια του γενετικού αλγορίθμου, ο μηχανισμός της διασταύρωσης είναι ο υπεύθυνος για την εξερεύνηση του συνόλου των λύσεων (exploration). Καθώς ο αλγόριθμος αρχίζει να συγκλίνει σε κάποιες λύσεις παρόμοιες μεταξύ τους, ο τελεστής της διασταύρωσης γίνεται λιγότερο παραγωγικός, διότι τα άτομα του πληθυσμού παρουσιάζουν πολλές ομοιότητες μεταξύ τους και επομένως οποιοσδήποτε συνδυασμός τους δεν επιφέρει σημαντικές βελτιώσεις στους "απογόνους". Συνεπώς, με το πέρας των γενεών και τη σύγκλιση του αλγορίθμου, η αύξηση της πιθανότητας μετάλλαξης επιτρέπει στον μηχανισμό της μετάλλαξης να αναζητά τοπικά, γύρω από μια αποδεδειγμένα καλή λύση (exploitation), νέους "απογόνους" και έτσι καθίσταται δυνατή η εύρεση ακόμα καλύτερης λύσης.

2.5 Σύγκλιση

Με τον όρο "σύγκλιση" εννοούμε την επικράτηση ενός χρωμοσώματος ή μικρών παραλλαγών αυτού σε μεγάλο ποσοστό στον πληθυσμό. Όταν ο αλγόριθμος διαφαίνεται ότι συγκλίνει σε κάποιο χρωμόσωμα, τότε θα ικανοποιηθεί και κάποιο κριτήριο τερματισμού και επομένως ως βέλτιστη λύση θα θεωρηθεί το χρωμόσωμα αυτό. Υπάρχουν αρκετά κριτήρια τερματισμού των γενετικών αλγορίθμων. Πιο συγκεκριμένα:

- **Μέγιστος αριθμός "γενεών"**

Ο γενετικός αλγόριθμος τερματίζεται όταν ο ολοκληρωθεί η δημιουργία του καθορισμένου από τον χρήστη αριθμού γενεών. Στην παρούσα εργασία, χρησιμοποιήθηκε αυτό το κριτήριο τερματισμού.

- **Σταθεροποίηση στην τιμή καταλληλότητας**

Ο γενετικός αλγόριθμος τερματίζεται όταν δεν υπάρχει αλλαγή στην καλύτερη λύση για έναν προκαθορισμένο αριθμό γενεών.

- **Σταθεροποίηση στην τιμή της αντικειμενικής συνάρτησης**

Ο γενετικός αλγόριθμος τερματίζεται όταν η τιμή της συνάρτησης προς βελτιστοποίηση δεν επιδέχεται περαιτέρω βελτίωση για μια ακολουθία διαδοχικών γενεών. Τότε ο γενετικός αλγόριθμος θεωρείται ότι έχει βρει το ολικό ακρότατο.

- **Ελάχιστη τιμή στη συνάρτηση καταλληλότητας**

Ο γενετικός αλγόριθμος σταματά την αναζήτηση στο χώρο των λύσεων όταν η τιμή στη συνάρτηση καταλληλότητας που έχει βρεθεί για το καλύτερο άτομο του πληθυσμού γίνει ίση ή μικρότερη από μια προκαθορισμένη τιμή. Αυτό το κριτήριο εγγυάται την εύρεση τουλάχιστον μιας σχετικά ικανοποιητικής λύσης.
Ο αλγόριθμος τερματίζεται όταν κάποιο από αυτά τα κριτήρια ικανοποιηθεί. Ωστόσο, ένας συνδυασμός αυτών προτείνεται για πιο έγκυρα αποτελέσματα.

2.5.1 Πρόωρη Σύγκλιση

Η πρόωρη σύγκλιση (premature Convergence) είναι ένα πρόβλημα που μπορεί να εμφανιστεί σε όλους τους εξελικτικούς αλγορίθμους. Κατά την πρόωρη σύγκλιση, ο αλγόριθμος πολύ γρήγορα συγκλίνει γύρω από κάποιο χρωμόσωμα, το οποίο όμως εκπροσωπεί λύση που αποτελεί τοπικό ακρότατο. Ισοδύναμα, πρόωρη σύγκλιση έχουμε όταν ο πληθυσμός "κυριαρχείται" από ένα χρωμόσωμα το οποίο δεν είναι το βέλτιστο, και επομένως κάθε άτομο του πληθυσμού είτε είναι πανομοιότυπο με αυτό, είτε διαφέρει σε ελάχιστα σημεία. Ως εκ τούτου, ο γενετικός αλγόριθμος δεν μπορεί να "ξεφύγει" από το τοπικό αυτό ακρότατο, παρά μόνο με τη διαδικασία της μετάλλαξης, η οποία πρακτικά έχει μηδενική πιθανότητα να συμβεί. Η διαδικασία της διασταύρωσης στη φάση της πρόωρης σύγκλισης που τα χρωμοσώματα είναι πανομοιότυπα ή παρουσιάζουν μικρές παραλλαγές, δεν είναι δυνατό να συμβάλλει στην ποικιλομορφία του πληθυσμού διότι κατά το "ζευγάρωμά" τους θα προκύψουν τα ίδια χρωμοσώματα. Αυτό έχει ως αποτέλεσμα η αναζήτηση να περιορίζεται σε μια μόνο "γειτονιά" του χώρου των λύσεων.
Το φαινόμενο αυτό εμφανίζεται σε περιπτώσεις που η συνάρτηση καταλληλότητας παρουσιάζει απότομες μεταβολές και έντονα τοπικά ακρότατα και μπορεί να αντιμετωπιστεί με δύο τρόπους. Ο πρώτος είναι η απεικόνιση της συνάρτησης καταλληλότητας σε μία νέα συνάρτηση, λιγότερο απότομη (fitness scaling). Ο δεύτερος είναι ο καθορισμός ελάχιστων και μέγιστων ορίων, όσον αφορά το πόσες φορές επιλέγεται ένα χρωμόσωμα προς αναπαραγωγή σε κάθε κύκλο ανανέωσης του πληθυσμού.

- **Fitness scaling**

Αποτελεί την πιο ευρέως χρησιμοποιούμενη μέθοδο απεικόνισης της συνάρτησης καταλληλότητας. Σύμφωνα με αυτή, αντί να χρησιμοποιηθούν οι πραγματικές τιμές της συνάρτησης καταλληλότητας (fitness function), χρησιμοποιούνται οι κανονικοποιημένες τιμές. Αυτό επιτυγχάνεται με την αφαίρεση μιας προκαθορισμένης τιμής από την τιμή της συνάρτησης καταλληλότητας του κάθε χρωμοσώματος και έπειτα από την διαίρεσή αυτής με τη μέση τιμή της συνάρτησης

καταλληλότητας όλων των χρωμοσωμάτων. Κατά συνέπεια, ελέγχεται και ο αριθμός των φορών που θα επιλεγεί ένα χρωμόσωμα για αναπαραγωγή, και άρα πολύ "ικανά" χρωμοσώματα χρησιμοποιουνται ελεγχόμενα για την παράγουν "απογόνων".

2.5.2 Αργή Σύγκλιση

Η αργή σύγκλιση (slow convergence) είναι ουσιαστικά το ακριβώς αντίθετο φαινόμενο της πρόωρης σύγκλισης. Αργή σύγκλιση παρουσιάζεται όταν ακόμη και μετά το πέρας μεγάλου αριθμού επαναλήψεων, ο πληθυσμός εξακολουθεί να μη συγκλίνει σε κάποιο χρωμόσωμα. Το φαινόμενο αυτό εμφανίζεται όταν η συνάρτηση καταλληλότητας έχει μικρές κλίσεις, με αποτέλεσμα τα μέγιστα και τα ελάχιστά της να έχουν αμελητέες διαφορές. Η λύση είναι και πάλι η απεικόνιση της συνάρτησης καταλληλότητας σε μια νέα, η οποία να έχει πιο έντονες διακυμάνσεις.

2.6 Σύγκριση με Κλασσικές Μεθόδους

Οι αιτιοκρατικές ή κλασσικές μέθοδοι εφαρμόζουν μια εξαντλητική αναζήτηση σε όλο το πεδίο ορισμού των μεταβλητών της συνάρτησης προς βελτιστοποίηση. Όπως θα αναλυθεί στο 3ο κεφάλαιο, το πρόβλημα που θα επιλύσουμε είναι στην ουσία ένα πρόβλημα ελαχιστοποίησης. Τα αποτελέσματα των αιτιοκρατικών μεθόδων είναι εγγυημένα τα βέλτιστα, δηλαδή η λύση που θα εντοπίσουν θα είναι με απόλυτη βεβαιότητα η βέλτιστη για το εκάστοτε πρόβλημα. Αυτό είναι και το κυριότερο πλεονέκτημα που παρουσιάζουν έναντι των στοχαστικών μεθόδων, στις οποίες ανήκει και ο γενετικός αλγόριθμος. Το σημαντικότερο μειονέκτημα των αιτιοκρατικών μεθόδων είναι το γεγονός ότι απαιτείται μεγάλη υπολογιστική ισχύς προκειμένου να φτάσουμε στο κριτήριο τερματισμού. Επίσης, είναι αποτρεπτικός παράγοντας ο χρόνος που απαιτείται για να εξετάσουμε ολόκληρο τον χώρο των δυνατών λύσεων, καθώς έχει παρατηρηθεί ότι ο χρόνος αυξάνει όσο αυξάνει το πλήθος των μεταβλητών του προβλήματος.
Οι στοχαστικές μέθοδοι σε αντίθεση με τις αιτιοκρατικές, δεν πραγματοποιούν εξαντλητική αναζήτηση σε όλο το πεδίο ορισμού των μεταβλητών της συνάρτησης προς βελτιστοποίηση. Ο τρόπος με τον οποίο εξετάζουν το πεδίο έρευνάς τους είναι να λαμβάνουν δείγματα από αυτό και να επικεντρώνουν την αναζήτησή τους σε ενδεικνυόμενες "γειτονιές" του, ανάλογα με τα αποτελέσματα που λαμβάνουν καθώς εκτελούνται. Επομένως οι γενετικοί αλγόριθμοι χρησιμοποιούν "πιθανοτικές μεταβάσεις" κατά τη διαδικασία αναζήτησης της βέλτιστης λύσης και άρα δεν λειτουργούν αιτιοκρατικά όπως οι κλασσικές μέθοδοι βελτιστοποίη-

σης. Κατά συνέπεια, τα αποτελέσματα που μας επιστρέφουν τελικά ως λύση δεν είναι τα βέλτιστα με απόλυτη βεβαιότητα.
Ιδιαίτερα οι γενετικοί αλγόριθμοι, αναζητούν τη βέλτιστη λύση σε ολόκληρο τον πληθυσμό των υποψήφιων λύσεων, και όχι σε μεμονωμένες λύσεις όπως οι συμβατικές μέθοδοι βελτιστοποίησης. Αυτό αυξάνει και την αξιοπιστία τους, διότι αυξάνει τις πιθανότητες για την εύρεση του ολικού μεγίστου/ελαχίστου και ταυτόχρονα συμβάλλει στο να αποφευχθεί το ενδεχόμενο εγκλωβισμού της διαδικασίας αναζήτησης σε τοπικά ακρότατα. Αξίζει να σημειωθεί ότι οι γενετικοί αλγόριθμοι χρησιμοποιούν αποκλειστικά τη συνάρτηση καταλληλότητας για τον εντοπισμό της βέλτιστης λύσης, και όχι παραγώγους. Το γεγονός αυτό τους καθιστά ιδιαίτερα εύχρηστους και σε διακριτά προβλήματα στοχαστικής βελτιστοποίησης.
Παρακάτω θα αναλύσουμε τη δομή των αιτιοκρατικών (κλασσικών) μεθόδων βελτιστοποίησης και θα αναλύσουμε τις προϋποθέσεις που απαιτούνται για την ορθή χρήση τους.

2.6.1 Κλασσικοί Αλγόριθμοι Βελτιστοποίησης

Για αρκετά χρόνια στη βιβλιογραφία κυριάρχησαν οι κλασικοί αλγόριθμοι βελτιστοποίησης (αιτιοκρατικές μέθοδοι), οι οποίοι είναι επαναληπτικοί αλγόριθμοι κατάβασης (descent) και έχουν την ακόλουθη γενική μορφή:
$x_{k+1} = x_k + \eta \cdot \Delta x_k$ (1) όπου $k = 0, 1, \ldots$ είναι ο αριθμός επανάληψης του αλγορίθμου, x_k και x_{k+1} είναι οι τιμές της μεταβλητής της αντικειμενικής συνάρτησης στην επανάληψη k και k+1 αντίστοιχα.
Τέλος, η παράμετρος η καλείται "μέγεθος βήματος" (step size), ενώ ο όρος Δx_k ονομάζετι "κατεύθυνση αναζήτησης" (search direction). Χαρακτηριστικό των επαναληπτικών αλγορίθμων κατάβασης είναι ότι σε κάθε επανάληψη ισχύει $f(x_{k+1}) < f(x_k)$, δηλαδή κάθε επανάληψη μειώνει την τιμή της αντικειμενικής συνάρτησης. Όταν η κατεύθυνση αναζήτησης περιγράφεται από την παράγωγο της αντικειμενικής συνάρτησης, δηλαδή $\Delta x_k = -\nabla f(x_k)$, τότε ο αλγόριθμος ονομάζεται "αλγόριθμος κατάβασης βαθμίδας" (gradient descent). Άλλες παραλλαγές του βασικού αλγορίθμου της εξίσωσης (1) αποτελούν η "μέθοδος του Νεύτωνα" (Newton's method) και οι αλγόριθμοι "συζυγούς κατάβασης" (conjugate descent) *(Καμπουρλάζος, Β. & Παπακώστας Γ. (2015). Εισαγωγή στην Υπολογιστική Νοημοσύνη, Κεφάλαιο 3. Αποθετήριο Κάλλιπος).*

2.6.2 Παραδοχές Κλασσικών Αλγορίθμων

Οι αλγόριθμοι που περιγράφηκαν πιο πάνω χρησιμοποιούνται ακόμη και σήμερα ευρύτατα στην βελτιστοποίηση συστημάτων και διαδικασιών, ωστόσο για να γίνει χρήση τους υπάρχουν κάποιες προϋποθέσεις που πρέπει να ικανοποι-

ηθούν.
Όταν δεν ικανοποιούνται, τότε συχνά καταφεύγουμε στη χρήση εξελικτικών αλγορίθμων όπως ο γενετικός. Συγκεκριμένα:

- Στην περίπτωση της ελαχιστοποίησης μιας αντικειμενικής συνάρτησης, μια βασική παραδοχή για την ύπαρξη (μοναδικού) ολικού ελαχίστου είναι αυτή της κυρτότητας. Κυρτή (convex) καλείται μια συνάρτηση f(x) για την οποία ισχύει:
  $f(\lambda \cdot x_1 + (1-\lambda) \cdot x_2) \leq \lambda \cdot f(x_1) + (1-\lambda) \cdot f(x_2)$ για $0 \leq \lambda \leq 1$.

  Το ελάχιστο μιας κυρτής συνάρτησης στο πεδίο ορισμού της είναι ολικό ελάχιστο και όχι μόνο τοπικό ελάχιστο. Ωστόσο, μια αντικειμενική συνάρτηση συχνά δεν είναι κυρτή. Επιπλέον, στους αλγορίθμους στους οποίους γίνεται χρήση της πρώτης παραγώγου της αντικειμενικής συνάρτησης είναι πιθανό το ενδεχόμενο εγκλωβισμού σε τοπικό ελάχιστο, όπου ο αλγόριθμος τερματίζει χωρίς να εγγυάται τον εντοπισμό του ολικού ελαχίστου. Συνεπώς ένας κλασικός αλγόριθμος μπορεί να συγκλίνει σε υποβέλτιστες λύσεις.

- Μια επιπλέον παραδοχή των κλασικών αλγορίθμων, οι οποίοι κάνουν χρήση της πρώτης παραγώγου της αντικειμενικής συνάρτησης, είναι και η ύπαρξη της δεύτερης παραγώγου. Ωστόσο, σε περιπτώσεις πολύπλοκων συστημάτων ο υπολογισμός των δύο παραγώγων είναι δύσκολος.

Η προϋπόθεση των παραπάνω παραδοχών αποτελεί το λόγο για τον οποίο οι κλασικοί αλγόριθμοι βελτιστοποίησης συχνά αποτυγχάνουν σε πολύπλοκα προβλήματα. Η ανάγκη ύπαρξης αποτελεσματικών αλγορίθμων, που να μπορούν να εφαρμοστούν σε πολύπλοκα προβλήματα βελτιστοποίησης αποτελεί ενεργό πεδίο της επιστημονικής έρευνας *(Καμπουρλάζος, Β. & Παπακώστας Γ. (2015). Εισαγωγή στην Υπολογιστική Νοημοσύνη, Κεφάλαιο 3. Αποθετήριο Κάλλιπος).*

2.7 Πλεονεκτήματα και Μειονεκτήματα Γενετικού Αλγορίθμου

Πλεονεκτήματα
Τα βασικότερα πλεονεκτήματα που παρουσιάζουν οι γενετικοί αλγόριθμοι είναι τα παρακάτω:

1. **Μπορούν να επιλύσουν πολύπλοκα προβλήματα σε εύλογο υπολογιστικό χρόνο.**

Ένας από τους σημαντικούς λόγους για τους οποίους γίνεται εκτεταμένη χρήση των γενετικών αλγορίθμων είναι η μεγάλη τους αποδοτικότητα. Έχει αποδειχθεί στην πράξη ότι προβλήματα που επιδέχονται περισσότερες από μια λύσεις, δύσκολα προσδιορισμένες, μπορούν να αντιμετωπιστούν επιτυχώς από του γενετικούς αλγορίθμους. Αξιοσημείωτο είναι επίσης το γεγονός ότι αντιμετωπίζουν με ευχέρεια συναρτήσεις που παρουσιάζουν μεγάλες διακυμάνσεις (noisy functions) οι οποίες συχνά καθιστούν ανεπαρκείς τις κλασικές μεθόδους βελτιστοποίησης στην εύρεση των ακροτάτων τους.

2. **Αποφεύγουν τοπικά ακρότατα.**
   Το πλεονέκτημα των γενετικών αλγορίθμων έναντι των παραδοσιακών μεθόδων είναι ότι δεν ψάχνουν τη βέλτιστη λύση από ένα σημείο μόνο, αλλά από έναν πληθυσμό σημείων ταυτόχρονα, δηλαδή ερευνούν το διάστημα αναζήτησης από πολλά σημεία παράλληλα. Με αυτόν τον τρόπο μπορούν να αποφεύγουν τοπικά ακρότατα, σε αντίθεση με τις κλασικές τεχνικές βελτιστοποίησης όπως η "μείωση κλίσης" (gradient descent), οι οποίες μπορούν εύκολα να εγκλωβιστούν σε τοπικό ελάχιστο/μέγιστο.

3. **Δεν έχουν ιδιαίτερες αναλυτικές απαιτήσεις από τις συναρτήσεις που επεξεργάζονται.**
   Ο κύριος λόγος που συχνά καθιστά τις κλασσικές μεθόδους "δύσκαμπτες" έως και ακατάλληλες για την επίλυση προβλημάτων βελτιστοποίησης είναι η απαίτησή τους για ομαλότητα της αντικειμενικής συνάρτησης (τουλάχιστον δύο φορές συνεχώς παραγωγίσιμη), ύπαρξη Εσσιανού πίνακα κ.ο.κ. Αντιθέτως, οι γενετικοί αλγόριθμοι εφαρμόζονται και σε συναρτήσεις που δεν είναι λείες (συνεχείς και παραγωγίσιμες).

4. **Εκτελούν ταυτόχρονα εξερεύνηση του χώρου των λύσεων και εκμετάλλευση της ήδη υπάρχουσας πληροφορίας.**
   Με το τυχαίο ψάξιμο γίνεται καλή εξερεύνηση του χώρου αναζήτησης λύσεων (exploration), αλλά δε γίνεται εκμετάλλευση της πληροφορίας (exploitation). Αντίθετα, με τη μέθοδο της ανάβασης λόφου (hill climbing) πραγματοποιείται καλή εκμετάλλευση της πληροφορίας (π.χ τιμή της αντικειμενικής συνάρτησης) αλλά όχι ικανοποιητική εξερεύνηση. Συνήθως τα δύο αυτά χαρακτηριστικά είναι ανταγωνιστικά και το επιθυμητό είναι να συνυπάρχουν και τα δύο προς όφελος της βελτιστοποίησης. Οι γενετικοί αλγόριθμοι περιέχουν τόσο το στοιχείο της εντατικοποίησης (intensification), δηλαδή με τοπικά βήματα επιδιώκουν να φτάσουν στο βέλτιστο, αλλά ταυτόχρονα περιέχουν και το στοιχείο της διαφοροποίησης (diversification), δηλαδή εκτελούν δραστικά αλλαγές προκειμένου να ξεφύγουν από το βέλτιστο της γειτονιάς στην οποία βρίσκονται. Με αυτόν

τον τρόπο, επιτυγχάνουν το βέλτιστο συνδυασμό εξερεύνησης και εκμετάλλευσης πληροφορίας, γεγονός που τους καθιστά ιδιαίτερα αποδοτικούς και ελκυστικούς.

5. **Επιδέχονται παράλληλη υλοποίηση.**
Οι γενετικοί αλγόριθμοι μπορούν να δεχθούν παράλληλη υλοποίηση, κατά την οποία γίνεται καταμερισμός του αρχικού προβλήματος σε πολλά υπόπροβλήματα (ή αλλιώς διεργασίες). Έτσι, αυξάνεται η αποδοτικότητά τους σε μεγάλα και σύνθετα προβλήματα βελτιστοποίησης. Η βασική ιδέα πίσω από κάθε παράλληλο πρόγραμμα είναι ο διαχωρισμός μια εργασίας σε επιμέρους κομμάτια και η ταυτόχρονη επίλυση των επιμέρους κομματιών χρησιμοποιώντας πολλούς επεξεργαστές. Μερικές μέθοδοι παραλληλοποίησης χρησιμοποιούν ένα μοναδικό πληθυσμό, ενώ άλλες διαχωρίζουν τον πληθυσμό σε αρκετά απομονωμένους υποπληθυσμούς.
Ενδεικτικά, σύμφωνα με τη μέθοδο παραλληλοποίησης "Συντονιστή-εργάτη" (Master-slave) ο αλγόριθμος χρησιμοποιεί έναν πληθυσμό και ο υπολογισμός της συνάρτησης καταλληλότητας ή και η εφαρμογή των γενετικών τελεστών γίνεται από άλλα υπολογιστικά στοιχεία. Για παράδειγμα, τα υπολογιστικά στοιχεία αυτά μπορεί να είναι ένα υπολογιστής με πολλαπλούς επεξεργαστές, μια ομάδα υπολογιστών που βρίσκονται στο ίδιο δίκτυο (cluster) ή συνδυασμός αυτών. Η παραλληλοποίηση μειώνει τον πραγματικό χρόνο εκτέλεσης του προγράμματος και ταυτόχρονα έχει το πλεονέκτημα στη διαθέσιμη μνήμη *(Deepa, S.N. & Sivanandam, S.N. (2008). Introduction to Genetic Algorithms. Springer-Verlag Berlin Heidelberg).*

6. **Είναι εύκολα επεκτάσιμοι και εξελίξιμοι.**
Οι γενετικοί αλγόριθμοι τροποποιούνται σχετικά εύκολα και μπορούν να επεκταθούν ανάλογα με την κρίση του σχεδιαστή και το πρόβλημα που στοχεύουν να επιλύσουν. Είναι δηλαδή αρκετά "ευέλικτοι" με την έννοια ότι μπορούν να υποστούν αλλαγές προς όφελος της απόδοσης. Παραδείγματος χάρη, είδαμε ότι έχουν αναπτυχθεί αρκετές μέθοδοι επιλογής των γονέων από τις οποίες μπορεί να επιλέξει ο σχεδιαστής του προγράμματος. Σε πολλές εφαρμογές όμως έχουν προστεθεί και λειτουργίες στο γενετικό αλγόριθμο που δεν είναι εμπνευσμένες από τη φύση.

7. **Μπορούν να συνυπάρξουν σε υβριδικές μορφές με άλλες μεθόδους.**
Σε μερικά προβλήματα όπου άλλες τεχνικές βελτιστοποίησης έχουν μεγαλύτερη αποδοτικότητα, υπάρχει η δυνατότητα χρησιμοποίησης ενός υβριδικού σχήματος γενετικού αλγορίθμου ή ακόμη ένα τμήμα αυτού. Αυτό δείχνει και τη μεγάλη ευελιξία των γενετικών αλγορίθμων. Για παράδειγμα,

μπορεί να χρησιμοποιηθεί μια τεχνική εντός του γενετικού αλγορίθμου που να βελτιώνει την απόδοσή του ως προς την ικανότητα αναζήτησης. Όταν ο γενετικός αλγόριθμος ως μέθοδος ολικής βελτιστοποίησης συνδυαστεί με μια άλλη μέθοδο ειδικά σχεδιασμένη για το πρόβλημα ενδιαφέροντος που αναζητά τη βέλτιστη λύση τοπικά, τότε η συνολική ικανότητα αναζήτησης της βέλτιστης λύσης βελτιώνεται σημαντικά. Ένα τέτοιο παράδειγμα υβριδικού γενετικού αλγορίθμου αποτελεί ο "Μιμητικός αλγόριθμος" (Memetic algorithm) *(Hart, W.E., Krasnogor, N.& Smith, J.E. (2005). Memetic Evolutionary Algorithms. Recent Advances in Memetic Algorithms. Studies in Fuzziness and Soft Computing, vol 166. Springer-Verlag Berlin Heidelberg*).

Μειονεκτήματα

Παρά τα πλεονεκτήματα που παρουσιάζουν, οι γενετικοί αλγόριθμοι, ωστόσο, συνδέονται με ορισμένα μειονεκτήματα και για τον λόγο αυτόν δεν ενδείκνυνται ανεπιφύλακτα για την επίλυση οποιουδήποτε προβλήματος βελτιστοποίησης. Τα κυριότερα από αυτά είναι τα εξής:

1. **Δεν εγγυώνται την εύρεση ολικού ακροτάτου.**
   Λόγω της στοχαστικής φύσης που έχει ο γενετικός αλγόριθμος, δεν εξασφαλίζεται η εύρεση της ολικής βέλτιστης λύσης, αλλά μπορούν να επιστρέψουν μια λύση που είναι κοντά στη βέλτιστη.

2. **Απαιτούν μεγάλη υπολογιστική ισχύ.**
   Ο μεγάλος αριθμός πράξεων που εκτελεί ο γενετικός αλγόριθμος τόσο για τον υπολογισμό της συνάρτησης καταλληλότητας, όσο και για την αντικειμενική συνάρτηση σε κάθε επανάληψη, αλλά και η πολυπλοκότητα της εκάστοτε συνάρτησης προς υπολογισμό, συνεπάγεται αρκετό υπολογιστικό χρόνο αλλά και μνήμη. Αρκεί να σκεφτούμε ότι οι υπολογισμοί αυτών των δύο συναρτήσεων γίνονται επαναληπτικά για κάθε χρωμόσωμα του πληθυσμού και αποθηκεύονται προσωρινά στη μνήμη προκειμένου να γίνονται οι απαραίτητες συγκρίσεις ώστε να εξελίσσεται ο πληθυσμός.

3. **Δυσκολία στην εύρεση κατάλληλης παραμετροποίησης.**
   Λόγω του ότι ο γενετικός αλγόριθμος δέχεται αρκετές παραμέτρους κατά το ξεκίνημά του, είναι χρονοβόρο και αρκετές φορές δύσκολο να βρεθεί ο συνδυασμός των παραμέτρων που δίνει τα καλύτερα αποτελέσματα για το εκάστοτε πρόβλημα. Στην περίπτωση που δεν υπάρχει καμία ένδειξη για την πραγματική λύση του προβλήματος βελτιστοποίησης που επιλύουμε, η επιλογή της παραμετροποίησης είναι τυχαία και αυτό σημαίνει ότι τα αποτελέσματα του γενετικού αλγορίθμου ενδέχεται να φέρουν μεγάλο

σφάλμα. Στην παρούσα διπλωματική εργασία και για το πρόβλημα επιλογής μεταβλητών στο πολλαπλό γραμμικό μοντέλο παλινδρόμησης, οι παράμετροι που εισάγουμε στον αλγόριθμο είναι επτά, και προκειμένου να βρεθεί ο συνδυασμός που δίνει το κατάλληλο μοντέλο παλινδρόμησης χρειάστηκαν αρκετές δοκιμές. Εν ολίγοις, η κατάλληλη παραμετροποίηση στις ευριστικές μεθόδους προκύπτει εμπειρικά με τη μέθοδο δοκιμής και σφάλματος (trial and error). Ωστόσο υπάρχουν ειδικά πακέτα που έχουν σχεδιαστεί έτσι ώστε να βρίσκουν τον βέλτιστο συνδυασμό παραμέτρων για αλγορίθμους βελτιστοποίησης με κυριότερα τα "irace","REVAC" και "SPOT" *(Rudolf, M. (2016). Parameter tuning for numerical optimization algorithms. Czech Technical University in Prague. Dept of Computer Science and Engineering. Prague).*

4. **Δυσκολίες κατά τον σχεδιασμό τους.**
   Εκτός της δυσκολίας εντοπισμού κατάλληλης παραμετροποίησης, σε πολλά προβλήματα βελτιστοποίησης υπάρχει δυσκολία στην εύρεση της συνάρτησης καταλληλότητας που πρέπει να χρησιμοποιηθεί, ακόμα και στην επιλογή της κωδικοποίησης.

5. **Κίνδυνος πρόωρης σύγκλισης.**
   Όπως προαναφέρθηκε, οι εξελικτικοί αλγόριθμοι, όπως και ο γενετικός, επεξεργάζονται πολλές λύσεις ταυτόχρονα και όχι μια και μοναδική, όπως οι κλασσικές μέθοδοι βελτιστοποίησης (μέθοδος ανάβασης λόφου-hill climbing), γεγονός που βοηθά στην αποφυγή εγκλωβισμού σε τοπικά ακρότατα. Ωστόσο υπάρχει ο κίνδυνος της πρόωρης σύγκλισης, δηλαδή η τάση του αλγορίθμου να συγκλίνει σε κάποιο τοπικό βέλτιστο, λόγω του ότι έχει αγνοήσει περιοχές του χώρου των λύσεων.

Κεφάλαιο 3

ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

3.1 Πολλαπλό γραμμικό μοντέλο παλινδρόμησης

Το πολλαπλό γραμμικό μοντέλο είναι η επέκταση του απλού γραμμικού μοντέλου παλινδρόμησης, όπου μια τυχαία μεταβλητή η οποία θα ονομάζεται και εξαρτημένη μεταβλητή, εξαρτάται γραμμικά από ένα σύνολο ανεξάρτητων τυχαίων μεταβλητών.Αναλυτικότερα, θεωρούμε ότι διαθέτουμε p επεξηγηματικές μεταβλητές τις οποίες και συμβολίζουμε με τη βοήθεια ενός τυχαίου διανύσματος $\mathbf{X} = (X_1, ..., X_p)^T$ και υποθέτουμε ότι όλες είναι ποσοτικές και συνδέονται γραμμικά με την ποσοτική μεταβλητή απόκρισης Υ. Κατ'επέκταση του απλού γραμμικού μοντέλου, θεωρούμε ότι το διάνυσμα $\mathbf{X}$ επηρεάζει γραμμικά την τιμή της τυχαίας μεταβλητής Υ, έτσι ώστε να ισχύει:

$$Y = a + b_1 X_1 + b_2 X_2 + ... + b_p X_p + \varepsilon, \tag{3.1}$$

όπου $\varepsilon \sim N(0, \sigma^2)$ είναι το τυχαίο σφάλμα.
Για δοθέν $\mathbf{X} = \mathbf{x} = (x_1, ..., x_p)^T$ και αν λάβουμε υπόψη ότι η μέση τιμή του τυχαίου σφάλματος ε είναι μηδέν, τότε η δεσμευμένη μέση τιμή της τυχαίας μεταβλητής Υ δοθέντος του $\mathbf{X} = \mathbf{x}$ δίνεται από τη σχέση:

$$E[Y|\mathbf{X} = \mathbf{x}] = a + b_1 x_1 + b_2 x_2 + ... + b_p x_p \,. \tag{3.2}$$

Η σχέση (3.2) καλείται πολλαπλό γραμμικό μοντέλο, αφού έχουμε περισσότερες από μια επεξηγηματικές μεταβλητές και η συνάρτηση που περιγράφει τη σχέση είναι γραμμική.

Αντιστοίχως, για το τυχαίο δείγμα $(Y_1, X_{11}..., X_{1p})^T, ..., (Y_1, X_{n1}..., X_{np})^T$ μεγέθους n ισχύει η σχέση:

$$Y_i = a + b_1 X_{i1} + b_2 X_{i2} + ... + b_p X_{ip} + \varepsilon_i, \tag{3.3}$$

όπου i=1,2,...n και $\varepsilon_i \sim N(0, \sigma^2)$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές που ονομάζονται τυχαία σφάλματα. Η ποσότητα σ^2 εκφράζει τη διασπορά των σφαλμάτων, η οποία θεωρείται σταθερή ανεξάρτητα της τιμής του τυχαίου διανύσματος $\mathbf{X}$. Το πολλαπλό γραμμικό μοντέλο της σχέσης (3.2) μπορεί να παρασταθεί με χρήση διανυσμάτων ως εξής:

$E[Y|\mathbf{X} = \mathbf{x}] = \tilde{\mathbf{x}}\mathbf{b}$, με $\tilde{\mathbf{x}} = (1, x_1, ..., x_p)^T$ και $\mathbf{b} = (a, b_1, ..., b_p)^T$

και η σχέση (3.3) με τη χρήση πινάκων γράφεται:

$\mathbf{Y} = \tilde{\mathbf{X}}\mathbf{b} + \epsilon$ με $\mathbf{Y} = (Y_1, ..., Y_n)^T$, $\mathbf{b} = (a, b_1, ..., b_p)^T$, $\epsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ και

$$\tilde{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}.$$

Οι σταθερές ποσότητες a, b_1,. . . , b_p ονομάζονται συντελεστές του μοντέλου και είναι πραγματικοί αριθμοί. Οι συντελεστές αυτοί μαζί με την άγνωστη διασπορά σ^2 του τυχαίου σφάλματος αποτελούν τους p+1 αγνώστους του μοντέλου, ονομάζονται συντελεστές του μοντέλου (model coefficients) και είναι πραγματικοί αριθμοί. Η παράμετρος b_j για j=1,2,. . . p εκφράζει το πόσο αναμένεται να μεταβληθεί (αυξηθεί ή μειωθεί) η τιμή της τυχαίας μεταβλητής Υ, αν η τυχαία μεταβλητή X_j αυξηθεί κατά μία μονάδα και υπό την προϋπόθεση ότι όλες οι άλλες επεξηγηματικές μεταβλητές διατηρούνται σταθερές. Το πρόσημο b_j του συντελεστή προσδιορίζει τη σχέση εξάρτησης μεταξύ των τυχαίων μεταβλητών Υ και X_j όταν όλες οι υπόλοιπες επεξηγηματικές μεταβλητές X_k , $k \neq j$ παραμείνουν σταθερές, δηλαδή δηλώνει την αύξηση ή την μείωση αντίστοιχα στη μεταβλητή απόκρισης.
Με τη βοήθεια των παρατηρήσεων που έχουμε στη διάθεσή μας $(y_1, x_{11}, ..., x_{1p})^T$

$...(y_n, x_{1n}, ...x_{np})^T$ εκτιμούμε τους άγνωστους συντελεστές του μοντέλου, με αποτέλεσμα να καταλήγουμε στην εκτιμώμενη ευθεία της πολλαπλής γραμμικής παλινδρόμησης, η οποία είναι η παρακάτω:

$$\hat{Y} = \hat{a} + \hat{b_1}x_1 + ... + \hat{b_p}x_n.$$

Η τυχαία μεταβλητή $\hat{Y}$ ονομάζεται προβλεπόμενη τιμή (predicted value) της τυχαίας μεταβλητής Υ με βάση το μοντέλο παλινδρόμησης που περιγράψαμε και είναι ίση με τη δεσμευμένη μέση τιμή της τυχαίας μεταβλητής Υ όταν $\mathbf{X} = \mathbf{x}$. Γενικεύοντας, για κάθε i=1,. . . ,n υπολογίζουμε τις παρατηρούμενες προβλεπόμενες τιμές ή προσαρμοσμένες τιμές (fitted values), δηλαδή τις προβλεπόμενες τιμές με βάση το δείγμα που διαθέτουμε, από τη σχέση $\hat{y}_i = \hat{a}+\hat{b_1}x_{i1}+...+\hat{b_p}x_{1p}$. Συμπερασματικά, οι ποσότητες $\hat{\varepsilon}_i = y_i - \hat{y}_i$ καλούνται υπόλοιπα (residuals), είναι δηλαδή οι εκτιμήσεις των τυχαίων σφαλμάτων ε_i *(Φουσκάκης, Δ. (2013). Ανάλυση Δεδομένων με Χρήση της R, Εκδόσεις Τσότρας, Αθήνα).*
Τέλος, για την εκτίμηση των συντελεστών του μοντέλου χρησιμοποιούμε τη μέθοδο ελαχίστων τετραγώνων (least squares method). Αποδεικνύεται ότι οι εκτιμητές δίνονται από τη σχέση:

$$\hat{\mathbf{b}} = (\tilde{X}^T \cdot \tilde{X})^{-1}\tilde{X}^T \cdot \mathbf{y}. \tag{3.4}$$

Για να ισχύει η σχέση (3.4), πρέπει οι επεξηγηματικές μεταβλητές να είναι γραμμικά ανεξάρτητες έτσι ώστε ο πίνακας να είναι αντιστρέψιμος. Σε αντίθετη περίπτωση, έχουμε το πρόβλημα της πολυσυγγραμμικότητας (multicollinearity) το οποίο θα αναλυθεί στην Ενότητα 3.4.

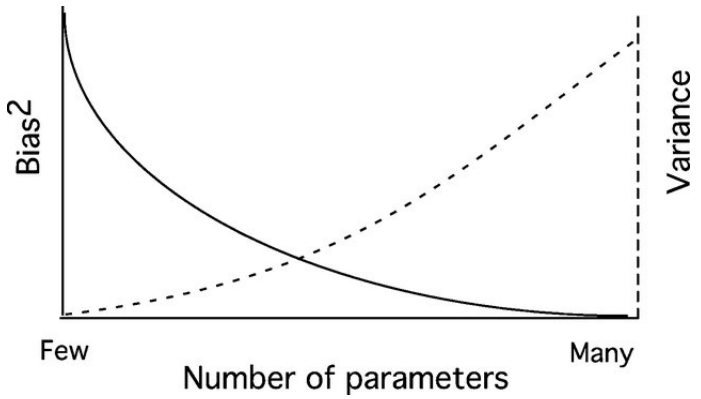
3.2 Επιλογή μεταβλητών στην παλινδρόμηση

Η επιλογή μεταβλητών είναι ένας τρόπος να μειωθούν οι μεταβλητές που συμμετέχουν στην κατασκευή ενός μοντέλου παλινδρόμησης κι επομένως να μειωθεί η πολυπλοκότητα του μοντέλου. Αποδεικνύεται ιδιαίτερα χρήσιμη προκειμένου να αποφευχθεί και το φαινόμενο της υπερπροσαρμογής (overfitting). Με τον όρο "υπερπροσαρμογή στα δεδομένα" καλούμε το φαινόμενο κατά το οποίο ένα μοντέλο, εν προκειμένω παλινδρόμησης, ενώ προσαρμόζεται πολύ ικανοποιητικά στο σύνολο δεδομένων που διαθέσαμε για την κατασκευή του (εκπαίδευση), παρουσιάζει μεγαλύτερο σφάλμα (χειρότερη απόδοση) όταν καλείται να διαχειριστεί ένα διαφορετικό σύνολο δεδομένων. Αυτό σημαίνει ότι το μοντέλο που κατασκευάσαμε παρουσιάζει πρόβλημα γενίκευσης, δηλαδή εξαρτάται σε μεγάλο βαθμό από τα δεδομένα που του δόθηκαν. Με τον όρο "επιλογή μεταβλητών" εννοούμε την εύρεση ενός κατάλληλου γνήσιου υποσυνόλου των επεξηγηματικών μεταβλητών του προβλήματος για την πρόβλεψη της μεταβλητής απόκρισης με υψηλή ακρίβεια. Απαντά στο ερώτημα: Αν προσθέσουμε την i-οστή μεταβλητή στο μοντέλο, αυξάνεται η αποδοτικότητά του σημαντικά; Σε

περίπτωση που αυτό δε συμβαίνει, τότε παραλείπουμε αυτή τη μεταβλητή διότι καθιστά το μοντέλο πιο πολύπλοκο ως προς τους υπολογισμούς και την ερμηνεία.
Είναι πολύ μεγάλο το πλήθος των δυνατών γραμμικών μοντέλων που μπορεί να προκύψει αν αναζητήσουμε όλους τους συνδυασμούς των επεξηγηματικών μεταβλητών που μπορούν να συμπεριληφθούν σε κάθε γραμμικό μοντέλο. Για p τυχαίες μεταβλητές, το πλήθος των γραμμικών μοντέλων που προκύπτουν είναι 2^p που σημαίνει ότι για προβλήματα γραμμικής παλινδρόμησης το πλήθος των δυνατών μοντέλων μεγαλώνει εκθετικά σε σχέση με το πλήθος των τυχαίων μεταβλητών. Για παράδειγμα, για 15 μεταβλητές έχουμε να εξετάσουμε $2^{15} = 32.768$ μοντέλα, και εδώ φαίνεται η αναγκαιότητα εναλλακτικών τρόπων εύρεσης του βέλτιστου μοντέλου, όπως ένας από αυτός είναι και ο γενετικός αλγόριθμος.
Συμπερασματικά, αναζητούμε απλά μοντέλα με μεγάλη προβλεπτική ικανότητα τα οποία να περιέχουν τόσες επεξηγηματικές μεταβλητές όσες είναι απαραίτητες για την επεξήγηση της Υ. Η ύπαρξη μικρού πλήθους ανεξάρτητων μεταβλητών σε ένα γραμμικό μοντέλο μπορεί να δημιουργήσει πρόβλημα μεροληψίας στο γραμμικό μοντέλο, καθώς λίγες τυχαίες μεταβλητές μπορεί να αλλοιώσουν τις προβλέψεις της παλινδρόμησης, ειδικά αν οι παρατηρήσεις που διαθέτουμε είναι λίγες. Αντιθέτως, η ύπαρξη μεγάλου αριθμού επεξηγηματικών μεταβλητών σε ένα γραμμικό μοντέλο μπορεί να δημιουργήσει πρόβλημα στην ακρίβεια των εκτιμήσεων, δηλαδή να αυξήσει τη διασπορά των εκτιμήσεων των παραμέτρων του μοντέλου και το μοντέλο να παρουσιάζει πρόβλημα γενίκευσης. Το πρόβλημα αυτό αναπαρίσταται από το ακόλουθο διάγραμμα και είναι χαρακτηριστικό στο σχεδιασμό ενός πολλαπλού γραμμικού μοντέλου.
Στόχος είναι να βρεθεί ένα γραμμικό μοντέλο με περισσότερες από μία επεξηγηματικές μεταβλητές, για να είναι πιο κοντά στην πραγματική τιμή η πρόβλεψη της Υ από τις X_i, το οποίο όμως να αποτελείται από όσο το δυνατόν μικρότερο αριθμό επεξηγηματικών μεταβλητών. Η παραπάνω πρόταση συνοψίζεται ως "Η Αρχή της Φειδωλότητας" (Principle of Parsimony). Στο Σχήμα 3.1 φαίνεται με παραστατικό τρόπο η Αρχή της φειδωλότητας και το ισοζύγιο μεροληψίας και διασποράς.

Τα πολλαπλά γραμμικά μοντέλα παλινδρόμησης, στα πλαίσια του προβλήματος της επιλογής μεταβλητών το οποίο μελετάμε, μπορούν εύκολα να παρασταθούν με χρήση δυαδικών διανυσμάτων μήκους p, όπου p ο αριθμός των επεξηγηματικών μεταβλητών. Πιο συγκεκριμένα, όλα τα υποψήφια μοντέλα αντιστοιχούν σε διανύσματα στα οποία η κάθε συντεταγμένη λαμβάνει την τιμή ένα, εάν η συγκεκριμένη μεταβλητή συμπεριλαμβάνεται στο μοντέλο, και μηδέν εάν δεν συμπεριλαμβάνεται. Η αναπαράσταση αυτή είναι ιδιαίτερα βολική για το συγκεκριμένο πρόβλημα και ταιριάζει με το διακριτό χαρακτήρα στοχαστικής

βελτιστοποίησης του προβλήματος. Τα εν λόγω διανύσματα σμυβολίζονται με $\vec{\gamma}$, και ένα τέτοιο διάνυσμα θα μπορούσε να είναι το:
$\vec{\gamma} \leftarrow (1,0,0,1,0,1,0,0,0,0,1,1,0,0,0)$, το οποίο αντιστοιχεί στο γραμμικό μοντέλο με την 1η, την 4η, την 6η, την 11η και τη 12η μεταβλητή.



Σχήμα 3.1: Η Αρχή της φειδωλότητας

3.3 Διαδικασίες επιλογής μεταβλητών με βήματα

Καθοριστικό ρόλο στη σύγκριση και επιλογή μοντέλων παίζει το άθροισμα τετραγώνων των υπολοίπων, δηλαδή το

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Η μεταβολή του SSE οδηγεί στον έλεγχο F ο οποίος είναι:

$$F = \frac{(SSE_0 - SSE_1)/q}{SSE_1/(n-p)} \sim F_{q,(n-p)},$$

όπου SSE_0 το άθροισμα τετραγώνων των υπολοίπων υπό την H_0: $b_1 = b_2 = \ldots = b_q = 0$, b_i, $i \neq 1, 2, .., q$ χωρίς περιορισμούς (μοντέλο M_0), ένω το SSE_1

το άθροισμα τετραγώνων των υπολοίπων υπό την H_1: όλα τα b_i χωρίς περιορισμούς (μοντέλο M_1). Επίσης n ο αριθμός των παρατηρήσεων, q ο αριθμός των μεταβλητών στις οποίες διαφέρουν τα δύο προς εξέταση μοντέλα, και p ο αριθμός των παραμέτρων υπό την H_1.
Η πρόσθεση μιας ή περισσότερων μεταβλητών σε ένα στατιστικό μοντέλο οδηγεί στον ίδιο ή σε ακόμα μεγαλύτερο βαθμό επεξήγησης της μεταβλητότητας της μεταβλητής απόκρισης. Αυτό πρακτικά σημαίνει ότι το άθροισμα τετραγώνων των υπολοίπων θα είναι μικρότερο από πριν. Συνεπώς, αν διαθέτουμε ένα σύνολο από υποψήφιες μεταβλητές προς εισαγωγή στο μοντέλο, τότε εξετάζουμε την εισαγωγή εκείνης της μεταβλητής που δίνει τη μεγαλύτερη διαφορά $SSE_0 - SSE_1$ διότι αν αυτή η μείωση δεν είναι στατιστικά σημαντική, τότε δε θα είναι ούτε και οι άλλες. Από την άλλη πλευρά, η αφαίρεση μιας ή περισσότερων μεταβλητών από ένα στατιστικό μοντέλο έχει ως αποτέλεσμα το νέο μοντέλο να εξηγεί σε μικρότερο ή στην καλύτερη περίπτωση στον ίδιο βαθμό τη συμπεριφορά των τιμών της εξαρτημένης μεταβλητής, και άρα το άθροισμα των τετραγώνων των υπολοίπων να αυξάνεται. Βέβαια, αν η μεταβλητή ή οι μεταβλητές που αφαιρούμε δεν είναι στατιστικά σημαντικές, η μεταβολή του SSE παίρνει τιμές κοντά στο μηδέν.

Μέθοδοι Επιλογής Μεταβλητών
Με βάση τα παραπάνω, ορίζονται τρεις μέθοδοι οι οποίες ακολουθούνται για την ανάπτυξη του βέλτιστου και οικονομικότερου μοντέλου. Οι μέθοδοι αυτές είναι οι εξής:

- Διαδικασία της διαδοχικής αφαίρεσης (Backward Elimination)
- Διαδικασία της διαδοχικής πρόσθεσης (Forward Selection)
- Κατά βήματα εμπρός-πίσω επιλογή (Stepwise Selection)

Διαδικασία της διαδοχικής αφαίρεσης

Η μέθοδος αυτή ξεκινάει συμπεριλαμβάνοντας στο μοντέλο όλες τις διαθέσιμες επεξηγηματικές μεταβλητές και αφαιρεί διαδοχικά μία προς μία τις μεταβλητές, αρχίζοντας από αυτή που δίνει τη μικρότερη αύξηση του SSE και εφόσον ο έλεγχος F που εκτελείται για τη συγκεκριμένη μεταβλητή είναι στατιστικά μη σημαντικός. Αυτή θα είναι και η μεταβλητή με τη μικρότερη συμβολή στο μοντέλο. Πιο συγκεκριμένα, τα βήματα αυτής της μεθόδου είναι τα ακόλουθα:
Β1. Εισάγουμε όλες τις διαθέσιμες επεξηγηματικές μεταβλητές στο μοντέλο.
Β2. Αφαιρούμε τη λιγότερο σημαντική μεταβλητή, αυτή δηλαδή που δε μεταβάλλει

σημαντικά την τιμή της διαφοράς $SSE_0 - SSE_1$ που είναι μέρος του αριθμητή του στατιστικού ελέγχου F.
Β3. Προσαρμόζουμε εκ νέου ένα μοντέλο παλινδρόμησης στα δεδομένα, παραλείποντας τη μεταβλητή που αφαιρέσαμε στο Β2.
Β4. Επαναλαμβάνουμε τα Β2 και Β3 μέχρις ότου η αφαίρεση μιας οποιασδήποτε από τις μεταβλητές να είναι στατιστικά σημαντική, οπότε και σταματάμε τη διαδικασία.

Διαδικασία διαδοχικής πρόσθεσης

Η μέθοδος αυτή έχει ως αφετηρία το μοντέλο $y = b_0$ και προσθέτει κάθε φορά την επεξηγηματική μεταβλητή εκείνη, η οποία μας δίνει τη μεγαλύτερη στατιστικά σημαντική τιμή της ελεγχοσυνάρτησης F. Έτσι, η μέθοδος αυτή μπορεί να συνοψιστεί από τα ακόλουθα βήματα:
Β1. Ξεκινάμε από το μοντέλο που περιέχει μόνο τον σταθερό όρο, δηλαδή το $y = b_0$ και μάλιστα ισχύει ότι $\hat{b_0} = \hat{y}$.
Β2. Προσθέτουμε τη μεταβλητή εκείνη, η οποία δίνει τη μεγαλύτερη στατιστικά σημαντική τιμή του ελέγχου F, ισοδύναμα εκείνη που συνεπάγεται τη μεγαλύτερη μείωση του SSE και άρα συμβάλλει περισσότερο στην επεξήγηση της Y σε σχέση με τις υπόλοιπες μεταβλητές.
Β3. Προσαρμόζουμε εκ νέου ένα μοντέλο παλινδρόμησης συμπεριλαμβάνοντας τη μεταβλητή που προσθέσαμε στο Β2, και διερευνάμε ποια θα είναι η επόμενη μεταβλητή που θα δώσει στατιστικά σημαντική μείωση του SSE σε σχέση με το μοντέλο που περιέχει μόνο την πρώτη μεταβλητή και έτσι την εισάγουμε στο μοντέλο.
Β4. Επαναλαμβάνουμε τα βήματα Β2 και Β3 μέχρις ότου η τιμή του ελέγχου F για την πρόσθεση οποιασδήποτε από τις εναπομείνασες μεταβλητές να μην είναι στατιστικά σημαντική, οπότε και σταματάμε τη διαδικασία.

Κατά βήματα εμπρός-πίσω επιλογή

Η μέθοδος αυτή αποτελεί μια "διορθωμένη" εκδοχή της διαδικασίας διαδοχικής πρόσθεσης, καθώς παρεμβάλλεται ένας επιπλέον έλεγχος σε κάθε επανάληψη της διαδικασίας διαδοχικής πρόσθεσης μιας μεταβλητής. Ο εν λόγω έλεγχος εξετάζει την περίπτωση στην οποία η πρόσθεση μιας νέας μεταβλητής στο μοντέλο οδηγεί στην εξασθένηση της στατιστικής σημαντικότητας κάποιας άλλης μεταβλητής, που είχε εισαχθεί σε προγενέστερο στάδιο, με συνέπεια να πρέπει να εξεταστεί η αξία της παραμονής της στο μοντέλο.

3.4 Αξιολόγηση μοντέλου παλινδρόμησης

Υπάρχουν αρκετοί τρόποι για να ελέγξουμε την αποτελεσματικότητα ενός γραμμικού μοντέλου παλινδρόμησης στην πρόβλεψη και την εξήγηση της μεταβλητής απόκρισης από τις επεξηγηματικές μεταβλητές που χρησιμοποιήθηκαν στο μοντέλο. Παρακάτω, αναλύονται οι πιο διαδεδομένοι στη βιβλιογραφία.

Συντελεστής Προσδιορισμού R^2

Η ποσότητα

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

καλείται συντελεστής προσδιορισμού (coefficient of determination) και εκφράζει το ποσοστό της μεταβλητότητας της τυχαίας μεταβλητής Y που περιγράφεται μέσω του μοντέλου παλινδρόμησης, δηλαδή φανερώνει το βαθμό στον οποίο η τυχαία μεταβλητή $\mathbf{X}$ ερμηνεύει με τη βοήθεια της γραμμικής παλινδρόμησης τη μεταβλητότητα της τυχαίας μεταβλητής Υ. Οι τιμές τις οποίες μπορεί να λάβει αυτός ο συντελεστής είναι στο διάστημα [0,1]. Ωστόσο, ο συντελεστής αυτός είναι καλό να μη χρησιμοποιείται ως μέτρο καλής προσαρμογής του μοντέλου στα δεδομένα, ή ακόμη και ως μέτρο σύγκρισης δύο μοντέλων διότι αν σε ένα γραμμικό μοντέλο προσθέσουμε μια επεξηγηματική μεταβλητή με ελάχιστη συνεισφορά στη μείωση της αβεβαιότητας ως προς τη μεταβλητή απόκρισης, τότε ο συντελεστής προσδιορισμού θα αυξηθεί δίνοντάς μας έτσι την εσφαλμένη εντύπωση ότι το νέο μοντέλο είναι περισσότερο κατάλληλο.
Για το λόγο αυτό, είναι προτιμότερο να υπολογίζουμε τον προσαρμοσμένο συντελεστή προσδιορισμού (adjusted coefficient of determination) ο οποίος εκτελεί την ίδια δουλειά με τον συντελεστή προσδιορισμού που αναλύσαμε ήδη και επιπρόσθετα λαμβάνει υπόψη και την πολυπλοκότητα του μοντέλου, δηλαδή τον αριθμό των επεξηγηματικών μεταβλητών. Κατά συνέπεια, αν προσθέσουμε μια τυχαία μεταβλητή στο μοντέλο, ενδέχεται το νέο πιο σύνθετο μοντέλο να έχει μικρότερη τιμή στον προσαρμοσμένο συντελεστή προσδιορισμού σε σχέση με το αρχικό απλούστερο μοντέλο. Η σχέση που συνδέει τους δύο συντελεστές είναι η εξής:

$$\tilde{R^2} = R^2 - \frac{(1-R^2)\cdot p}{n-p-1},$$

όπου p ο αριθμός των επεξηγηματικών μεταβλητών και n ο αριθμός των παρατηρήσεων.

Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error)

Το μέσο τετραγωνικό σφάλμα (MSE) αποτελεί ίσως το πιο αντικειμενικό κριτήριο καλής προσαρμογής ενός μοντέλου παλινδρόμησης. Η ποσότητα αυτή εκτιμάει την άγνωστη διασπορά του τυχαίους σφάλματος και αποτελεί εκτιμητή του. Δίνεται από τη σχέση:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Η θετική τετραγωνική ρίζα του MSE καλείται τυπικό σφάλμα της παλινδρόμησης (standard error of regression) και όσο μικρότερη είναι η τιμή της, τόσο καλύτερη προσαρμογή έχουμε στα δεδομένα. Αξίζει να αναφερθεί ότι συχνά, αντί για το MSE υπολογίζουμε την τετραγωνική του ρίζα, δηλαδή το RMSE (Root Mean Squared Error), διότι λαμβάνει τιμές στην ίδια κλίμακα με τη μεταβλητή απόκρισης Y και έτσι μας καλύτερη εικόνα για το μέγεθος του σφάλματος του μοντέλου που κατασκευάστηκε.

Akaike Information Criterion (AIC)

Το AIC αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου με όσο το δυνατόν μικρότερο αριθμό παραμέτρων. Στη γενική περίπτωση, ορίζεται από τη σχέση:

$$AIC(\vec{\gamma}) = 2 \cdot d_{\vec{\gamma}} - 2 \cdot lnL_{\vec{\gamma}},$$

όπου $d_{\vec{\gamma}}$ το πλήθος των παραμέτρων του μοντέλου $\vec{\gamma}$ και $L_{\vec{\gamma}}$ η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμηθέν μοντέλο.
Το προτιμητέο μοντέλο με βάση αυτό το κριτήριο είναι εκείνο με το μικρότερο AIC. Είναι γνωστό ότι η εισαγωγή επιπλέον μεταβλητών στο μοντέλο βελτιώνει την προσαρμογή του μοντέλου στα δεδομένα, ανεξάρτητα από το αν αυτές είναι στατιστικά σημαντικές ή όχι. Αυτό σημαινει ότι η προσθήκη μεταβλητών στο μοντέλο αυξάνει τον όρο $lnL_{\vec{\gamma}}$, άρα ο δεύτερος όρος του κριτηρίου AIC μειώνεται. Όμως, ταυτόχρονα αυξάνεται ο πρώτος όρος του. Τελικά, η εισαγωγή επιπλέον παραμέτρων στο μοντέλο μειώνει την τιμή του AIC μόνο αν αυτές βελτιώνουν την προσαρμογή του μοντέλου σε βαθμό που υπερβαίνει τον αυξημένο πρώτο όρο, δηλαδή το γινόμενο $2 \cdot d_{\vec{\gamma}}$.

Bayesian Information Criterion (BIC)

Το BIC προτάθηκε από τον Schwartz (1978) και αποτελεί ένα ακόμη κριτήριο επιλογής του βέλτιστου μοντέλου παλινδρόμησης ανάμεσα σε μοντέλα με διαφορετικό αριθμό παραμέτρων, όπως και το AIC. Η βασική τους διαφορά έγκειται στο ότι η εισαγωγή επιπρόσθετων μεταβλητών αποθαρρύνεται σε μεγαλύτερο βαθμό από το BIC, διότι περιέχει έναν όρο ποινικοποίησης για κάθε προσθήκη μεταβλητής. Στη γενική περίπτωση, ορίζεται από τη σχέση:

$$BIC(\vec{\gamma}) = -2 \cdot lnL_{\vec{\gamma}} + d_{\vec{\gamma}} \cdot ln(n),$$

Όπου και πάλι $d_{\vec{\gamma}}$ είναι το πλήθος των μεταβλητών που συμμετέχουν στο μοντέλο και n το πλήθος των παρατηρήσεων (*Heuvel, E., Romeijn, J.W. & Wit, E. (2012). All models are wrong...': an introduction to model uncertainty. Statistica Neerlandica, 66(3), 217-236*).
Το κριτήριο αυτό θα χρησιμοποιηθεί στην παρούσα διπλωματική εργασία ως αντικειμενική συνάρτηση στο πρόβλημα βελτιστοποίησης που θα επιλύσουμε.

Πολυσυγγραμμικότητα

Με τον όρο πολυσυγγραμμικότητα (multicollinearity) εννοούμε την ύπαρξη έντονης συσχέτισης μεταξύ δύο ή περισσότερων επεξηγηματικών μεταβλητών X_i, ισοδύναμα αυτό σημαίνει ότι μια επεξηγηματική μεταβλητή είναι γραμμικά συσχετισμένη με άλλες επεξηγηματικές μεταβλητές με την έννοια ότι εκφράζεται ως γραμμικός συνδυασμός τους. Με το μαθηματικό φορμαλισμό, αυτό γράφεται ως εξής:

$$x_j = \sum_{i \neq j} c_i \cdot x_i + d.$$

Με άλλα λόγια, το φαινόμενο της πολυσυγραμμικότητας προκύπτει όταν η μια μεταβλητή προβλέπεται σε μεγάλο βαθμό όταν γνωρίζουμε την τιμή μιας άλλης μεταβλητής, έτσι ώστε ουσιαστικά μας παρέχουν τις ίδιες πληροφορίες και ο διαχωρισμός των επιδράσεών τους καθίσταται δύσκολος. Αποτελεί συχνό φαινόμενο, ιδιαίτερα σε δεδομένα που προέρχονται από οικονομικές και κοινωνικές μελέτες, διότι συνήθως υπάρχει αλληλεξάρτηση κοινωνικών χαρακτηριστικών

και οικονομικών μεγεθών. Η παρουσία πολυσυγγραμμικότητας οδηγεί σε αυξημένα τυπικά σφάλματα των συντελεστών παλινδρόμησης $\hat{\mathbf{b}}$ με αποτέλεσμα να οδηγεί σε ασταθείς εκτιμήσεις σχετικά με την επίδραση της κάθε επεξηγηματικής μεταβλητής στην μεταβλητή απόκρισης. Αξίζει να σημειωθεί ότι σε αυτήν την περίπτωση καθίσταται δύσκολος ο εντοπισμός των στατιστικά σημαντικών μεταβλητών διότι η τιμή του στατιστικού ελέγχου (ελεγχοσυνάρτηση) $t = \frac{\hat{\mathbf{b}}}{se(\hat{\mathbf{b}})}$ υπό την $H_0 : \mathbf{b} = 0$ είναι αρκετά μικρή. Η πολυσυγγραμμικότητα είναι μια από τις κυριότερες αιτίες που ευθύνεται για την εξαγωγή λανθασμένων συμπερασμάτων στην ανάλυση του πολλαπλού γραμμικού μοντέλου, λόγω της συνεπαγόμενης αύξησης των τυπικών σφαλμάτων των συντελεστών *(Καρώνη, Χ. & Οικονόμου, Π. (2010). Στατιστικά Μοντέλα Παλινδρόμησης. Εκδόσεις Συμεών. Αθήνα).*

Κεφάλαιο 4

Ο ΓΕΝΕΤΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ ΣΤΟ ΠΡΟΒΛΗΜΑ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

Στο παρόν κεφάλαιο θα αναλυθεί πλήρως η δομή και οι επιμέρους λειτουργίες που εκτελούνται στον γενετικό αλγόριθμο προκειμένου να επιλύσει το πρόβλημα της επιλογής μεταβλητών σε μοντέλα παλινδρόμησης.

4.1 Δομή και ανάλυση αλγορίθμου

4.1.1 Εξωτερικές μεταβλητές

Οι παράμετροι-μεταβλητές που δέχεται η συνάρτηση "genetic" για να ξεκινήσει ο αλγόριθμος παρουσιάζονται παρακάτω:

genetic ← function(X, Y, $pop_{size} = 50$, eval=evaluation, maxiter=25,$p_{cross} = 1$, $type_{cross} = 1$, elitism=TRUE, $p_{old} = 0.30$, $p_{mut} = 0.1$)
Οι τιμές που δόθηκαν είναι ενδεικτικές μιας δοκιμής του γενετικού αλγορίθμου και ο χρήστης μπορεί να δώσει άλλες τιμές.

- X
  Είναι ένας πίνακας διάστασης n x p, όπου n ο αριθμός των παρατηρήσεων και p ο αριθμός των επεξηγηματικών μεταβλητών. Για το πρώτο σύνολο

δεδομένων με βάση το οποίο και σχεδιάστηκε ο γενετικός αλγόριθμος, οι μεταβλητές κατασκευάζονται έτσι ώστε να είναι ανεξάρτητες, ενώ στη συνέχεια θα κατασκευαστούν σύνολα δεδομένων στα οποία οι p επεξηγηματικές μεταβλητές παρουσιάζουν συσχετίσεις μεταξύ τους. Οι γραμμές του πίνακα X αντιστοιχούν στις n παρατηρήσεις κάθε μεταβλητής, ενώ οι στήλες αντιστοιχούν στις p διαφορετικές μεταβλητές για τις οποίες ο γενετικός αλγόριθμος αναζητά το βέλτιστο συνδυασμό για το γραμμικό μοντέλο παλινδρόμησης. Ενδεικτικά, το (i,j) στοιχείο του πίνακα αντιστοιχεί στην i-οστή παρατήρηση της j-οστής μεταβλητής , όπου το i μπορεί να πάρει τιμές από το 1 έως και το n, ενώ το j από το 1 έως και το p.

- Y
  Είναι ένα διάνυσμα μήκους n, οι τιμές του οποίου αντιστοιχούν στις παρατηρήσεις για την εξαρτημένη μεταβλητή (μεταβλητή απόκρισης).

- pop_{size}
  Πρόκειται για την παράμετρο του γενετικού που δηλώνει το μέγεθος του πληθυσμού. Διατηρείται σταθερή σε όλο τον αλγόριθμο και αυτό σημαίνει ότι σε κάθε επανάληψη ο αλγόριθμος "κρατάει" τόσα χρωμοσώματα, όσα και η τιμή αυτής της παραμέτρου. Σε όλες τις δοκιμές, η τιμή αυτή επιλέχθηκε να είναι ίση με $N = 50$.

- eval
  Είναι η συνάρτηση που καλεί ο γενετικός αλγόριθμος για να υπολογίσει την "καταλληλότητα" του κάθε χρωμοσώματος. Στη συνάρτηση αυτή, γίνεται ο υπολογισμός της αντικειμενικής συνάρτησης με βάση την οποία επιλύεται το πρόβλημα επιλογής μεταβλητών στην παλινδρόμηση. Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, επιλέχθηκε η συνάρτηση BIC ως η συνάρτηση αξιολόγησης. Η δομή της θα αναλυθεί παρακάτω.

- maxiter
  Η παράμετρος αυτή δηλώνει τον αριθμό των επαναλήψεων που θα εκτελέσει ο αλγόριθμος μέχρι να τερματιστεί. Ισοδύναμα, με όρους γενετικού αλγορίθμου, η παράμετρος αυτή αντιστοιχεί στον αριθμό των γενεών που παράγονται κατά τη διάρκεια του αλγορίθμου. Σε αρκετές δοκιμές δόθηκε η τιμή 25 σε αυτήν την παράμετρο, ωστόσο διαπιστώνεται ότι ο αλγόριθμος συγκλίνει σε μία λύση σε αρκετά μικρότερο αριθμό επαναλήψεων.

- p_{cross}
  Η μεταβλητή αυτή δηλώνει την πιθανότητα να γίνει μια διασταύρωση μεταξύ των χρωμοσωμάτων του πληθυσμού. Σε περίπτωση που δεν πραγματοποιηθεί διασταύρωση μεταξύ δύο "γονέων", αυτοί περνούν στην επό-

μενη γενιά, ελαφρώς μεταλλαγμένοι, σε περίπτωση που συμβεί μετάλλαξη. Η πιθανότητα διασταύρωσης παίρνει διαφορετική τιμή κατά το πέρας των επαναλήψεων και συγκεκριμένα μετά την 11η επανάληψη γίνεται μικρότερη.

- $type_{cross}$
  Η μεταβλητή αυτή δηλώνει τον τρόπο με τον οποίο επιλέγει ο χρήστης να εκτελείται η διασταύρωση. Μπορεί να πάρει 3 διακριτές τιμές που η καθεμία αντιστοιχεί σε διαφορετική μέθοδο διασταύρωσης. Συγκεκριμένα, για $type_{cross} = 1$, ο χρήστης επιλέγει να πραγματοποιείται διασταύρωση ενός σημείου (one-point crossover), για $type_{cross} = 2$ πραγματοποιείται διασταύρωση δύο σημείων (two-point crossover) ενώ για $type_{cross} = 3$ πραγματοποιείται ομοιόμορφη διασταύρωση (uniform crossover) με τη βοήθεια μάσκας όπως έχει αναλυθεί σε προηγούμενο κεφάλαιο. Η τιμή της παραμέτρου αυτής επιλέγεται να παραμένει σταθερή σε όλο τον αλγόριθμο, έτσι ώστε μετά από πολλές δοκιμές να καταλήξουμε σε έναν τύπο διασταύρωσης που να εξυπηρετεί καλύτερα το πρόβλημα επιλογής μεταβλητών το οποίο επιλύουμε.

- elitism
  Η μεταβλητή αυτή δηλώνει την εφαρμογή ή μη του ελιτισμού μετά από κάθε διασταύρωση. Αναλυτικότερα, αν ο χρήστης επιθυμεί να επιλέγονται τα δύο καλύτερα χρωμοσώματα από τους δύο "γονείς" και τους δύο "απογόνους" μετά τη διασταύρωση, τότε θα πρέπει να δηλώσει ως TRUE τη μεταβλητή αυτή, διαφορετικά αν την δηλώσει ως FALSE ο αλγόριθμος θα μεταβιβάζει στην επόμενη γενιά πάντα τους δύο "απογόνους" είτε είναι καλύτεροι από τους "γονείς" τους είτε όχι. Ο ελιτισμός γενικά επιταχύνει την ταχύτερη σύγκλιση του αλγορίθμου και για το λόγο αυτό στις περισσότερες δοκιμές επιλέχθηκε η τιμή TRUE.

- p_{old}
  Η μεταβλητή αυτή δηλώνει το ποσοστό των καλύτερων χρωμοσωμάτων από τον πληθυσμό, το οποίο θέλουμε να διατηρεί ο αλγόριθμος αναλλοίωτο σε κάθε επανάληψη-γενιά. Με άλλα λόγια, μετά το πέρας των διασταυρώσεων και της μετάλλαξης σε κάθε επανάληψη, το ποσοστό του έως τότε συνολικού πληθυσμού (παλαιάς και νέας γενιάς) που ορίζεται μέσω της μεταβλητής αυτής αντιπροσωπεύει το τμήμα του πληθυσμού που ο αλγόριθμος "κρατάει" για να μεταβιβάσει στην επόμενη γενιά. Τα χρωμοσώματα που ανήκουν σε αυτό το τμήμα του πληθυσμού θα έχουν την ευκαιρία να διασταυρωθούν στην επόμενη επανάληψη και να μεταβιβάσουν τα χαρακτηριστικά τους. Το υπόλοιπο τμήμα του πληθυσμού, δηλαδή το συμπληρωματικό ποσοστό, θα ανανεωθεί με τυχαία χρωμοσώματα, με

τον ίδιο ακριβώς τρόπο που έγινε και η αρχικοποίηση του πληθυσμού στο πρώτο στάδιο του αλγορίθμου. Λαμβάνει τιμές στο διάστημα [0,1], ενώ προτιμώνται σχετικά μικρές τιμές (μικρότερες του 0.5) προκειμένου να γίνεται καλύτερη εξερεύνηση στο χώρο των λύσεων. Ενδεικτικά, για τιμή $p_{old} = 0.3$, το 70% των όχι καλύτερων ατόμων του πληθυσμού χάνεται και ανανεώνεται με τυχαίο τρόπο, ενώ διατηρείται το 30% των καλύτερων ατόμων για να μεταβιβαστεί στην επόμενη γενιά.

- p_{mut}
  Η μεταβλητή αυτή χρησιμοποιείται για να ορίσει την πιθανότητα με την οποία θα πραγματοποιείται μετάλλαξη σε κάποιο γονίδιο του χρωμοσώματος. Συγκεκριμένα, η πιθανότητα αυτή εκφράζει την πιθανότητα να γίνει μετάλλαξη σε κάθε ένα από τα γονίδια του χρωμοσώματος, και δίνεται ως όρισμα στη συνάρτηση "mutation" για να πραγματοποιηθεί η μετάλλαξη. Ενδεικτικά, για $p_{mut} = 0.05$ υπάρχει πιθανότητα 5% να γίνει μετάλλαξη σε κάθε ένα από τα p γονίδια του εκάστοτε χρωμοσώματος.

4.1.2 Ο γενετικός αλγόριθμος σε βήματα

Αν θέλουμε να περιγράψουμε σε βήματα τις διαδικασίες που ακολουθούνται κατά τη δημιουργία του γενετικού αλγορίθμου, τότε οφείλουμε να επισημάνουμε τα παρακάτω:

1. Παράγεται ο αρχικός τυχαίος πληθυσμός χρωμοσωμάτων από τα οποία θα γίνει και η επιλογή των "γονέων" για την πρώτη διασταύρωση.

2. Για κάθε ένα χρωμόσωμα του πληθυσμού αυτού, υπολογίζεται η τιμή της αντικειμενικής συνάρτησης (τιμή του κριτηρίου BIC). Αυτή η τιμή είναι ενδεικτική της καταλληλότητας του εκάστοτε χρωμοσώματος ως γραμμικού μοντέλου για το συγκεκριμένο σύνολο δεδομένων. Στόχος είναι να βρεθεί το μοντέλο με τη χαμηλότερη τιμή BIC.

3. Ορίζεται η πιθανότητα επιλογής του εκάστοτε χρωμοσώματος ως "γονέας", βάσει της τιμής BIC που του αντιστοιχεί. Η επιλογή γίνεται μέσω τυχαίας δειγματοληψίας με επανάθεση στον αρχικό πληθυσμό με τις καθορισμένες πιθανότητες που εξαρτώνται από την καταλληλότητα του εκάστοτε χρωμοσώματος. Τα χρωμοσώματα που θα επιλεγούν είναι και οι "γονείς" της επόμενης γενιάς.

4. Πραγματοποιείται η διασταύρωση των γονέων που επιλέχθηκαν με κάποια πιθανότητα, σύμφωνα με τον τύπο διασταύρωσης που έχει επιλεγεί από

τον χρήστη και δόθηκε ως παράμετρος στον γενετικό αλγόριθμο. Αν δε γίνει διασταύρωση, τότε ως απόγονοι ορίζονται οι ίδιοι οι "γονείς".

5. Μετά τη διασταύρωση, εκτελείται με κάποια πιθανότητα μετάλλαξη στους απογόνους.

6. Έπειτα, και εφόσον έχει επιλεχθεί από τον χρήστη η εφαρμογή της ελιτίστικης στρατηγικής, οι δύο "γονείς" και οι δύο "απόγονοί" τους συγκρίνονται και αξιολογούνται έτσι ώστε να κρατηθεί το καλύτερο ζευγάρι αυτών. Ως καλύτερο ζευγάρι εννοείται αυτό που δίνει τις 2 χαμηλότερες τιμές στο κριτήριο BIC. Επομένως, στη νέα γενιά μεταβιβάζονται 2 από τα 4 χρωμοσώματα, αυτά που ευνοήθηκαν από τη σύγκριση.

7. Σε αυτό το σημείο υλοποιείται μια δεύτερη μορφή ελιτισμού, υποχρεωτικά και ανεξάρτητα από το αν προηγήθηκε ήδη ο ελιτισμός. Συγκεκριμένα, συγκεντρώνεται η παλιά και η νέα γενιά και επιλέγονται τα καλύτερα χρωμοσώματα έπειτα από κατάταξη από το καλύτερο προς το χειρότερο. Τελικά, κρατούνται τα καλύτερα από αυτά στον αριθμό του προκαθορισμένου από τον χρήστη πληθυσμού που έχει δοθεί ως όρισμα της συνάρτησης. Έτσι δίνεται η ευκαιρία σε προηγούμενες αποδεδειγμένα καλές λύσεις να βελτιωθούν περαιτέρω και να μη χαθούν στο πέρασμα των γενεών.

8. Από τον διαμορφωμένο αυτόν πληθυσμό, ο αλγόριθμος κρατάει ένα προκαθορισμένο ποσοστό, το οποίο δηλώθηκε ως παράμετρος στην αρχή του αλγορίθμου. Τα υπόλοιπα χρωμοσώματα του πληθυσμού παράγονται τυχαία, με αποτέλεσμα να ανανεώνεται ένα μέρος του πληθυσμού με καινούρια χρωμοσώματα. Τα χρωμοσώματα που παράγονται εκ νέου είναι τόσα ώστε να συμπληρωθεί το 100% του πληθυσμού.

9. Για κάθε γενιά, δηλαδή μετά το πέρας κάθε επανάληψης, υπολογίζονται και επιστρέφονται στον χρήστη τα 3 καλύτερα μοντέλα με τις τιμές BIC που τους αντιστοιχούν. Αυτό βοηθάει στο να παρατηρούμε τη βελτίωση που σημειώνεται στην εύρεση της βέλτιστης λύσης καθώς εξελίσσεται ο αλγόριθμος, δηλαδή να μελετήσουμε τη σύγκλισή του.

10. Με την ολοκλήρωση του αριθμού των επαναλήψεων, ο αλγόριθμος επιστρέφει τον τελικό πληθυσμό των χρωμοσωμάτων ακολουθούμενα από τις τιμές που δίνουν στην αντικειμενική συνάρτηση. Στον τελικό πληθυσμό, επιστρέφονται τα χρωμοσώματα χωρίς τυχόν αντίγραφά τους.

Οι παραπάνω διαδικασίες προκειμένου να πραγματοποιηθούν, απαιτούν τον ορισμό και τον υπολογισμό αρκετών εσωτερικών μεταβλητών, δηλαδή μεταβλητών που

δεν δηλώνονται εξωτερικά στον αλγόριθμος ως παράμετροι, αλλά υφίστανται μόνο κατά την εκτέλεση του αλγορίθμου. Για τον γενετικό αλγόριθμο χρειάζονται αρκετές τέτοιες μεταβλητές τις οποίες αναλύουμε στην Ενότητα 4.1.3.

4.1.3 Εσωτερικές μεταβλητές

Παρακάτω παρουσιάζονται οι μεταβλητές που χρησιμοποιούνται εσωτερικά στον γενετικό αλγόριθμο προκειμένου να εκτελεστούν όλες οι απαραίτητες λειτουργίες του.

- var_counter
  Είναι ο ακέραιος αριθμός που δηλώνει το (μέγιστο) πλήθος των τυχαίων μεταβλητών του γραμμικού μοντέλου. Ορίζεται ως η δεύτερη συντεταγμένη της διάστασης του πίνακα X. Επίσης, ταυτίζεται με το μήκος του διανύσματος του εκάστοτε χρωμοσώματος, διότι κάθε συντεταγμένη ενός χρωμοσώματος αντιστοιχεί και σε μια τυχαία μεταβλητή.

- best_evaluations
  Πρόκειται για το διάνυσμα στο οποίο αποθηκεύεται η καλύτερη τιμή του κριτηρίου BIC για κάθε επανάληψη. Η καλύτερη τιμή αντιστοιχεί στη μικρότερη τιμή διότι λύνουμε πρόβλημα ελαχιστοποίησης. Ορίζεται στην αρχή του αλγορίθμου με την εντολή best_evaluations ← rep(0,maxiter), δηλαδή αρχικοποιείται ως ένα διάνυσμα με μηδενικά και μήκος ίσο με τον αριθμό επαναλήψεων. Στο τέλος κάθε επανάληψης αποθηκεύεται η μικρότερη τιμή BIC που αντιστοιχεί σε κάποιο από τα χρωμοσώματα του τρέχοντος πληθυσμού.

- BIC_gen
  Είναι το διάνυσμα στο οποίο αποθηκεύονται οι τιμές του κριτηρίου BIC για κάθε χρωμόσωμα της τρέχουσας γενιάς. Ορίζεται με την εντολή BIC_gen ← rep(0,pop_size) όπου αρχικοποιείται με μηδενικά και μήκος ίσο με τον αριθμό των ατόμων του πληθυσμού. Καθώς προχωράει ο αλγόριθμος, οι τιμές του διανύσματος αυτού ανανεώνονται με τη δημιουργία κάθε επόμενης γενιάς. Η αποθήκευση στη μνήμη όλων των τιμών BIC που αντιστοιχούν σε όλα τα χρωμοσώματα που κατασκευάζει επαναληπτικά ο αλγόριθμος θα σήμαινε μεγάλο υπολογιστικό κόστος για τον αλγόριθμο.

- new_BIC_gen
  Πρόκειται για το διάνυσμα όπου αποθηκεύονται οι τιμές BIC δύο διαδοχικών γενεών, δηλαδή ολόκληρης της παλαιάς γενιάς ("γονείς" αλλά και τα υπόλοιπα χρωμοσώματα που δεν επιλέχθηκαν για διασταύρωση) αλλά και της νέας γενιάς που αποτελείται από τους "απογόνους" που προέκυψαν από τη διασταύρωση των "γονέων". Δημιουργείται με την εντολή new_BIC_gen ← rep(0,2 · pop_size). Αρχικοποιείται με μηδενικές τιμές και έχει μέγεθος ίσο με δύο πληθυσμούς, δηλαδή εκπροσωπεί δύο γενιές. Η δημιουργία αυτής της μεταβλητής είναι απαραίτητη προκειμένου να μπορούν να συγκριθούν οι τιμές της αντικειμενικής συνάρτησης και να κρατηθούν εν τέλει τα καλύτερα χρωμοσώματα σε μέγεθος ενός πληθυσμού που κατασκευάστηκαν κατά την τρέχουσα επανάληψη.

- population_chromo
  Είναι ο πίνακας που αποθηκεύονται τα χρωμοσώματα του πληθυσμού κάθε γενιάς. Οι διαστάσεις του ορίζονται να είναι τέτοιες ώστε το πλήθος των γραμμών του να αντιστοιχεί στο πλήθος των ατόμων του πληθυσμού, και οι στήλες του να είναι ίσες με το μέγιστο πλήθος των επεξηγηματικών μεταβλητών του μοντέλου παλινδρόμησης. Έτσι, ορίζεται με την εντολή population_chromo ← matrix(c(0),$nrow = pop\_size$,$ncol = var\_counter$) και αρχικοποιείται με μηδενικά διανύσματα. Ο πίνακας αυτός αρχικά περιέχει τα αρχικοποιημένα τυχαία χρωμοσώματα σε μορφή διανυσμάτων αλλά καθώς προχωράει ο αλγόριθμος οι τιμές του ανανεώνεται και αποθηκεύονται οι γονείς που έχουν επιλεγεί για διασταύρωση ώστε να προκύψει η επόμενη γενιά. Αυτός ο πίνακας στο σημείο της επιλογής των γονέων αποκτά και μια επιπλέον στήλη, στην οποία αποθηκεύονται οι τιμές BIC του εκάστοτε χρωμοσώματος.

- elit_4_eval
  Είναι ένα διάνυσμα με μήκος ίσο με 4 το οποίο κατασκευάζεται για να χρησιμοποιηθεί σε περίπτωση που επιλεγεί από τον χρήστη να εφαρμοστεί ελιτίστικη στρατηγική. Δημιουργείται με την εντολή elit_4_eval ← rep(c(0),4) και αρχικοποιείται με μηδενικές τιμές. Πιο συγκεκριμένα, σε αυτό το διάνυσμα αποθηκεύονται οι τιμές του κριτηρίου BIC για τα 4 άτομα που προκύπτουν μετά τη διασταύρωση, τους δύο "γονείς" και τους δύο "απογόνους". Στην ελιτίστικη στρατηγική, χρειάζεται να συγκριθούν οι τιμές της αντικειμενικής συνάρτησης των 4 ατόμων, έτσι ώστε να κρατηθεί το καλύτερο ζευγάρι των ατόμων και να περάσει στην επόμενη γενιά. Για να υλοποιηθεί αυτό, χρειάζεται το διάνυσμα elit_4_eval

ώστε να αποθηκεύει τις τιμές BIC που θα συγκριθούν.

- merged_matrix
  Αποτελεί μια βοηθητική μεταβλητή στην οποία συνενώνεται το διάνυσμα population_chromo που έχει ως γραμμές τα χρωμοσώματα του τρέχοντος πληθυσμού με δυαδική κωδικοποίηση που δηλώνει την ύπαρξη ή μη της εκάστοτε μεταβλητής με 1 ή 0 αντίστοιχα, με το διάνυσμα BIC_gen στο οποίο αποθηκεύονται οι τιμές BIC του πληθυσμού αυτού. Αυτός ο πίνακας είναι αρκετά χρήσιμος διότι μας επιτρέπει να βλέπουμε την τιμή BIC που αντιστοιχεί σε κάθε χρωμόσωμα.

- new_pop_chromo
  Πίνακας που κατασκευάζεται κατά αντιστοιχία με το population_chromo, με τη διαφορά ότι σε αυτόν αποθηκεύονται τα χρωμοσώματα της νέας γενιάς, δηλαδή αυτής που προέκυψε μετά τη διασταύρωση, τη μετάλλαξη και την ελιτίστικη στρατηγική, εφόσον αυτή επιλέχθηκε. Οι τιμές του ανανεώνονται σε κάθε επανάληψη ενώ δεν αποκλείεται να αποθηκευτούν σε αυτόν χρωμοσώματα της προηγούμενης γενιάς που μέσω του ελιτισμού καταφέρνουν να "περάσουν" στην επόμενη γενιά. Αρχικά ορίζεται με την εντολή
  new_pop_chromo $\leftarrow$ matrix(c(0),$nrow = pop\_size$,$ncol = var\_counter$)
  και με μηδενικά διανύσματα.

- cross_type_1
  Λογική μεταβλητή που δηλώνει τη διασταύρωση ενός σημείου. Ο ορισμός της γίνεται μέσω ελέγχου συνθήκης και συγκεκριμένα εκχωρείται σε αυτήν η τιμή Αληθής (TRUE) ή ψευδής (FALSE) που προκύπτει από τον έλεγχο για το αν η μεταβλητή type_cross είναι ίση με τη μονάδα. Η αντίστοιχη εντολή είναι η cross_type_1 $\leftarrow$ $type\_cross == 1$. Όμοια, κατασκευάζονται και οι μεταβλητές cross_type_2 και cross_type_3 για τη διασταύρωση δύο σημείων και την ομοιόμορφη διασταύρωση αντίστοιχα.

- fitness_score
  Η μεταβλητή αυτή δημιουργείται ώστε να αποδοθούν οι πιθανότητες επιλογής των χρωμοσωμάτων του τρέχοντος πληθυσμού ως "γονείς" της επόμενης γενιάς. Οι πιθανότητες αυτές είναι ανάλογες των τιμών BIC των χρωμοσωμάτων και ο τρόπος κατασκευής της περιγράφεται στην

Ενότητα 4.1.4.

- parent_indices
  Είναι η μεταβλητή στην οποία αποθηκεύονται τα χρωμοσώματα που επιλέχθηκαν από την τρέχουσα γενιά να είναι οι "γονείς" της επόμενης γενιάς. Πιο συγκεκριμένα, είναι ένα διάνυσμα οι συντεταγμένες του οποίου είναι οι δείκτες που δηλώνουν τη θέση του χρωμοσώματος που θα είναι "γονέας" στο σύνολο του πληθυσμού. Ορίζεται με τη χρήση της εντολής sample, δηλαδή γίνεται τυχαία δειγματοληψία με επανάθεση στον πληθυσμό (population_chromo) και με πιθανότητες ανάλογες των τιμών του διανύσματος fitness_score. Επισημαίνεται ότι εφόσον η δειγματοληψία γίνεται με επανάθεση και ο αριθμός των "γονέων" που θα επιλεχθούν είναι ίσος με το μέγεθος του πληθυσμού, κάποιοι δείκτες θα επιλεγούν περισσότερες από μια φορές, και άρα θα υπάρξουν χρωμοσώματα-"γονείς" που θα διασταυρωθούν παραπάνω από μια φορές. Αφού προσδιοριστούν οι δείκτες των επιλεχθέντων χρωμοσωμάτων, το σύνολο των "γονέων" αποθηκεύεται στον πίνακα population_chromo με τις αντίστοιχες τιμές BIC ως τελευταία στήλη του πίνακα.

- parent_1, parent_2
  Διανύσματα διάστασης 1xp που αποθηκεύουν τους "γονείς" που διασταυρώνονται σε κάθε επανάληψη με τη σειρά που αυτοί έχουν επιλεχθεί. Στη μεταβλητή parent_1 αποθηκεύονται οι γονείς με περιττό δείκτη, δηλαδή ο πρώτος, τρίτος, πέμπτος κ.ο.κ ενώ στη μεταβλητή parent_2 αποθηκεύονται οι γονείς με άρτιο δείκτη, δηλαδή ο δεύτερος, τέταρτος, έκτος κ.ο.κ. Αφού οι "γονείς" είναι και αυτοί χρωμοσώματα, το μήκος των διανυσμάτων που τους εκπροσωπούν θα είναι και αυτό ίσο με τον μέγιστο αριθμό επεξηγηματικών μεταβλητών του προβλήματος, δηλαδή p.

- locus
  Είναι το διάνυσμα με μήκος ένα ή δύο που δηλώνει τη θέση/τις θέσεις στην/στις οποία/οποίες γίνεται τομή των χρωμοσωμάτων των δύο "γονέων" που διασταυρώνονται σύμφωνα με τη διασταύρωση ενός ή δύο σημείων αντίστοιχα. Κατά τον ορισμό του, γίνεται χρήση των εντολών floor και runif για τη διασταύρωση ενός σημείου, έτσι ώστε να επιλέγεται ένας ακέραιος αριθμός από την ομοιόμορφη κατανομή με άκρα το ένα και το (var_counter-1), διότι θέλουμε το σημείο που θα επιλεχθεί να είναι εσωτερικό του χρωμοσώματος. Για τη διασταύρωση δύο σημείων, έγινε χρήση της εντολής sample προκειμένου να επιλεχθούν δύο εσωτερικά

σημεία στο χρωμόσωμα ώστε να γίνει η τομή. Η δειγματοληψία γίνεται χωρίς επανάθεση για να εξασφαλίσουμε ότι μια θέση δε θα επιλεγεί δύο φορές και εξαιρούνται η πρώτη και η τελευταία θέση που θεωρούνται "εξωτερικές" για το χρωμόσωμα. Επίσης χρησιμοποιήθηκε και η εντολή sort στη διασταύρωση δύο σημείων, έτσι ώστε το επιστρεφόμενο διάνυσμα να περιέχει τις θέσεις διατεταγμένες, δηλαδή από τη μικρότερη προς τη μεγαλύτερη για να έχει νόημα η διασταύρωση.

- part_a1, part_a2, part_b1, part_b2
  Βοηθητικές μεταβλητές για τη διασταύρωση ενός σημείου των γονέων. Είναι διανύσματα το μέγεθος των οποίων είναι μεταβλητό και καθορίζεται από την τιμή του locus. Συγκεκριμένα, το part_a1 περιλαμβάνει τα γονίδια του πρώτου γονέα από το πρώτο μέχρι και το σημείο διασταύρωσης, ενώ το part_a2 περιλαμβάνει τα γονίδια του δεύτερου γονέα από τη θέση μετά το σημείο διασταύρωσης μέχρι και το τέλος του χρωμοσώματος. Τα δύο αυτά διανύσματα συνενώνονται και τελικά σχηματίζουν τον πρώτο "απόγονο". Όμοια και για τις μεταβλητές part_b1, part_b2 που σχηματίζουν τον δεύτερο "απόγονο". Για την διασταύρωση δύο σημείων χρησιμοποιούνται κατά αντιστοιχία και δύο ακόμη διανύσματα, τα part_a3 και part_b3.

- off1, off2
  Είναι τα διανύσματα τα οποία αναπαριστούν τους "απογόνους" που προέκυψαν από τη διασταύρωση. Ορίζονται βάσει των επιλογών που έχει κάνει ο χρήστης σχετικά με τον τύπο διασταύρωσης και τον ελιτισμό, ενώ ενδέχεται να συμβεί και μετάλλαξη.

- mask
  Διάνυσμα με μήκος όσο και το μήκος των χρωμοσωμάτων και δυαδικά ψηφία ως συντεταγμένες. Δημιουργείται σε περίπτωση που ο χρήστης διαλέξει την κωδικοποίηση 3 για την διασταύρωση, δηλαδή την ομοιόμορφη διασταύρωση. Οι συντεταγμένες του δηλώνουν τον "γονέα" από τον οποίο θα κληρονομήσει το χρωμόσωμα το εκάστοτε γονίδιο και για τη δημιουργία της μάσκας έγινε και πάλι χρήση της εντολής sample.

- elit_4_chromo
  Πίνακας που έχει ως γραμμές τα τέσσερα χρωμοσώματα που πρέπει να συγκριθούν σε περίπτωση που εφαρμοστεί ελιτισμός, δηλαδή τους δύο "γονείς" με τους αντίστοιχους "απογόνους" τους. Η μεταβλητή αυτή

είναι βοηθητική έτσι προκειμένου να μπορούν να συγκριθούν και τελικά να επιλεγούν τα δύο από τα τέσσερα καλύτερα χρωμοσώματα που θα μεταβιβαστούν στην επόμενη γενιά.

- old_new_100
  Είναι ένας πίνακας με διαστάσεις Nxp στον οποίο αποθηκεύονται προσωρινά τα χρωμοσώματα δύο διαδοχικών γενεών, προηγούμενης γενιάς και "απογόνων". Σε κάθε επανάληψη ανανεώνεται και δρα βοηθητικά, έτσι ώστε να αξιολογηθούν τα άτομα συνολικά και να προωθηθεί στην επόμενη γενιά ένα ποσοστό των καλύτερων.

- old_size
  Μεταβλητή που καθορίζει το πλήθος των ατόμων του πληθυσμού που θα μεταβιβαστεί στην επόμενη γενιά, έπειτα από τη συγκέντρωση των καλύτερων ατόμων δύο διαδοχικών γενεών και την αξιολόγησή τους. Ορίζεται με την εντολή old_size ← pop_size·p_old και θεωρητικά μπορεί να πάρει τιμές στο διάστημα [1,N]. Ωστόσο προτιμώνται τιμές στο κοντά στο 20% ή 30%, προκειμένου να υπάρχει ένα μέρος του πληθυσμού που ανανεώνεται με τυχαίο τρόπο.

- best_population
  Είναι ο πίνακας που επιστρέφει στο τέλος κάθε επανάληψης τα χρωμοσώματα του πληθυσμού διατεταγμένα σε αύξουσα σειρά ως προς την τιμή BIC τους και χωρίς διπλά αντίγραφα. Ανανεώνεται σε κάθε επανάληψη και είναι ο πίνακας που επιστρέφεται τελικά από τη συνάρτηση genetic. Η πρώτη γραμμή του περιλαμβάνει το καλύτερο άτομο του πληθυσμού που αντιστοιχεί στο καλύτερο γραμμικό μοντέλο. Ορίζεται με τη βοήθεια των συναρτήσεων unique και order της R.

4.1.4 Σχολιασμός επιμέρους διαδικασιών

Παρακάτω παρατίθενται τα τμήματα κώδικα και αναλύεται η υλοποίησή τους για την καλύτερη κατανόηση του γενετικού αλγορίθμου.
Προκειμένου να κατασκευαστεί ο γενετικός αλγόριθμος και να ελεγχθεί η σωστή λειτουργία του για την εύρεση του βέλτιστου γραμμικού μοντέλου, προσωμοιώθηκαν δεδομένα τα οποία βασίζονται στην κανονική κατανομή. Πιο αναλυτικά, παράχθηκαν n=50 παρατηρήσεις από p=15 τυχαίες μεταβλητές που βασίζονται στην κανονική κατανομή με μέση τιμή μηδέν και τυπική απόκλιση 1 (τυποποιημένη κανονική κατανομή). Οι 15 αυτές τυχαίες μεταβλητές παίζουν

το ρόλο των ανεξάρτητων τυχαίων μεταβλητών για το γραμμικό μοντέλο με μεταβλητές τις X_i i=1,2,. . . ,p.
Οι εντολές που χρησιμοποιήθηκαν για την παραγωγή των παραπάνω αναγράφονται στο Σχήμα 4.1.

```
#set.seed(123) #for reproducible results
Ncol=15
Nrow=50
X<-matrix(rep(0,Nrow*Ncol),ncol=Ncol)  #create a matrix
for (i in 1:Ncol){
  X[,i]<-rnorm(Nrow,0,1)
}
#columns are the observations(50) of each variable
mean=6+8*X[,1]+3*X[,4]+10*X[,7]-12*X[,12]+4*X[,15]
Y<-rnorm(Nrow,mean,1.5)
#generate 50 response variables based on 4 out of 15 predictive variables
```

Σχήμα 4.1: Παραγωγή αρχικού συνόλου δεδομένων

Όπως φαίνεται, οι παρατηρήσεις καταχωρούνται στον πίνακα X όπου κάθε στήλη του περιέχει τις 50 τιμές της εκάστοτε μεταβλητής. Προσομοιώσαμε επίσης 50 τιμές της μεταβλητής απόκρισης Υ, από την κανονική κατανομή με βάση ένα συγκεκριμένο μοντέλο (μέση τιμή), δηλαδή γνωρίζουμε και καθορίζουμε εκ των προτέρων τον τρόπο με τον οποίο παράχθηκαν οι τιμές αυτές, και τυπική απόκλιση 1.5. Συγκεκριμένα, η μεταβλητή Υ βασίζεται σε 5 από τις 15 μεταβλητές, στην 1η, στην 4η, στην 7η, στην 12η και στην 15η. Στόχος είναι να κατασκευαστεί ο γενετικός αλγόριθμος έτσι ώστε να βρίσκει ως βέλτιστο μοντέλο αυτό που βασίζεται στην 1η,4η,7η,12η και 15η μεταβλητή και να γενικευτεί ώστε να εντοπίζει το βέλτιστο μοντέλο όταν εφαρμοστεί σε άλλα σύνολα δεδομένων.
Εκτός όμως από το αρχικό αυτό σύνολο δεδομένων, η απόδοση του αλγορίθμου θα εξεταστεί στη συνέχεια και σε άλλα σύνολα δεδομένων με ελαφρώς διαφορετική δομή και συγκεκριμένα με μεγαλύτερη διασπορά στη μεταβλητή απόκρισης Υ και σε δεδομένα με μεταβλητές που εμφανίζουν πολυσυγγραμμικότητα. Αναλυτικότερα, το δεύτερο σύνολο δεδομένων κατασκευάζεται με όμοιο τρόπο, με τη διαφορά ότι η μεταβλητή απόκρισης προέρχεται από την κανονική κατανομή με μέση τιμή που περιγράφεται από το ίδιο μοντέλο με πριν αλλά με τυπική απόκλιση ελαφρώς μεγαλύτερη, δηλαδή 2.5. Θέλουμε να ελέγξουμε το βαθμό βεβαιότητας με τον οποίο ο γενετικός επιστρέφει το βέλτιστο μοντέλο, όταν υπάρχει μεγαλύτερος "θόρυβος" στα δεδομένα.

Ο κώδικας που παράγει αυτά τα δεδομένα βρίσκεται στο Σχήμα 4.2.

```
#set.seed(123) #for reproducible results
Ncol=15
Nrow=50
X<-matrix(rep(0,Nrow*Ncol),ncol=Ncol)  #create a matrix
for (i in 1:Ncol){
  X[,i]<-rnorm(Nrow,0,1)
}
#columns are the observations(50) of each variable
mean=6+8*X[,1]+3*X[,4]+10*X[,7]-12*X[,12]+4*X[,15]
Y<-rnorm(Nrow,mean,2.5)
```

Σχήμα 4.2: Παραγωγή δεύτερου συνόλου δεδομένων

Κατόπιν, θα ελεγχθεί ο αλγόριθμος σε δύο ακόμη σύνολα δεδομένων τα οποία παρουσιάζουν το φαινόμενο της πολυσυγγραμμικότητας. Στην περίπτωση αυτή, οι πρώτες 10 τυχαίες μεταβλητές είναι ανεξάρτητες και προέρχονται κατά τα γνωστά από την τυποποιημένη κανονική κατανομή, ενώ οι υπόλοιπες 5 είναι γραμμικοί συνδυασμοί των προηγούμενων, με καθορισμένη μέση τιμή και τυπική απόκλιση τη μονάδα. Η μεταβλητή απόκρισης Y προέρχεται από το μοντέλο με την 1η, την 5η, την 7η, την 11η, την 13η και την 15η μεταβλητή, και για το τρίτο σύνολο δεδομένων η τυπική της απόκλιση ορίστηκε στο 1.5. Στο Σχήμα 4.3 παρουσιάζεται ο κώδικας για την παραγωγή του τρίτου συνόλου δεδομένων.

```
Ncol=15
Nrow=50
X<-matrix(rep(0,Nrow*Ncol),ncol=Ncol)  #create a matrix
for (i in 1:10){
  X[,i]<-rnorm(Nrow,0,1)
}
#columns are the observations(50) of each variable
mean1<-0.3*X[,1]+0.5*X[,2]+0.7*X[,3]+0.9*X[,4]+1.1*X[,5]
for(i in 11:15){
  X[,i]<-rnorm(Nrow,mean1,1)
}
mean2<-4+2*X[,1]-1*X[,5]+1.5*X[,7]+1*X[,11]+0.5*X[,13]+X[,15]
Y<-rnorm(Nrow,mean2,1.5)
```

Σχήμα 4.3: Παραγωγή τρίτου συνόλου δεδομένων

Στο τελευταίο σύνολο δεδομένων προσομειώνουμε 10 τυχαίες μεταβλητές οι οποίες είναι ανεξάρτητες και προέρχονται από την τυποποιημένη κανονική κατανομή, ενώ οι υπόλοιπες 5 προέρχονται από γραμμικούς συνδυασμούς των προηγούμενων, με καθορισμένη μέση τιμή και τυπική απόκλιση τη μονάδα. Η μεταβλητή απόκρισης Y προέρχεται από το μοντέλο με την 1η, την 5η, την 7η, την 11η, την 13η και την 15η μεταβλητή, και για το τέταρτο σύνολο δεδομένων η τυπική της απόκλιση ορίστηκε στο 2.5. Στο Σχήμα 4.4 παρουσιάζεται ο κώδικας για την παραγωγή του τέταρτου συνόλου δεδομένων.

```
Ncol=15
Nrow=50
X<-matrix(rep(0,Nrow*Ncol),ncol=Ncol)  #create a matrix
for (i in 1:10){
  X[,i]<-rnorm(Nrow,0,1)
}
#columns are the observations(50) of each variable
mean1<-0.3*X[,1]+0.5*X[,2]+0.7*X[,3]+0.9*X[,4]+1.1*X[,5]
for(i in 11:15){
  X[,i]<-rnorm(Nrow,mean1,1)
}
mean2<-4+2*X[,1]-1*X[,5]+1.5*X[,7]+1*X[,11]+0.5*X[,13]+X[,15]
Y<-rnorm(Nrow,mean2,2.5)
```

Σχήμα 4.4: Παραγωγή τέταρτου συνόλου δεδομένων

Απαραίτητο στοιχείο του αλγορίθμου είναι ο ορισμός της αντικειμενικής συνάρτησης, η οποία καλείται συχνά σε διαφορετικά σημεία του αλγορίθμου. Η συνάρτηση αυτή κατασκευάζεται με τον τρόπο που παρουσιάζεται στο Σχήμα 4.5.

```
evaluation<-function(chromo){
#set.seed(123)
Nrow=dim(X)[1] #goes through the number of observations
for(j in 1:Ncol){
 if(chromo[j]==0){
  bic<-AIC(lm(Y~1),k=log(Nrow))
 }else{
  bic<-AIC(lm(Y~X[,chromo==1]),k=log(Nrow)) #needs the number of observations
 }
}
return(bic)
}
```

Σχήμα 4.5: Αντικειμενική συνάρτηση BIC

Η συνάρτηση "evaluation" δέχεται ως όρισμά της ένα δυαδικό χρωμόσωμα και υπολογίζει την τιμή BIC του μοντέλου παλινδρόμησης που αυτό αναπαριστά. Εξετάζει και την περίπτωση που το χρωμόσωμα είναι το μηδενικό διάνυσμα, που σημαίνει ότι αναπαριστά το μοντέλο μόνο με τη σταθερά, και του αποδίδει την τιμή BIC που του αντιστοιχεί. Η συνάρτηση "διαβάζει" το πλήθος των παρατηρήσεων με την εντολή Nrow=dim(X)[1] , δηλαδή σαρώνει την πρώτη διάσταση του πίνακα X ο οποίος περιέχει τις παρατηρήσεις. Για τον υπολογισμό της αντικειμενικής συνάρτησης, καλείται η συνάρτηση AIC() της R, με την προσθήκη του όρου k=log(Nrow) προκειμένου να λάβει υπόψη τον αριθμό των παρατηρήσεων για τον υπολογισμό του BIC. Ο όρος αυτός συνιστά και την

“ποινή” (penalty) που επιβάλλει η αύξηση των παραμέτρων προς υπολογισμό στο κριτήριο BIC. Η εντολή lm(Y ∼X[,chromo==1]) περιγράφει την προσαρμογή του γραμμικού μοντέλου στα δεδομένα, με τη συνθήκη X[,chromo==1]) να επιβάλλει την προσθήκη της μεταβλητής στο υπό εξέταση μοντέλο, μόνο εάν το χρωμόσωμα στην αντίστοιχη θέση έχει μονάδα.
Το τμήμα κώδικα του Σχήματος 4.6 αποτελεί εισαγωγικό μέρος του κυρίως προγράμματος, δηλαδή του γενετικού αλγορίθμου.

```
var_counter<-dim(X)[2]  #max number of variables
best_evaluations<-rep(0,maxiter) #the best(min) BIC values for each iteration
BIC_gen<-rep(0,pop_size)
new_BIC_gen<-rep(0,2*pop_size)
population_chromo<-matrix(c(0),nrow=pop_size,ncol=var_counter)
elit_4_eval<-rep(c(0),4)  #store the BIC scores of the 2 parents and the 2 offspring

### INITIALISATION OF POPULATION ###
for (i in 1:pop_size){
  population_chromo[i,]<-sample(c(0,1),prob=c(0.5,0.5),size=var_counter,rep=TRUE)
}

print("initial population")
print(population_chromo)

### COMPUTE BIC FOR THE POPULATION ###
for (i in 1:pop_size){
  BIC_gen[i]<-evaluation(population_chromo[i,])
}


merged_matrix<-cbind(population_chromo,BIC_gen) #each chromosome with its BIC score
print("initial population with bic values")
print(merged_matrix)  #last column is bic
new_pop_chromo<-matrix(nrow=pop_size,ncol=var_counter)
cross_type_1<-type_cross==1
cross_type_2<-type_cross==2
cross_type_3<-type_cross==3
```

Σχήμα 4.6: Αρχικό μέρος του αλγορίθμου

Στο Σχήμα 4.6 παρουσιάζεται το αρχικό κομμάτι του γενετικού αλγορίθμου, το οποίο περιλαμβάνει τις αρχικοποιήσεις των μεταβλητών, την αρχικοποίηση του πληθυσμού και την αξιολόγηση αυτού. Η εντολή set.seed() χρησιμοποιείται στην R προκείμενου να βεβαιωθούμε ότι τα τυχαία δεδομένα που παράγουμε καθ’όλη τη διάρκεια του αλγορίθμου στη συνέχεια μπορούν να αναπαραχθούν και σε επόμενη δοκιμή του αλγορίθμου. Μέχρι αυτό το σημείο ο αλγόριθμος έχει τις απαραίτητες μεταβλητές που χρειάζονται προκειμένου να αρχίσει η επαναληπτική διαδικασία. Τυπώνονται στην οθόνη του χρήστη τόσο ο αρχικός πληθυσμός, όσο και ο πίνακας που περιέχει τον αρχικό πληθυσμό συνοδευό-

μενο από τις τιμές του κριτηρίου BIC για κάθε χρωμόσωμα.

```
### STARTING THE ITERATIONS ###
for(i in 1:maxiter){
  #rescaling the bic scores to have values between 0 and 1 (probabilities)
  #formula: xnew=(x-xmin)/(xmax-xmin)
  a=merged_matrix[,var_counter+1]-min(merged_matrix[,var_counter+1])
  b=max(merged_matrix[,var_counter+1])-min(merged_matrix[,var_counter+1])
  fitness_score=1-a/b  #The individual which has a high distance from the min should
  #be converted into a low probability
  #select the parents based on the fitness scores
  parent_indices<-sample(1:length(fitness_score),size=pop_size,replace=TRUE,
                         prob=fitness_score)
  population_chromo<-merged_matrix[parent_indices,]
  print(population_chromo)
```

Σχήμα 4.7: Επιλογή των γονέων

Στο τμημά κώδικα του Σχήματος 4.7 ξεκινούν οι επαναλήψεις του γενετικού αλγόριθμου. Αρχικά, εντός των επαναλήψεων ορίζονται οι πιθανότητες επιλογής των γονέων που οδηγούν στον προσδιορισμό των γονέων και τον επανακαθορισμό του πίνακα population_chromo. Όπως έχει αναλυθεί παραπάνω, οι πιθανότητες αυτές υπολογίζονται με βάση την τιμή BIC που αντιστοιχεί στο κάθε χρωμόσωμα-υποψήφιο μοντέλο γραμμικής παλινδρόμησης και οι τιμές των πιθανοτήτων αποθηκεύονται στη μεταβλητή fitness_score.
Πιο αναλυτικά, η επιλογή των γονέων γίνεται ως εξής:
Εντοπίζεται η μικρότερη τιμή του κριτηρίου BIC στον πληθυσμό και για κάθε ένα άτομο του πληθυσμού υπολογίζεται η διαφορά της τιμής BIC του από την μικρότερη τιμή BIC που μέχρι τώρα έχει βρεθεί. Έπειτα, η διαφορά αυτή διαιρείται με το διαφορά της μικρότερης τιμής από τη μεγαλύτερη, δηλαδή γίνεται διαίρεση με το εύρος τιμών του κριτηρίου. Με αυτόν τον τρόπο, οι τιμές που αποθηκεύονται στη μεταβλητή fitness_score δηλώνουν και τις πιθανότητες επιλογής, με την έννοια ότι τα άτομα που απέχουν μεγάλη απόσταση από τη μικρότερη τιμή BIC είναι αδύναμα και άρα αντιστοιχούν σε μικρή πιθανότητα επιλογής.
Η μεταβλητή parent_indices αποθηκεύει τους δείκτες που δηλώνουν τους γονείς που επιλέγονται από τον πληθυσμό μέσω της παραπάνω διαδικασίας ενώ η αποθήκευση των γονέων γίνεται στην ήδη υπάρχουσα μεταβλητή population_chromo, με την έννοια ότι ανανεώνονται οι τιμές της. Σε αυτό το σημείο, ο αλγόριθμος τυπώνει στην κονσόλα του χρήστη τον πληθυσμό των γονέων που τελικά επιλέχθηκαν.
Στο επόμενο στάδιο που είναι αυτό της διασταύρωσης, οι γονείς θα διασταυ-

ρωθούν διαδοχικά και σειριακά, δηλαδή ο πρώτος με τον δεύτερο, ο τρίτος με τον τέταρτο κ.ο.κ και χωρίς να αλλάξει η σειρά με την οποία αυτοί προέκυψαν από τη δειγματοληψία με επανάθεση. Το επαναληπτικό μέρος που αφορά τη διασταύρωση παρουσιάζεται στο Σχήμα 4.8.

```
for(n in 1:(pop_size/2)){
  #name the parents
  parent_1<-population_chromo[2*n-1,1:var_counter] #odd-numbered parents
  parent_2<-population_chromo[2*n,1:var_counter]   #even-numbered parents
  if(i>10){ #change the probabilities according to the stage of the algorithm
    p_cross<-0.5
    p_mut<-0.1
  }
```

Σχήμα 4.8: Διασταύρωση γονέων

Πιο συγκεκριμένα, οι διασταυρώσεις γίνονται με έναν από τους τρεις γνωστούς τρόπους και αφήνεται στον χρήστη να επιλέξει ποια μέθοδο θα χρησιμοποιήσει. Πριν όμως γίνει αυτό, το εκάστοτε χρωμόσωμα-γονέας αποθηκεύεται προσωρινά στη μνήμη ως parent_1 ή ως parent_2 ανάλογα με το αν κατέχει άρτια ή περιττή θέση στον πίνακα population_chromo. Η διαδικασία αυτή γίνεται επαναληπτικά για κάθε ένα ζεύγος χρωμοσωμάτων που θα διασταυρωθεί και είναι απαραίτητη προκειμένου να μπορεί να υλοποιηθεί ο εκάστοτε τύπος διασταύρωσης. Αξίζει να σημειωθεί ότι οποιεσδήποτε και να είναι οι τιμές των πιθανοτήτων διασταύρωσης και μετάλλαξης που θα δώσει ο χρήστης στον αλγόριθμο ως ορίσματα, αυτές τροποποιούνται από την 11η επανάληψη και μετά έτσι ώστε να γίνεται μεγαλύτερη εξερεύνηση του χώρου των λύσεων κατά τα πρώτα στάδια του αλγορίθμου και καλύτερη εκμετάλλευση του χώρου στα τελικά στάδια στα οποία προσεγγίζεται η λύση.
Η επιλογή του τύπου διασταύρωσης που θα εφαρμοστεί πραγματοποιείται με τη χρήση των εντολών if και else if.
Πιο αναλυτικά, για τη διασταύρωση ενός σημείου (one-point crossover) έχουμε τον κώδικα του Σχήματος 4.9.

```
###CROSSOVER###
#a randomly generated uniform number between 0 and 1 is compared to the crossove
#probability and if it is less that the probability,crossover is performed
#otherwise offspring are identical to the parents
if(type_cross==1){ #1-point crossover
  if (runif(1)<=p_cross){
    locus<-floor(runif(1,min=1,max=var_counter-1))  #select point of crossover
    #1st offspring
    part_a1<-parent_1[1:locus]
    part_a2<-parent_2[(locus+1):var_counter]
    #2nd offspring
    part_b1<-parent_2[1:locus]
    part_b2<-parent_1[(locus+1):var_counter]
    off1<-c(part_a1,part_a2)
    off2<-c(part_b1,part_b2)
  } else {
```

Σχήμα 4.9: Διασταύρωση ενός σημείου

Με χρήση της εντολής runif(1) παράγεται ένας αριθμός στο διάστημα (0,1) από την ομοιόμορφη κατανομή και συγκρίνεται με την πιθανότητα διασταύρωσης p_cross. Αν η τιμή αυτή είναι μικρότερη από την πιθανότητα διασταύρωσης p_cross, τότε πραγματοποιείται η διασταύρωση, διαφορετικά ως απόγονοι ορίζονται οι "γονείς" και άρα δεν πραγματοποιείται διασταύρωση. Αν πραγματοποιηθεί η συγκεκριμένη διασταύρωση, τότε γίνεται η επιλογή του εσωτερικού σημείου του χρωμοσώματος στο οποίο θα γίνει η εναλλαγή του γενετικού υλικού των "γονέων" και καταχωρείται στη μεταβλητή locus. Στη συνέχεια, με τη βοήθεια των μεταβλητών part_a1, part_a2, part_b1, part_b2 που αποτελούν τα τμήματα των "γονέων" που εναλλάσσονται, προκύπτουν οι δύο "απόγονοι" που αποθηκεύονται στις μεταβλητές off1 και off2.

Στο Σχήμα 4.10 παρουσιάζεται η υλοποίηση της διασταύρωσης δύο σημείων (two-point crossover). Συγκεκριμένα, γίνεται η επιλογή δύο εσωτερικών σημείων του χρωμοσώματος στο οποίο θα γίνει η τομή και η εναλλαγή των τμημάτων των γονέων. Ως αποτέλεσμα, παράγονται τα δύο νέα χρωμοσώματα ως απόγονοι των γονέων.
Όσον αφορά τον τελευταίο τύπο διασταύρωσης, δηλαδή την ομοιόμορφη διασταύρωση (uniform crossover), ο κώδικας που την υλοποιεί παρουσιάζεται στο Σχήμα 4.11.

Δημιουργείται η "μάσκα" διασταύρωσης με την εντολή sample από ένα δείγμα μεγέθους 15, δηλαδή ίσο με το μέγιστο πλήθος όλων των επεξηγηματικών μεταβλητών το οποίο αποτελείται από 0 και 1. Τα δυαδικά ψηφία στη μάσκα

```
} else if(type_cross==2){ #2-point crossover
  if (runif(1)<=p_cross){
    locus<-sort(sample(2:(var_counter-1),2,replace=FALSE))  #select 2 points
    part_a1<-parent_1[1:(locus[1]-1)]
    part_a2<-parent_2[locus[1]:(locus[2]-1)]
    part_a3<-parent_1[(locus[2]):var_counter]
    off1<-c(part_a1,part_a2,part_a3)
    part_b1<-parent_2[1:(locus[1]-1)]
    part_b2<-parent_1[locus[1]:(locus[2]-1)]
    part_b3<-parent_2[(locus[2]):var_counter]
    off2<-c(part_b1,part_b2,part_b3)
  } else {
    off1<-parent_1
    off2<-parent_2
  }  #end of 2-point crossover
```

Σχήμα 4.10: Διασταύρωση δύο σημείων

```
} else if (type_cross==3) { #uniform crossover
  mask<-sample(c(0,1),size=var_counter,prob=c(0.5,0.5),replace=TRUE)
  off1<-rep(0,var_counter)
  off2<-rep(0,var_counter)
  for(k in 1:length(parent_1)){
    if(mask[k]==1){
      off1[k]<-parent_1[k]
      off2[k]<-parent_2[k]
    } else {
      off1[k]<-parent_2[k]
      off2[k]<-parent_1[k]
    }
  } #end of uniform crossover
}#end of if-else statement for types of crossover
```

Σχήμα 4.11: Ομοιόμορφη διασταύρωση

δηλώνουν την επιλογή της αντίστοιχης συντεταγμένης-γονίδιο του χρωμοσώματος-"απογόνου" από τον πρώτο "γονέα" (αν το ψηφίο είναι 1) ή από τον δεύτερο "γονέα" (αν το ψηφίο είναι 0) αντίστοιχα. Για τον δεύτερο "απόγονο", ισχύουν τα ίδια αν αντιμεταθέσουμε τους "γονείς".
Μετά τη διασταύρωση, ακολουθεί η μετάλλαξη, η οποία υλοποιείται μέσω μιας συνάρτησης που καλείται από το κύριο μέρος του αλγορίθμου μέσω των εντολών του Σχήματος 4.12.

```
#mutation
off1<-mutation(off1,p_mut)
off2<-mutation(off2,p_mut)
```

Σχήμα 4.12: Κλήση της συνάρτησης mutation

Η διαδικασία της μετάλλαξης πραγματοποιείται μέσω του κώδικα του Σχήματος 4.13.

```
mutation<-function(chromo,p_mut){
#set.seed(123)
  for(i in 1:length(chromo)){
    if(runif(1)<=p_mut){
      chromo[i]<-1-chromo[i]
    }
  }
return(chromo) #returns the same or the mutated chromosome
}
```

Σχήμα 4.13: Μετάλλαξη

Δέχεται στο όρισμά της το υποψήφιο χρωμόσωμα προς μετάλλαξη και την πιθανότητα μετάλλαξης p_mut. Για κάθε συντεταγμένη του χρωμοσώματος, η συνάρτηση αυτή αλλάζει την υπάρχουσα τιμή της συντεταγμένης από 0 σε 1 και αντίστροφα, σύμφωνα με την τιμή της πιθανότητας μετάλλαξης. Υπενθυμίζεται ότι η παράμετρος αυτή παίρνει τιμή στο [0,1] ως πιθανότητα, και για την ακραία τιμή 0 δεν πραγματοποιείται καθόλου μετάλλαξη. Πιο αναλυτικά, για κάθε μια συντεταγμένη του χρωμοσώματος, παράγεται μια τιμή στο διάστημα [0,1] από την ομοιόμορφη κατανομή και ελέγχεται αν η τιμή αυτή είναι μικρότερη, ίση ή μεγαλύτερη από την δοσμένη πιθανότητα μετάλλαξης. Στην ομοιόμορφη κατανομή κάθε αριθμός έχει την ίδια πιθανότητα να παραχθεί και άρα κάθε αριθμός αναπαριστά ένα ισοπίθανο ενδεχόμενο. Σε περίπτωση που αυτός ο αριθμός είναι μικρότερος ή ίσος από την πιθανότητα μετάλλαξης, τότε γίνεται αλλαγή του ψηφίου του χρωμοσώματος από 0 σε 1 ή αντίστροφα. Σε περίπτωση όμως που ο αριθμός που παράχθηκε είναι μεγαλύτερος από την πιθανότητα μετάλλαξης, τότε η συγκεκριμένη συντεταγμένη παραμένει αναλλοίωτη και δεν πραγματοποιείται μετάλλαξη. Η συνάρτηση αυτή υλοποιείται για οποιοδήποτε μέγεθος χρωμοσώματος διότι μπορεί να το αναγνωρίσει μέσω της εντολής length(). Στο τέλος, επιστρέφει το τελικό χρωμόσωμα, έτσι όπως έχει διαμορφωθεί κατόπιν της μετάλλαξης, το οποίο ενδέχεται να είναι και ταυτόσημο με το αρχικό.
Το επόμενο μέρος του αλγορίθμου υλοποιείται σε περίπτωση που επιλεγεί από το χρήστη η ελιτίστικη στρατηγική (elitism). Εάν ο χρήστης επιλέξει την εφαρ-

μογή της ελιτίστικης στρατηγικής θέτοντας τη μεταβλητή elitism ως TRUE, τότε εκτελούνται οι εντολές που επιλέγουν ως νέα μέλη της νέας γενιάς τα καλύτερα δύο χρωμοσώματα από τα τέσσερα, δηλαδή από τους δύο "γονείς" και τους δύο "απογόνους" τους. Αντιθέτως, αν η μεταβλητή elitism επιλεγεί ως FALSE, τότε δεν υλοποιείται καμία αλλαγή και ως απόγονοι επιλέγονται τα δύο παιδιά που προέκυψαν από τις διαδικασίες της διασταύρωσης και κατόπιν της μετάλλαξης.
Οι εντολές που υλοποιούν τα παραπάνω καταγράφονται στο Σχήμα 4.14 που ακολουθεί.

```
if (elitism==TRUE){
  elit_4_chromo<-rbind(parent_1,parent_2,off1,off2) #compare the chromosomes
  elit_4_eval<-c(evaluation(parent_1),evaluation(parent_2),evaluation(off1),
                 evaluation(off2))
  #concatenate the BIC scores of the 4 chromosomes for testing
  #elit_4_eval_sorted<-sort(elit_4_eval)
  off1<-elit_4_chromo[order(elit_4_eval)[1],]
  off2<-elit_4_chromo[order(elit_4_eval)[2],]
}



new_pop_chromo[2*n-1,]<-off1
new_pop_chromo[2*n,]<-off2

} #end of for-loop n
```

Σχήμα 4.14: Ελιτισμός

Οι δύο τελευταίες εντολές καταχωρούν τα δύο παιδιά που προέκυψαν είτε με εφαρμογή της ελιτίστικης στρατηγικής είτε χωρίς, στις αντίστοιχες θέσεις του νέου πληθυσμού, δηλαδή της νέας γενιάς.
Σε αυτό το σημείο απομένει στον γενετικό αλγόριθμο να συγκεντρώσει τους πληθυσμούς της παλιάς και της νέας γενιάς και να επιλέξει τα καλύτερα άτομα από το σύνολο αυτό ώστε να τα αποθηκεύσει και να μη χαθούν με το πέρας των επαναλήψεων. Η διαδικασία αυτή μπορεί να χαρακτηριστεί ως ένας δεύτερος τύπος ελιτισμού προκειμένου να εξασφαλιστεί ότι δε θα χαθούν οι λύσεις που αντιστοιχούν σε μοντέλα παλινδρόμησης με το καλύτερο υποσύνολο επεξηγηματικών μεταβλητών. Συγκεκριμένα, συνενώνονται σε έναν κοινό νέο πίνακα old_new_100 τα στοιχεία των πινάκων merged_matrix και new_pop_chromo δηλαδή τα χρωμοσώματα παλαιάς και νέας γενιάς προκειμένου να καταστεί δυνατή η ταξινόμησή τους από το καλύτερο προς το χειρότερο με βάση την τιμή του κριτηρίου BIC. Για το σκοπό αυτό, καλείται εκ νέου η συνάρτηση

evaluation που υπολογίζει την τιμή BIC για το κάθε μοντέλο παλινδρόμησης και τελικά τα στοιχεία του νέου πίνακα αναδιατάσσονται από το καλύτερο άτομο προς το χειρότερο. Μετά την αναδιάταξη, ο αλγόριθμος κρατάει ένα τμήμα του συνολικού πληθυσμού, και συγκεκριμένα ίσο σε μέγεθος του πληθυσμού που έχει ορίσει ο χρήστης στην ανάλογη παράμετρο (το μέγεθος του πληθουσμού διατηρείται από επανάληψη σε επανάληψη). Το Σχήμα 4.15 που ακολουθεί υλοποιεί τη συγκέτρωση παλαιάς και νέας γενιάς στον αλγόριθμο.

```
old_new_100<-rbind(merged_matrix[,1:var_counter],new_pop_chromo)
new_BIC_gen<-rep(NA,dim(old_new_100)[1])
#bic for the old_new_100
for(v in 1:dim(old_new_100)[1]){
  new_BIC_gen[v]<-evaluation(old_new_100[v,])

}
#old population is now replaced by the 50 best of the old_new_100
merged_matrix[,1:var_counter]<-old_new_100[order(new_BIC_gen)[1:pop_size],] #sort
```

Σχήμα 4.15: Συγκέντρωση παλαιάς και νέας γενιάς

Απαραίτητο τώρα είναι να ανανεωθεί μέρος του πληθυσμού ώστε να αναζητηθούν νέες λύσεις και να εξερευνηθεί περαιτέρω ο χώρος των λύσεων. Ο αλγόριθμος σχεδιάστηκε έτσι ώστε να κρατάει ένα συγκεκριμένο ποσοστό στη νέα γενιά και να αντικαθιστά το υπόλοιπο μέρος του πληθυσμού με καινούργια τυχαία χρωμοσώματα. Η ανανέωση του πληθυσμού γίνεται και πάλι με χρήση της εντολής sample όπως έγινε και κατά την αρχικοποίηση του πληθυσμού. Επίσης, υπάρχουν και εντολές που πραγματοποιούν τον εκ νέου υπολογισμό της τιμής του κριτηρίου BIC για όλα τα χρωμοσώματα, καθώς υπάρχουν τα νέα χρωμοσώματα για τα οποία δεν έχει γίνει ακόμη ο υπολογισμός. Σε τελευταίο στάδιο, καταχωρούνται στη μεταβλητή best_population τα χρωμοσώματα του πληθυσμού ταξινομημένα από το καλύτερο προς το χειρότερο και χωρίς διπλότυπα, δηλαδή αποκλείονται τα χρωμοσώματα που ταυτίζονται και επομένως αντιπροσωπεύουν την ίδια ακριβώς λύση. Στο τέλος κάθε επανάληψης τυπώνεται το διάνυσμα best_evaluations το οποίο περιέχει τη μικρότερη τιμή του κριτηρίου BIC που εντοπίστηκε σε κάθε επανάληψη. Αυτό το διάνυσμα μας επιτρέπει να παρακολουθήσουμε την πορεία του αλγορίθμου ως προς τη μείωση και τελικά τη σταθεροποίηση της τιμής BIC. Ταυτόχρονα, ο αλγόριθμος τυπώνει στην κονσόλα του χρήστη τα 3 καλύτερα χρωμοσώματα που προέκυψαν από την κάθε επανάληψη ακολουθούμενα από τις αντίστοιχες τιμές BIC στην τελευταία

συντεταγμένη του διανύσματος. Στο Σχήμα 4.16 αποτυπώνονται οι εντολές για την ανανέωση του τμήματος του πληθυσμού.

Οι τελευταίες εντολές του γενετικού αλγορίθμου, αφού ολοκληρώθηκε η επαναληπτική διαδικασία αλλά και όλοι οι γενετικοί τελεστές, απλώς τυπώνουν στην οθόνη του χρήστη τις τιμές του κριτηρίου BIC για τα καλύτερα χρωμοσώματα της κάθε γενιάς που εντοπίζονται σε κάθε επανάληψη και τον τελικό πληθυσμό των χρωμοσωμάτων που αποτελείται από τα καλύτερα χρωμοσώματα που εντοπίστηκαν στο τέλος της κάθε επανάληψης μαζί με την τιμή BIC που τους αντιστοιχούν ως τελευταία στήλη. Ενδέχεται ο αριθμός των χρωμοσωμάτων που επιστρέφονται στον χρήστη να είναι μικρότερος από το καθορισμένο μέγεθος του πληθυσμού (pop_size) που στην περίπτωσή μας ορίσθηκε να είναι ίσο με 50, διότι αποκλείονται οι λύσεις που εμφανίζονται περισσότερες από μια φορές. Στο Σχήμα 4.17 βλέπουμε τις εντολές εκτύπωσης.

```
  old_size<-pop_size*p_old
  for(k in(old_size+1):pop_size){
    merged_matrix[k,1:var_counter]<-sample(c(0,1),prob=c(0.5,0.5),size=var_counter,
                                           rep=TRUE)


  }
  #evaluate from scratch
  for(l in 1:pop_size){

    BIC_gen[l]<-evaluation(merged_matrix[l,1:var_counter])
  }
  merged_matrix[,var_counter+1]<-BIC_gen
  best_population<-unique(merged_matrix[order(BIC_gen),])
  best_evaluations[i]<-min(BIC_gen)
  print("BEST UNIQUE POPULATION 1:3")
  print(best_population[1:3,])
} # END OF ITERATIONS
```

Σχήμα 4.16: Ανανέωση τμήματος του πληθυσμού

```
print("best evaluations through all iterations")
print(best_evaluations)
print("final population")
print(best_population)
```

Σχήμα 4.17: Εντολές εκτύπωσης

4.2 Εξαντλητική Αναζήτηση

Όπως έχει ήδη αναφερθεί, ο γενετικός αλγόριθμος όπως και κάθε αλγόριθμος στοχαστικής βελτιστοποίησης δεν εγγυάται την εύρεση ολικού μεγίστου ή ελαχίστου. Αυτό σημαίνει ότι είναι δυνατόν ο γενετικός αλγόριθμος να συγκλίνει σε μια λύση σε πεπερασμένο χρόνο η οποία να μην είναι η βέλτιστη, καθώς δεν έχει εξερευνηθεί εξονυχιστικά ολόκληρος ο χώρος των λύσεων αλλά ένα ικανοποιητικό μέρος αυτού. Μια ενδελεχής μελέτη του χώρου των υποψήφιων λύσεων συνεπάγεται και μεγάλο υπολογιστικό κόστος, και αυτός είναι ο λόγος που προτιμώνται εναλλακτικοί αλγόριθμοι όπως ο γενετικός έναντι της μεθόδου του εξονυχιστικού ψαξίματος ή πλήρους αναζήτησης (exhaustive search ή full enumeration). Η χρήση της μεθόδου που εξετάζει όλες τις δυνατές λύσεις συχνά καθίσταται απαγορευτική, ειδικά σε προβλήματα όπως αυτό για το οποίο έχουμε 15 επεξηγηματικές μεταβλητές και επομένως θα χρειαστεί η εξέταση όλων των 2^{15} μοντέλων γραμμικής παλινδρόμησης. Ωστόσο, στα πλαίσια της παρούσας διπλωματικής εργασίας υλοποιείται και αυτός ο αλγόριθμος προκειμένου να γίνει επαλήθευση της λύσης που προτείνεται από τον γενετικό και να διαπιστώσουμε αν πράγματι μας παρέχει τη βέλτιστη λύση ή έστω κάποια λύση που να είναι κοντά στην πραγματική. Στο Σχήμα 4.18 υπάρχει ο κώδικας που υλοποιεί τον αλγόριθμο της πλήρους αναζήτησης για το σύνολο δεδομένων που προσομοιώθηκαν.

Συγκεκριμένα, με την εντολή expand.grid() δημιουργείται ένα πλαίσιο δεδομένων (dataframe) που σε κάθε γραμμή του περιέχει όλους τους δυνατούς συνδυασμούς των ορισμάτων που δέχεται. Στην περίπτωσή μας δέχεται ως όρισμα 15 φορές το διάνυσμα (0,1) που σημαίνει ότι θα δημιουργήσει ένα πλαίσιο δεδομένων με 2^{15} γραμμές, όσες είναι δηλαδή και όλες οι δυνατές αναδιατάξεις των διανυσμάτων (0,1) και 15 στήλες, όσα είναι και τα ορίσματα που δέχεται. Έπειτα, το πλαίσιο δεδομένων μετατρέπεται με την εντολή as.matrix() σε πίνακα έτσι ώστε να μπορούν να χρησιμοποιηθούν όλες οι γνωστές συναρτήσεις των πινάκων στην R. Στη συνέχεια, καλείται η συνάρτηση "evaluation" που κατασκευάστηκε για τον γενετικό αλγόριθμο, προκειμένου να αξιολογηθούν για το συγκεκριμένο σύνολο δεδομένων όλα τα δυνατά μοντέλα, που κωδικοποιούνται ως γραμμές του πίνακα all_models και η τιμή αξιολόγησής τους αποθηκεύεται στο διάνυσμα BIC_values. Δημιουργείται ένας πίνακας με όλα τα μοντέλα ακολουθούμενα από την τιμή της συνάρτησης αξιολόγησης ως 16η στήλη ο οποίος στη συνέχεια ταξινομείται από το καλύτερο προς το χειρότερο μοντέλο, και επιστρέφεται τελικά η θέση του καλύτερου μοντέλου στον αρχικό πίνακα (all_models) και το μοντέλο που εκπροσωπείται σε αυτή τη θέση. Επίσης, επιστρέφονται και οι τιμές του κριτηρίου BIC ταξινομημένες από τη μικρότερη προς τη μεγαλύτερη για όλα τα δυνατά μοντέλα πολλαπλής γραμμικής παλινδρόμησης του συνόλου δεδομένων. Ο αλγόριθμος αυτός απαιτεί αρκετή παρα-

πάνω ώρα για να τρέξει σε σχέση με τον γενετικό, καθώς για το πρόβλημα των 15 συνολικά επεξηγηματικών μεταβλητών χρειάζεται περίπου 9 λεπτά σε υπολογιστή intel Core επεξεργαστή i7.

```
#full enumeration
p=15
full=2^p
regressors<-c("x1","x2","x3","x4","x5","x6","x7","x8","x9","x10","x11","x12",
              "x13","x14","x15")


all_models<-expand.grid(c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),
                        c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1))


all_models<-expand.grid(c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),
                        c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1))
all_models<-all_models[1:full,]
all_models_mat<-as.matrix(all_models)


BIC_values<-rep(0,full)
for (i in 1:full){
 BIC_values[i]<-evaluation(all_models[i,])
}


merged<-cbind(all_models_mat,BIC_values)
merged_sort<-merged[order(merged[,16]),]
min_pos<-which.min(BIC_values)  #location of the best model
best_model<-all_models[min_pos,]
print("best model")
print(best_model)
sorted_BIC<-sort(BIC_values,decreasing=FALSE)
print(sorted_BIC)
```

Σχήμα 4.18: Αλγόριθμος πλήρους αναζήτησης

Κεφάλαιο 5

ΑΠΟΤΕΛΕΣΜΑΤΑ

5.1 Έλεγχος αλγορίθμου

Όταν σχεδιάζεται ένας αλγόριθμος, κρίνεται απαραίτητο να ελεγχθεί ως προς λογικά και συντακτικά λάθη. Τα συντακτικά λάθη ανιχνεύονται κατά τη διαδικασία της μεταγλώττισης και αφορούν παραβιάσεις του τυπολογικού και συντακτικού της γλώσσας (π.χ. μια εντολή έχει γραφεί συντακτικώς λανθασμένα). Στην περίπτωση αυτή, ο προγραμματιστής πρέπει να επιστρέψει στον πηγαίο κώδικα, να διορθώσει τα λάθη και να το υποβάλλει εκ νέου για μεταγλώττιση. Τα λογικά λάθη, από την άλλη πλευρά, είναι τα πλέον δύσκολα στην ανίχνευσή τους και σχετίζονται με σφάλματα που έχουν γίνει στη λογική επίλυσης του προβλήματος (π.χ το πρόγραμμα παράγει άλλα αποτελέσματα και όχι τα ζητούμενα). Ο υπολογιστής, εκτελεί όλες τις πράξεις ενός προγράμματος χωρίς να μπορεί να κάνει έλεγχο της λογικής ορθότητας του προγράμματος.
Για τον έλεγχο της ορθότητας του αλγορίθμου και των αποτελεσμάτων που αυτός παράγει, εκτελείται έλεγχος μέσω προσομοιωμένων δεδομένων στα οποία γνωρίζουμε τον τρόπο με τον οποίον παράχθηκαν, δηλαδή γνωρίζουμε το μοντέλο από το οποίο προέρχονται. Με άλλα λόγια, παράγονται τυχαία δεδομένα για τα οποία ο τρόπος δημιουργίας τους είναι προκαθορισμένος και η μεταξύ τους εξάρτηση είναι γνωστή, κατά συνέπεια γνωρίζουμε το αποτέλεσμα που αναμένεται από τον γενετικό αλγόριθμο, αν υποθέσουμε ότι δουλεύει επιτυχώς. Καλώντας λοιπόν τον αλγόριθμο να τρέξει για τα συγκεκριμένα δεδομένα, είμαστε σε θέση να ελέγξουμε την αποδοτικότητά του. Ο αλγόριθμος προγραμματίστηκε με βάση το πρώτο σύνολο δεδομένων το οποίο περιέχει n=50 παρατηρήσεις από p=15 επεξηγηματικές μεταβλητές και μία εξαρτημένη μεταβλητή Υ (διάνυσμα) για την οποία έχει οριστεί η σχέση της με 5 από τις 15 συνολικά μεταβλητές, και συγκεκριμένα τις X_1, X_4, X_7, X_{12} και X_{15}. Αναμένουμε λοιπόν ο γενετικός αλγόριθμος να δώσει ορθά αποτελέσματα, δηλαδή

τα αποτελέσματά του να είναι τέτοια ώστε το καλύτερο χρωμόσωμα να είναι αυτό που περιέχει άσσους στην πρώτη, τέταρτη, έβδομη, δωδέκατη και δέκατη πέμπτη θέση, δηλαδή να εκπροσωπεί το μοντέλο με αυτές τις μεταβλητές. Με αυτόν τον τρόπο θα ελέγξουμε τον αλγόριθμο ως προς τα αποτελέσματα που δίνει και θα προτείνουμε τιμές των παραμέτρων εισόδου για τις οποίες δύναται να επιλύσει το πρόβλημα επιλογής μεταβλητών σε πολλαπλά γραμμικά μοντέλα παλινδρόμησης με υψηλότερη αποδοτικότητα.
Όσον αφορά την αποδοτικότητα του γενετικού αλγορίθμου, υπάρχουν δύο τρόποι να μετρηθεί για τους εξελικτικούς αλγορίθμους: ο ένας σχετίζεται με την ποιότητα της λύσης και ο άλλος με την ταχύτητα σύγκλισης του αλγορίθμου. Συνήθως, οι προγραμματιστές ενδιαφέρονται για το συνδυασμό των δύο αυτών προσεγγίσεων και επιδιώκουν να βελτιστοποιήσουν τον αλγόριθμο έτσι ώστε να ικανοποιούνται και οι δύο σε κάποιο βαθμό. Η ποιότητα της λύσης μπορεί να μετρηθεί μέσω της συνάρτησης αξιολόγησης που χρησιμοποιεί ο αλγόριθμος, ενώ η ταχύτητα σύγκλισης μπορεί να μετρηθεί από τον αριθμό των υπολογισμών στη συνάρτηση αξιολόγησης, τον χρόνο CPU κλπ. Για μια και μοναδική εκτέλεση του αλγορίθμου, η αποδοτικότητα συνίσταται στα παρακάτω τρία στοιχεία:

- Δοθέντος ενός μέγιστου χρόνου εκτέλεσης (running time-computational effort), η αποδοτικότητα του αλγορίθμου προσδιορίζεται από τη βέλτιστη τιμή της συνάρτησης αξιολόγησης που επιστρέφεται.

- Δοθείσης μιας κατώτατης τιμής της συνάρτησης αξιολόγησης, η αποδοτικότητα του αλγορίθμου προσδιορίζεται ως ο χρόνος εκτέλεσης που απαιτείται για να συγκλίνει στην τιμή αυτή.

- Δοθέντος ενός μέγιστου χρόνου εκτέλεσης και μιας κατώτατης τιμής στη συνάρτηση αξιολόγησης, η αποδοτικότητα του αλγορίθμου προσδιορίζεται ως ο συνδυασμός των δύο ενδεχομένων: μια εκτέλεση κρίνεται ως επιτυχής εάν ο αλγόριθμος συγκλίνει στην κατώτατη τιμή εντός του δοθέντος χρόνου, διαφορετικά αποτυγχάνει.

Προφανώς, λόγω της στοχαστικής φύσης του γενετικού αλγορίθμου, πολλαπλές εκτελέσεις του αλγορίθμου είναι απαραίτητες προκειμένου να λάβουμε μια καλή εκτίμηση της αποδοτικότητας. Αν συναθροίσουμε τα παραπάνω κριτήρια για ένα συγκεκριμένο αριθμό εκτελέσεων, αποκτάμε τρία μέτρα αποδοτικότητας (performance metrics) που χρησιμοποιούνται συχνά στον προγραμματισμό εξελικτικών προγραμμάτων: τη μέση βέλτιστη τιμή αξιολόγησης MBF (Mean Best Fitness), τον μέσο αριθμό αξιολογήσεων μέχρι την εύρεση της λύσης AES (Average Number of Evaluations to Solution) και το ρυθμό επιτυχίας

SR (Success Rate). Για το πρόβλημα της επιλογής μεταβλητών που επιλύει η συγκεκριμένη διπλωματική, επιλέχθηκε ο ρυθμός επιτυχούς εύρεσης λύσης ή ισοδύναμα το ποσοστό στο οποίο ο αλγόριθμος εντόπισε τη βέλτιστη λύση σε ένα δεδομένο αριθμό εκτέλεσής του, εν προκειμένω για 10 εκτελέσεις.
Βέβαια, αξίζει να αναφερθεί ότι η απόδοση του γενετικού αλγορίθμου επηρεάζεται σημαντικά από τον παράγοντα της τύχης που υπεισέρχεται στον αλγόριθμο, με την έννοια ότι τα παραγόμενα δεδομένα είναι τυχαία. Αυτό σημαίνει ότι παράγονται από ένα συγκεκριμένο μοντέλο με κάποιον θόρυβο ή σφάλμα, και αυτό μπορεί να επιφέρει αλλοιώσεις στα αποτελέσματα του αλγορίθμου, ανεξαρτήτως των παραμέτρων που δίνονται ως είσοδος στον αλγόριθμο. Το πρώτο σύνολο δεδομένων με βάση το οποίο και κατασκευάστηκε ο αλγόριθμος περιλαμβάνει 50 παρατηρήσεις από 15 μεταβλητές, που προέρχονται από την κανονική κατανομή με μέση τιμή ίση με μηδέν και τυπική απόκλιση ένα. Οι 15 τυχαίες μεταβλητές διαδραματίζουν το ρόλο των ανεξάρτητων τυχαίων μεταβλητών για το πολλαπλό γραμμικό μοντέλο με μεταβλητές τις X_i (i=1,...,15). Προσομοιώνονται και 50 παρατηρήσεις της μεταβλητής απόκρισης Υ η οποία βασίζεται σε 5 από τις 15 τυχαίες επεξηγηματικές μεταβλητές σύμφωνα με τη σχέση:

$Y_i = 6 + 8 \cdot X_{i,1} + 3 \cdot X_{i,4} + 10 \cdot X_{i,7} - 12 \cdot X_{i,12} + 4 \cdot X_{i,15} + e_i$, για i=1,...,50 με $e_i \sim N(0, 1.5^2)$.
Για το συγκεκριμένο σύνολο δεδομένων το χρωμόσωμα που αντιστοιχεί στο βέλτιστο γραμμικό μοντέλο είναι προφανώς το (**1**,0,0,**1**,0,0,**1**,0,0,0,0,**1**,0,0,**1**).
Ένας πρώτος έλεγχος του αλγορίθμου γίνεται δίνοντας ως παραμέτρους τις παρακάτω:

- pop_size=50
- maxiter=30
- p_cross=1
- p_old=0.20
- elitism=TRUE
- p_mut=0.1

Τα αποτελέσματα που προκύπτουν με βάση την παραμετροποίηση αυτή παρουσιάζονται στο Σχήμα 5.1.

```
[1] "final population"
                                                    BIC_gen
 [1,] 1 0 0 1 0 0 1 0 0 0 0 1 0 0 1 216.5034
 [2,] 1 0 1 1 0 0 1 0 0 0 1 1 0 1 1 224.9584
 [3,] 1 0 1 1 0 1 1 1 1 0 1 1 1 0 1 236.1700
 [4,] 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 243.6382
 [5,] 1 0 0 0 1 0 1 0 0 1 0 1 0 0 1 287.5633
 [6,] 0 0 0 1 0 0 1 1 1 0 1 1 0 0 1 369.9576
 [7,] 0 0 0 0 0 1 1 1 0 1 0 1 1 0 1 385.5448
 [8,] 1 0 0 0 0 0 0 0 0 1 0 0 1 1 0 1 396.9270
 [9,] 1 1 1 0 1 0 0 0 1 0 1 1 0 0 1 403.3694
[10,] 1 0 0 1 0 1 0 0 0 1 1 1 1 1 1 405.9932
[11,] 1 0 1 0 0 1 0 1 0 0 1 1 0 0 1 409.5983
[12,] 1 1 0 0 0 1 1 1 0 1 1 0 0 1 1 420.1798
[13,] 1 0 1 0 0 0 1 0 0 0 1 0 1 1 1 420.2359
[14,] 0 1 0 1 1 0 0 0 1 0 1 1 0 0 1 420.6681
[15,] 1 1 0 1 0 1 0 1 1 0 0 1 0 1 0 424.4746
[16,] 0 0 1 1 0 0 0 1 0 0 1 0 1 0 0 424.4746
[17,] 0 1 0 0 0 0 0 0 1 1 1 0 1 1 0 424.4746
[18,] 0 1 1 1 0 1 1 1 0 1 0 1 0 1 0 424.4746
[19,] 1 1 1 0 1 1 0 0 1 0 0 1 1 1 0 424.4746
[20,] 0 1 0 0 1 0 1 1 1 1 1 1 0 1 0 424.4746
[21,] 0 1 0 1 1 0 0 1 0 0 1 0 0 0 0 424.4746
[22,] 0 1 1 1 1 1 0 0 0 0 0 1 0 0 0 424.4746
[23,] 0 1 1 1 1 1 1 1 0 1 1 1 1 0 0 424.4746
[24,] 0 0 1 1 0 1 1 1 0 0 1 1 0 1 0 424.4746
[25,] 1 0 0 0 1 1 0 0 0 1 1 1 1 0 0 424.4746
[26,] 0 0 1 0 1 0 1 0 1 1 0 0 0 0 1 427.6374
[27,] 0 1 1 0 0 0 0 0 0 1 1 1 1 0 0 1 429.0361
[28,] 1 1 1 0 1 1 0 1 0 1 0 0 0 1 1 441.5320
[29,] 0 0 1 0 0 1 0 1 1 1 0 0 0 0 1 442.0832
[30,] 1 1 0 0 1 0 0 0 1 0 1 0 0 1 1 442.5842
[31,] 0 0 0 0 1 0 0 0 0 0 0 1 0 1 1 1 444.1332
[32,] 1 0 1 1 0 1 0 0 1 1 1 0 1 1 1 447.4984
[33,] 0 0 1 0 1 1 0 0 0 1 1 0 1 0 1 447.7151
[34,] 0 1 0 1 1 0 0 1 0 1 1 0 1 0 1 447.7366
[35,] 0 0 1 0 0 0 0 0 0 1 1 1 0 1 1 1 448.1436
[36,] 0 1 0 0 1 0 0 1 1 1 1 0 0 1 1 451.3623
```

Σχήμα 5.1: Αποτελέσματα γενετικού αλγορίθμου για την 1η παραμετροποίηση

Παρατηρούμε ότι εμφανίζονται μόνο 36 χρωμοσώματα σε μορφή διανυσμάτων ακολουθούμενα από την τιμή του κριτηρίου αξιολόγησης που τους αντιστοιχεί διότι ο αλγόριθμος κρατά σε κάθε γενιά χρωμοσώματα που εμφανίζονται μια μόνο φορά αφαιρώντας τυχόν επαναλήψεις τους. Ως καλύτερο χρωμόσωμα ο γενετικός επιστρέφει το βέλτιστο χρωμόσωμα το οποίο αναμέναμε και το οποίο αντιστοιχεί στο γραμμικό μοντέλο παλινδρόμησης με τις μεταβλητές X_1, X_4, X_7, X_{12} και X_{15}. Η τιμή του κριτηρίου BIC για το εν λόγω μοντέλο είναι μικρότερη από αυτές των υπόλοιπων μοντέλων που διερεύνησε ο γενετικός αλγόριθμος και είναι ίση με 216.5034 για το συγκεκριμένο σύνολο δεδομένων. Αυτός είναι ένας πρώτος έλεγχος που μας οδηγεί στο συμπέρασμα ότι ο γενετικός αλγόριθμος δύναται να παρέχει ικανοποιητικά αποτελέσματα για το συγκεκριμένο πρόβλημα επιλογής μεταβλητών, ωστόσο κρίνεται αναγκαίο να ελεγχθεί αρκετές φορές και με διαφορετικά δεδομένα. Εκτός από τα αποτελέσματα της τελευταίας

επανάληψης τα οποία επέστρεψε ο αλγόριθμος, επιστρέφονται τα 3 καλύτερα χρωμοσώματα που εντοπίζονται σε κάθε γενιά αλλά και οι τιμές BIC των καλύτερων χρωμοσωμάτων κατά τη διάρκεια εκτέλεσης της διαδικασίας εύρεσης του καλύτερου χρωμοσώματος. Συγκεκριμένα, λαμβάνουμε το αποτέλεσμα του Σχήματος 5.2 στο οποίο παρατηρούμε ότι από την 4η κιόλας επανάληψη (γενιά) ο γενετικός αλγόριθμος εντόπισε το βέλτιστο μοντέλο.

```
[1] "best evaluations through all iterations"
[1] 224.8243 220.4125 219.9991 216.5034 216.5034 216.5034 216.5034 2
16.5034 216.5034
[10] 216.5034 216.5034 216.5034 216.5034 216.5034 216.5034 216.5034
216.5034 216.5034
[19] 216.5034 216.5034 216.5034 216.5034 216.5034 216.5034 216.5034
```

Σχήμα 5.2: Τιμές της αντικειμενικής συνάρτησης για το καλύτερο χρωμόσωμα της κάθε γενιάς

Ένας δεύτερος έλεγχος της απόδοσης του αλγορίθμου με ελαφρώς διαφορετική παραμετροποίηση (p_old=0.20 αντί p_old=0.30) δίνει τα αποτελέσματα που καταγράφονται στο Σχήμα 5.3.

```
[1] "final population"
                                                BIC_gen
 [1,] 1 0 0 1 0 0 1 0 0 0 0 1 0 0 1 203.1312
 [2,] 1 1 0 1 1 0 1 1 0 0 1 1 0 0 1 213.9069
 [3,] 1 1 1 1 1 0 0 1 1 1 0 1 1 0 0 1 219.5605
 [4,] 1 1 1 0 0 1 1 1 0 0 1 1 0 0 1 292.5840
 [5,] 0 1 0 1 0 0 1 0 0 0 1 1 0 0 1 379.1239
```

Σχήμα 5.3: Αποτελέσματα γενετικού αλγορίθμου για τη 2η παραμετροποίηση

Ενδεικτικά, παρουσιάζονται τα 5 πρώτα χρωμοσώματα που είναι και τα καλύτερα ως προς την τιμή BIC. Εδώ το καλύτερο χρωμόσωμα έχει τιμή αξιολόγησης ίση με 203.1312 που είναι διαφορετική από αυτήν του προηγούμενου βέλτιστου μοντέλου. Αυτό είναι λογικό, αφού το σύνολο δεδομένων είναι τυχαίο κάθε φορά και άρα αλλάζει. Όμοια, οι τιμές του BIC που χαρακτηρίζουν το βέλτιστο χρωμόσωμα σε κάθε γενιά παρουσιάζονται στο Σχήμα 5.4.
Ο γενετικός αλγόριθμος από την 3η κιόλας γενιά "επισκέπτεται" το βέλτιστο μοντέλο και το διατηρεί ως το καλύτερο για το συγκεκριμένο σύνολο δεδομένων μέχρι και την τελευταία γενιά.

```
[1] "best evaluations through all iterations"
 [1] 213.2989 210.7364 203.1312 203.1312 203.1312 203.1312 203.1312
203.1312 203.1312
[10] 203.1312 203.1312 203.1312 203.1312 203.1312 203.1312 203.1312
203.1312 203.1312
[19] 203.1312 203.1312 203.1312 203.1312 203.1312 203.1312 203.1312
```

Σχήμα 5.4: Τιμές της αντικειμενικής συνάρτησης για το καλύτερο χρωμόσωμα της κάθε γενιάς

5.2 Παραμετροποιήσεις γενετικού αλγορίθμου

Στην προηγούμενη ενότητα παρουσιάστηκαν δύο δυνατοί συνδυασμοί παραμέτρων και δοκιμάστηκε η απόδοση του γενετικού αλγορίθμου για τα προσομοιωμένα δεδομένα. Ωστόσο, αν ληφθεί υπόψη ότι υπάρχουν 7 παράμετροι στις οποίες ο χρήστης καλείται να δώσει τιμή (tuning), οι συνδυασμοί που μπορούν να δοκιμαστούν είναι πραγματικά πολλοί. Για μια πιο ενδελεχή μελέτη της απόδοσης του αλγορίθμου, δοκιμάστηκαν μερικοί συνδυασμοί παραμέτρων για κάθε ένα από τα 4 σύνολα δεδομένων που παράχθηκαν και κάθε συνδυασμός δοκιμάστηκε 10 φορές με τυχαία δεδομένα που προέρχονται από το εκάστοτε μοντέλο. Έτσι, βρέθηκε ένα ποσοστό επιτυχούς εύρεσης του βέλτιστου μοντέλου για κάθε μια παραμετροποίηση, και με αυτόν τον τρόπο ο χρήστης αποκτά μια καλύτερη εποπτεία σχετικά με την αποτελεσματικότητα του αλγορίθμου. Έτσι για παράδειγμα, αν για κάποιο συνδυασμό παραμέτρων έχουμε στις 10 δοκιμές, 7 επιτυχείς ευρέσεις του βέλτιστου μοντέλου, αποδίδουμε στον αλγόριθμο για τον συγκεκριμένο συνδυασμό παραμέτρων ποσοστό επιτυχίας 70%.

5.2.1 Πρώτο σύνολο δεδομένων

Θεωρούμε ένα δύνολο δεδομένων n=50 παρατηρήσεων από p=15 επεξηγηματικές μεταβλητές και υλοποιούμε 4 διαφορετικά σενάρια. Υπό το πρώτο σενάριο, ισχύει η ανεξαρτησία των μεταβλητών καθώς όλες οι μεταβλητές X_{ij} παράγονται από την πολυδιάστατη κανονική κατανομή με μέση τιμή 0 και πίνακα συνδιασποράς I_{15}, δηλαδή $X_{ij} \sim N(0, I_{15})$ ενώ η μεταβλητή απόκρισης προέρχεται από:

$Y_i \sim N(6+8\cdot X_{i,1}+3\cdot X_{i,4}+10\cdot X_{i,7}-12\cdot X_{i,12}+4\cdot X_{i,15}, 1.5^2)$, για i=1,...,50.

Το πρώτο σύνολο δεδομένων είναι η εύκολη περίπτωση στην οποία θα δοκιμαστεί ο γενετικός, και στο εξής θα καλείται ως **"εύκολο"** σύνολο δε-

δομένων.
Αρχικά, δοκιμάστηκαν συνδυασμοί παραμέτρων και για τους 3 διαφορετικούς τύπους διασταύρωσης με πιθανότητα διασταύρωσης ίση με 1, δηλαδή σε κάθε επανάληψη και άρα σε κάθε γενιά είναι σίγουρο ότι οι "γονείς" θα διασταυρωθούν. Στο Σχήμα 5.5 καταγράφονται ορισμένοι συνδυασμοί παραμέτρων που δοκιμάστηκαν για το εν λόγω σύνολο δεδομένων.

Pop_size	maxiter	Type_cross	P_old	elitism	P_mut	ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΟΥΣ ΕΥΡΕΣΗΣ
50	25	**1**	0.30	TRUE	0.1	100%
50	25	**2**	0.30	TRUE	0.1	40%
50	25	**3**	0.30	TRUE	0.1	40%
50	25	**1**	0.20	FALSE	0.05	80%
50	25	**2**	0.20	FALSE	0.05	40%
50	25	**3**	0.20	FALSE	0.05	40%

Σχήμα 5.5: Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου και το αντίστοιχο ποσοστό επιτυχούς εύρεσης λύσης για το πρώτο προσομοιωμένο σύνολο δεδομένων.

Για τα παραπάνω αλλά και για όλες τις δοκιμές και σε όλα τα σύνολα δεδομένων ο πληθυσμός σταθεροποιήθηκε στην τιμή 50, που σημαίνει ότι σε κάθε γενιά τα χρωμοσώματα που τελικά εξετάζει και μεταβιβάζει ο αλγόριθμος είναι 50. Η τιμή αυτή επιλέχθηκε με βάση τον αριθμό των μεταβλητών που εμπλέκονται στο πρόβλημα, και μετά από αρκετούς πειραματισμούς διαπιστώθηκε ότι το μέγεθος του πληθυσμού πρέπει να είναι μεγαλύτερο από το διπλάσιο του αριθμού των μεταβλητών για να παρέχει ικανοποιητικά αποτελέσματα ο γενετικός αλγόριθμος, και άρα θα πρέπει να είναι τουλάχιστον 30. Δεδομένου ότι ο αλγόριθμος μετά από μονοψήφιο αριθμό επαναλήψεων, δηλαδή γενεών, συγκλίνει σε κάποιο μοντέλο, επιλέχθηκε το μέγεθος του πληθυσμού να είναι ελαφρώς μεγαλύτερο

από το απαιτούμενο, αφού δεν προσθέτει σημαντικό χρόνο στην εκτέλεση.

Για την παράμετρο που καθορίζει τον τύπο διασταύρωσης, δοκιμάστηκαν και οι 3 μέθοδοι διασταύρωσης και φαίνεται ότι αποδίδει καλύτερα ο πρώτος τύπος διασταύρωσης για το πρώτο σύνολο δεδομένων. Όσον αφορά την παράμετρο που ορίζει το ποσοστό διατήρησης των παλαιών χρωμοσωμάτων, επιλέχθηκαν δύο τιμές, 20% και 30% αντίστοιχα. Η πρώτη είναι σχετικά μικρή και επιτρέπει την προσθήκη περισσότερων τυχαίων προς εξέταση χρωμοσωμάτων στον αλγόριθμο, ενώ η δεύτερη τιμή αξιοποιεί σε ποσοστό 30% τα ήδη υπάρχοντα χρωμοσώματα του πληθυσμού. Για τον συγκεκριμένο τύπο δεδομένων, το ποσοστό 30% δίνει περισσότερες φορές το βέλτιστο μοντέλο παλινδρόμησης και άρα προτιμάται. Η παράμετρος που ελέγχει την εφαρμογή ή μη του ελιτισμού μπορεί προφανώς να πάρει μόνο δύο τιμές, TRUE (αληθής) ή FALSE (ψευδής). Έπειτα από δοκιμές, καλύτερα αποτελέσματα προκύπτουν με εφαρμογή της ελιτίστικης στρατηγικής, κάτι το οποίο είναι αναμενόμενο, διότι έτσι εξασφαλίζεται ότι αποδεδειγμένα καλές λύσεις δε θα χαθούν με το πέρας των επαναλήψεων.
Τέλος, η παράμετρος της πιθανότητας μετάλλαξης έλαβε τιμές 0.05 και 0.1 αντίστοιχα, που σημαίνει ότι κάθε συντεταγμένη του εκάστοτε χρωμοσώματος μπορεί να αλλάξει τιμή από 0 σε 1 και αντίστροφα με πιθανότητα 5% ή 10%. Από τη μία πλευρά η μικρή πιθανότητα μετάλλαξης βοηθά στο να μην αλλοιώνονται τα υπάρχοντα χρωμοσώματα, αλλά από την άλλη πλευρά αυτό μειώνει τη δυνατότητα διερεύνησης του χώρου των λύσεων.

Δεδομένου ότι ο τύπος διασταύρωσης ενός σημείου βρίσκει περισσότερες φορές το βέλτιστο μοντέλο για το πρώτο σύνολο δεδομένων, συνεχίζουμε με περισσότερες δοκιμές παραμέτρων για τον εν λόγω τύπο διασταύρωσης και τα αποτελέσματα συγκεντρώνονται στο Σχήμα 5.6.

Pop_size	maxiter	Type_cross	P_old	elitism	P_mut	ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΟΥΣ ΕΥΡΕΣΗΣ
50	25	1	0.30	TRUE	0.1	100%
50	25	1	0.20	TRUE	0.1	80%
50	25	1	0.20	FALSE	0.1	70%
50	25	1	0.30	FALSE	0.05	60%
50	25	1	0.10	TRUE	0.2	60%

Σχήμα 5.6: Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου για διασταύρωση ενός σημείου και το αντίστοιχο ποσοστό επιτυχούς εύρεσης λύσης για το πρώτο προσομοιωμένο σύνολο δεδομένων.

Από το Σχήμα 5.6 και για τύπο διασταύρωσης ενός σημείου (κωδικοποίηση 1), ο γενετικός αλγόριθμος βρίσκει σε ποσοστό 60% έως και 100% το βέλτιστο μοντέλο. Συγκεκριμένα, όταν η πιθανότητα μετάλλαξης είναι αρκετά μικρή (περίπτωση 5%) ή αρκετά μεγάλη (περίπτωση 20%) ο αλγόριθμος αποδίδει σε ποσοστό 60% ενώ όταν παίρνει μια ενδιάμεση τιμή (περίπτωση 10%) το ποσοστό επιτυχίας του αυξάνεται και κυμαίνεται από 70% μέχρι και 100%. Επίσης, η αύξηση του ποσοστού παλαιών χρωμοσωμάτων από 20% σε 30% αυξάνει την απόδοση του αλγορίθμου και αυτό αντικατοπτρίζεται στο ποσοστό εύρεσης βέλτιστης λύσης που από 80% γίνεται 100%, όταν εφαρμόζεται ελιτισμός και η πιθανότητα μετάλλαξης οριστεί ίση με 0.1. Όσον αφορά την εφαρμογή της ελιτίστικης στρατηγικής, φαίνεται ότι αυξάνει την αποδοτικότητα του αλγορίθμου και αυτό φαίνεται από την σύγκριση των παραμετροποιήσεων στις οποίες μεταβάλλεται μόνο η παράμετρος που ελέγχει τον ελιτισμό.
Η τιμή της πιθανότητας διασταύρωσης επιλέχθηκε να είναι ίση με τη μονάδα σε όλες τις δοκιμές του Σχήματος 5.6, καθώς διαπιστώθηκε ότι αυτή η τιμή καθιστά τον γενετικό αλγόριθμο πιο αποδοτικό. Αυτό είναι αναμενόμενο, διότι όπως διατυπώθηκε και στο 2ο Κεφάλαιο, η μεγάλη πιθανότητα διασταύρωσης ερμηνεύεται ως πληρέστερη εξερεύνηση του χώρου των υποψήφιων

λύσεων. Ωστόσο, υπενθυμίζεται ότι η πιθανότητα αυτή τροποποιείται και σταθεροποιείται σε άλλη τιμή (p_cross=0.5) από την 11η επανάληψη και μετά, όταν δηλαδή για τα συγκεκριμένα προβλήματα της επιλογής μεταβλητών ο αλγόριθμος φαίνεται να προσεγγίζει την βέλτιστη γι αυτόν λύση. Το συμπέρασμα αυτό επιβεβαιώνεται και πειραματικά στα Σχήματα 5.7 και 5.8 τα οποία συνοψίζουν τις δοκιμές των παραμέτρων για p_cross=0.5 και p_cross=0.8 αντίστοιχα. Οι δοκιμές έγιναν για τύπο διασταύρωσης ενός σημείου, ο οποίος αποδείχθηκε καταλληλότερος για το συγκεκριμένο τύπο δεδομένων.

Pop_size	maxiter	Type_cross	P_old	elitism	P_mut	ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΟΥΣ ΕΥΡΕΣΗΣ
50	25	1	0.30	TRUE	0.1	60%
50	25	1	0.20	TRUE	0.1	50%

Σχήμα 5.7: Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου και ποσοστό επιτυχούς εύρεσης λύσης για το πρώτο σύνολο δεδομένων και για πιθανότητα διασταύρωσης ίση με 50%.

Pop_size	maxiter	Type_cross	P_old	elitism	P_mut	ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΟΥΣ ΕΥΡΕΣΗΣ
50	25	1	0.30	TRUE	0.1	80%
50	25	1	0.20	TRUE	0.1	70%

Σχήμα 5.8: Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου και ποσοστό επιτυχούς εύρεσης λύσης για το πρώτο σύνολο δεδομένων και για πιθανότητα διασταύρωσης ίση με 80%.

Συμπερασματικά λοιπόν, η υψηλότερη πιθανότητα διασταύρωσης καθιστά τον αλγόριθμο πιο αποδοτικό, με την έννοια ότι έχει περισσότερες πιθανότητες να επιτύχει εντοπίζοντας το σωστό μοντέλο. Από τους τρεις παραπάνω πίνακες αποδεικνύεται ότι μια παραμετροποίηση που είναι αξιόπιστη σε ποσοστό 100% είναι η εξής: pop_size=50, maxiter=25, p_cross=1, type_cross=1, p_old=0.30, elitism=TRUE, p_mut=0.1 Αξίζει να τονιστεί ότι μπορεί να υπάρχουν και άλλες παραμετροποιήσεις που να βρίσκουν το βέλτιστο μοντέλο σε κάθε δοκιμή του αλγορίθμου.

5.2.2 Δεύτερο σύνολο δεδομένων

Στο δεύτερο σύνολο δεδομένων προσομοιώνουμε και πάλι n=50 παρατηρήσεις από p=15 τυχαίες μεταβλητές. Υπό το δεύτερο σενάριο, ισχύει η ανεξαρτησία των μεταβλητών καθώς όλες οι μεταβλητές X_{ij} παράγονται από την πολυδιάστατη κανονική κατανομή με μέση τιμή 0 και πίνακα συνδιασποράς I_{15} , δηλαδή $X_{ij} \sim N(0, I_{15})$ ενώ η μεταβλητή απόκρισης προέρχεται από:

$Y_i \sim N(6+8\cdot X_{i,1}+3\cdot X_{i,4}+10\cdot X_{i,7}-12\cdot X_{i,12}+4\cdot X_{i,15}, 2.5^2)$, για i=1,...,50.

Το δεύτερο σύνολο δεδομένων είναι μικρής δυσκολίας περίπτωση στην οποία θα δοκιμαστεί ο γενετικός, και στο εξής θα καλείται ως **"μικρής δυσκολίας"** σύνολο δεδομένων, λόγω της ελαφρώς αυξημένης διασποράς στα Y_i.
Η μόνη διαφορά με το προηγούμενο σύνολο δεδομένων είναι ότι τώρα η τυπική απόκλιση της μεταβλητής Υ αυξήθηκε κατά μια μονάδα και έγινε ίση με 2.5.
Για τις δοκιμές των παραμέτρων που φαίνονται στο Σχήμα 5.9 η τιμή της πιθανότητας διασταύρωσης (p_cross) σταθεροποιήθηκε στη μονάδα. Στο Σχήμα 5.9 γίνονται ενδεικτικά δοκιμές και για τους τρεις τύπους διασταύρωσης, έτσι ώστε να ελέγξουμε ποιος βρίσκει με μεγαλύτερο ποσοστό επιτυχίας το βέλτιστο μοντέλο.

Pop_size	maxiter	Type_cross	P_old	elitism	P_mut	ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΟΥΣ ΕΥΡΕΣΗΣ
50	25	**1**	0.30	TRUE	0.1	50%
50	25	**2**	0.30	TRUE	0.1	80%
50	25	**3**	0.30	TRUE	0.1	50%
50	25	**1**	0.20	FALSE	0.05	70%
50	25	**2**	0.20	FALSE	0.05	60%
50	25	**3**	0.20	FALSE	0.05	50%

Σχήμα 5.9: Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου και το αντίστοιχο ποσοστό επιτυχούς εύρεσης λύσης για το δεύτερο προσομοιωμένο σύνολο δεδομένων.

Όπως φαίνεται από το Σχήμα 5.9 ο δεύτερος συνδυασμός παραμέτρων με τον δεύτερο τύπο διασταύρωσης είναι πιο αποδοτικός για το συγκεκριμένο σύνολο δεδομένων με την έννοια ότι βρίσκει σε ποσοστό 80% το βέλτιστο γραμμικό μοντέλο παλινδρόμησης. Για το λόγο αυτό προχωράμε σε περισσότερες δοκιμές που αφορούν τον συγκεκριμένο τύπο διασταύρωσης στο Σχήμα 5.10.

Pop_size	maxiter	Type_cross	P_old	elitism	P_mut	ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΟΥΣ ΕΥΡΕΣΗΣ
50	25	2	0.30	TRUE	0.1	80%
50	25	2	0.20	TRUE	0.1	60%
50	25	2	0.30	FALSE	0.1	80%
50	25	2	0.40	TRUE	0.05	100%
50	25	2	0.10	FALSE	0.05	50%

Σχήμα 5.10: Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου για διασταύρωση δύο σημείων και το αντίστοιχο ποσοστό επιτυχούς εύρεσης λύσης για το δεύτερο προσομοιωμένο σύνολο δεδομένων

Από το Σχήμα 5.10 και για τύπο διασταύρωσης δύο σημείων (κωδικοποίηση 2), ο γενετικός αλγόριθμος βρίσκει σε ποσοστό 100% το βέλτιστο μοντέλο όταν επιλεγεί ελιτισμός, το ποσοστό της παλαιάς γενιάς που μεταφέρεται στη νέα είναι το 40% του πληθυσμού και συμβαίνει μετάλλαξη με πιθανότητα 5% σε κάθε γονίδιο του χρωμοσώματος. Για το συγκεκριμένο σύνολο δεδομένων για τα οποία υπάρχει υψηλότερη διασπορά των παρατηρήσεων της μεταβλητής απόκρισης, προκύπτει ότι όσο αυξάνεται το ποσοστό των χρωμοσωμάτων που μεταβιβάζεται στην επόμενη γενιά, τόσο αυξάνεται και η αποδοτικότητα του γενετικού.

5.2.3 Τρίτο σύνολο δεδομένων

Στο τρίτο σύνολο δεδομένων προσομοιώνουμε n=50 παρατηρήσεις από p=15 τυχαίες μεταβλητές. Υπό το τρίτο σενάριο, η μεταβλητή απόκρισης παράγεται από:
$Y_i \sim N(4 + 2 \cdot X_{i,1} - 1 \cdot X_{i,5} + 1.5 \cdot X_{i,7} + 1 \cdot X_{i,11} + 0.5 \cdot X_{i,13} + 1 \cdot X_{i,15}, 1.5^2)$,
για i=1,...,50.

Ωστόσο, αυτή τη φορά μόνο οι 10 πρώτες μεταβλητές παράγονται παράγονται από την πολυδιάστατη κανονική κατανομή με μέση τιμή 0 και πίνακα συνδιασποράς I_{10}, ενώ οι υπόλοιπες 5 προέρχονται από:
$X_{ij} \sim N(0.3 \cdot X_{i,1} + 0.5 \cdot X_{i,2} + 0.7 \cdot X_{i,3} + 0.9 \cdot X_{i,4} + 1.1 \cdot X_{i,5}, 1)$ για j=11,...,15, i=1,...,50.
Εμφανίζεται δηλαδή το φαινόμενο της πολυσυγγραμμικότητας.
Το τρίτο σύνολο δεδομένων είναι μέτριας δυσκολίας περίπτωση στην οποία θα δοκιμαστεί ο γενετικός, και στο εξής θα καλείται ως **"μέτριας δυσκολίας"** σύνολο δεδομένων, λόγω της πολυσυγγγραμμικότητας στα X_{ij}.
Στο Σχήμα 5.11 γίνονται ενδεικτικά κάποιες δοκιμές και για τους τρεις τύπους διασταύρωσης, έτσι ώστε να ελέγξουμε ποιος βρίσκει με μεγαλύτερο ποσοστό επιτυχίας το βέλτιστο μοντέλο.

Pop_size	maxiter	Type_cross	P_old	elitism	P_mut	ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΟΥΣ ΕΥΡΕΣΗΣ
50	25	**1**	0.30	TRUE	0.1	40%
50	25	**2**	0.30	TRUE	0.1	70%
50	25	**3**	0.30	TRUE	0.1	60%
50	25	**1**	0.20	FALSE	0.05	10%
50	25	**2**	0.20	FALSE	0.05	30%
50	25	**3**	0.20	FALSE	0.1	30%

Σχήμα 5.11: Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου και το αντίστοιχο ποσοστό επιτυχούς εύρεσης λύσης για το τρίτο προσομοιωμένο σύνολο δεδομένων

Όπως υποδεικνύεται από το Σχήμα 5.11, ο δεύτερος τύπος διασταύρωσης (διασταύρωση δύο σημείων) δίνει μεγαλύτερο ποσοστό εύρεσης της βέλτιστης λύσης που σημαίνει ότι με μεγαλύτερη πιθανότητα και συγκεκριμένα με πιθανότητα 70% ο γενετικός αλγόριθμος θα επιστρέψει ως έξοδο το βέλτιστο μοντέλο. Για το λόγο αυτό, συνεχίζουμε στο Σχήμα 5.12 με περισσότερες παραμετροποιήσεις που πραγματοποιούν αυτόν τον τύπο διασταύρωσης.

Pop_size	maxiter	Type_cross	P_old	elitism	P_mut	ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΟΥΣ ΕΥΡΕΣΗΣ
50	25	**2**	0.30	TRUE	0.08	70%
50	25	**2**	0.20	TRUE	0.1	80%
50	25	**2**	0.20	FALSE	0.1	100%
50	25	**2**	0.30	FALSE	0.05	80%
50	25	**2**	0.10	TRUE	0.1	60%

Σχήμα 5.12: Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου για διασταύρωση δύο σημείων και το αντίστοιχο ποσοστό επιτυχούς εύρεσης λύσης για το τρίτο προσομοιωμένο σύνολο δεδομένων

Από το Σχήμα 5.12 συμπεραίνουμε ότι για τύπο διασταύρωσης δύο σημείων, με ποσοστό παλαιών χρωμοσωμάτων 20%, χωρίς εφαρμογή ελιτίστικης στρατηγικής και με πιθανότητα μετάλλαξης ίση με 10% για κάθε γονίδιο του χρωμοσώματος, ο αλγόριθμος εντοπίζει το βέλτιστο μοντέλο σε κάθε σύνολο δεδομένων με τα συγκεκριμένα χαρακτηριστικά της παρούσας ενότητας. Μάλιστα, σε αντίθεση με τα δύο προηγούμενα σύνολα δεδομένων στα οποία η εφαρμογή της ελιτίστικης στρατηγικής διευκόλυνε τον αλγόριθμο στην εύρεση της λύσης, εδώ η μη εφαρμογή της φαίνεται να αποδίδει καλύτερα. Συγκεκριμένα, για ποσοστό μεταφοράς χρωμοσωμάτων στην επόμενη γενιά ίσο με 20% και πιθανότητα μετάλλαξης ίση με 10%, ο γενετικός βρίσκει με μεγαλύτερη βεβαιότητα κατά 20% τη βέλτιστη λύση όταν δεν εφαρμόζεται ελιτίστικη στρατηγική σε σχέση με την περίπτωση στην οποία εφαρμόζεται. Βέβαια, δεδομένης της τυχαιότητας που υπεισέρχεται σε κάθε επανάληψη του αλγορίθμου αλλά και στον τρόπο που παράχθηκαν τα δεδομένα, αυτό το συμπέρασμα δεν είναι απόλυτο και ενδέχεται να είναι αποτέλεσμα που προέκυψε από τα συγκεκριμένα δεδομένα που έτυχε να παραχθούν στο συγκεκριμένο "τρέξιμο" του αλγορίθμου.

5.2.4 Τέταρτο σύνολο δεδομένων

Στο τέταρτο σύνολο δεδομένων προσομοιώνουμε n=50 παρατηρήσεις από p=15 τυχαίες μεταβλητές. Υπό το τέταρτο σενάριο, η μεταβλητή απόκρισης παράγεται από:
$Y_i \sim N(4 + 2 \cdot X_{i,1} - 1 \cdot X_{i,5} + 1.5 \cdot X_{i,7} + 1 \cdot X_{i,11} + 0.5 \cdot X_{i,13} + 1 \cdot X_{i,15}, 2.5^2)$, για i=1,...,50.

Και αυτή τη φορά μόνο οι 10 πρώτες μεταβλητές παράγονται από την πολυδιάστατη κανονική κατανομή με μέση τιμή 0 και πίνακα συνδιασποράς I_{15} , ενώ οι υπόλοιπες 5 προέρχονται από:
$X_{ij} \sim N(0.3 \cdot X_{i,1} + 0.5 \cdot X_{i,2} + 0.7 \cdot X_{i,3} + 0.9 \cdot X_{i,4} + 1.1 \cdot X_{i,5}, 1)$ για j=11,...,15, i=1,...,50.
Εμφανίζεται και εδώ το φαινόμενο της πολυσυγγραμμικότητας.
Το τέταρτο σύνολο δεδομένων είναι μεγάλης δυσκολίας περίπτωση στην οποία θα δοκιμαστεί ο γενετικός, και στο εξής θα καλείται ως **"μεγάλης δυσκολίας"** σύνολο δεδομένων, λόγω της πολυσυγγγραμμικότητας στα X_{ij} αλλά και της μεγάλης διασποράς στα Y_i.
Στο Σχήμα 5.13 γίνονται ενδεικτικά δοκιμές και για τους τρεις τύπους διασταύρωσης με δεδομένα που έχουν τα παραπάνω χαρακτηριστικά. Παρατηρούμε ότι τα ποσοστά επιτυχίας του γενετικού αλγορίθμου είναι απογοητευτικά, διότι αποτυγχάνει παταγωδώς στην εύρεση του μοντέλου από τα οποία προήλθαν τα δεδομένα. Ενδεικτικά, για μία εκτέλεση του αλγορίθμου με τιμές των παραμέτρων όπως ορίζονται από την 3η γραμμή του Σχήματος 5.13, προκύπτει το χρωμόσωμα [1,0,0,0,1,0,1,0,0,0,1,0,0,0,1] που περιέχει όλες τις μεταβλητές του αληθινού μοντέλου, εκτός από την 13η. Στο μοντέλο αυτό η τιμή BIC βρέθηκε ίση με 272.9643, ενώ η τιμή BIC για το πραγματικό μοντέλο είναι ίση με 274.8592. Παρατηρούμε λοιπόν ότι ο γενετικός "προτείνει" ένα πιο φειδωλό μοντέλο, με μία μεταβλητή λιγότερη και παρά το γεγονός ότι αποτυγχάνει σε κάθε προσπάθειά του να εντοπίσει το σωστό μοντέλο με τη δεδομένη παραμετροποίηση, η τιμή της αντικειμενικής συνάρτησης είναι πολύ κοντά στο ολικό ελάχιστο. Αυτό το περιμέναμε και δικαιολογείται πλήρως, διότι ο γενετικός έχει ως κριτήριο την ελαχιστοποίηση ως προς BIC επομένως επιδιώκει να εντοπίσει το μοντέλο με τη μικρότερη τιμή BIC. Επίσης, αξίζει να παρατηρήσουμε ότι η επίδραση της 13η μεταβλητής την οποία αποτυγχάνει να συμπεριλάβει ο γενετικός αλγόριθμος στο μοντέλο, είναι ασθενής σε σχέση με αυτή των υπόλοιπων μεταβλητών στο πραγματικό μοντέλο από το οποίο προήλθαν τα δεδομένα, διότι ο συντελεστής της είναι ίσος με 0.5 ενώ οι υπόλοιποι συντελεστές είναι 2, 1 ή 1.5. Ενδέχεται, κάποιος άλλος αλγόριθμος στοχαστικής βελτιστοποίησης όπως η βελτιστοποίηση σμήνους σωματιδίων (particle swarm optimisation) ή η προσομοιωμένη ανόπτηση (simulated annealing) να απέδιδε

Pop_size	maxiter	Type_cross	P_old	elitism	P_mut	ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΟΥΣ ΕΥΡΕΣΗΣ
50	25	1	0.30	TRUE	0.08	10%
50	25	2	0.20	FALSE	0.1	10%
50	25	3	0.30	TRUE	0.1	0%
50	25	1	0.30	FALSE	0.05	10%
50	25	2	0.30	FALSE	0.05	0%
50	25	3	0.20	TRUE	0.1	0%

Σχήμα 5.13: Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου και το αντίστοιχο ποσοστό επιτυχούς εύρεσης λύσης για το τέταρτο προσομοιωμένο σύνολο δεδομένων

καλύτερα για σύνολα δεδομένων με υψηλή μεταβλητότητα και πολυσυγγραμμικότητα *(Eberhart, R. C. & Kennedy, J.(2001). Swarm Intelligence, Morgan Kaufmann)*, *(Jeong, I.-S.et al. (2016). A Feature Selection Approach Based on Simulated Annealing for Detecting Various Denial of Service Attacks. River Publishers)*.

Συμπερασματικά, από την ανάλυση των αποτελεσμάτων που προηγήθηκε φαίνεται πως ο γενετικός αλγόριθμος αποδίδει καλύτερα όταν:

1. Η πιθανότητα διασταύρωσης είναι κοντά στο 1, αν όχι ίση με 1, για την καλύτερη εξερεύνηση του χώρου των λύσεων. Διαφορετικά, θα πρέπει να αυξηθεί το πλήθος των επαναλήψεων σε συνδυασμό ίσως με το μέγεθος του πληθυσμού και αυτό θα αυξήσει τον υπολογιστικό χρόνο του αλγορίθμου.

2. Ο τύπος διασταύρωσης είναι συνήθως δύο σημείων, δηλαδή τύπος διασταύρωσης ελαφρώς πιο σύνθετος από του ενός σημείου

3. Δεν είναι απαραίτητο να εφαρμόζεται ελιτισμός καθώς οι υπόλοιπες παράμετροι και κυρίως το ποσοστό των ικανών χρωμοσωμάτων που μεταφέρεται στην επόμενη γενιά βοηθούν τον αλγόριθμο να αποδώσει ικανοποιητικά

4. Το μέγεθος του πληθυσμού δεν είναι πολύ μικρό, ώστε να μπορούν να γίνονται αρκετές διασταυρώσεις σε κάθε επανάληψη

5. Η πιθανότητα μετάλλαξης κυμαίνεται στο $5 - 10\%$, καθώς μεγαλύτερες τιμές αλλοιώνουν σημαντικά τα χρωμοσώματα και καθυστερούν τη σύγκλιση, ενώ μικρότερες τιμές είναι δυνατό να εγκλωβίσουν τον αλγόριθμο σε περιορισμένες περιοχές του χώρου των λύσεων και το ψάξιμο να μην είναι επαρκές

6. Το ποσοστό των χρωμοσωμάτων που διατηρούνται στην επόμενη γενιά κυμαίνεται στο $20 - 40\%$, έτσι ώστε ο αλγόριθμος να οδηγείται σε πιο επιτυχημένες διασταυρώσεις με το πέρας των γενεών και να μην χάνονται αποδεδειγμένα καλά άτομα καθώς προχωράει ο αλγόριθμος

Συνοψίζουμε τα παραπάνω συμπεράσματα για κάθε ένα σύνολο δεδομένων στο Σχήμα 5.14.

	Pop_size	maxiter	P_cross	Type_cross	P_old	elitism	P_mut	Ποσοστό Επιτυχίας
Εύκολο	50	25	100%	1	20-30%	**NAI**	10%	80-100%
Μικρής δυσκολίας	50	25	100%	2	**30-40%**	NAI/OXI	5-10%	80-100%
Μέτριας δυσκολίας	50	25	100%	2	20-30%	**OXI**	5-10%	80-100%
Μεγάλης δυσκολίας	50	25	100%	1,2	20-30%	NAI/OXI	8-10%	10%

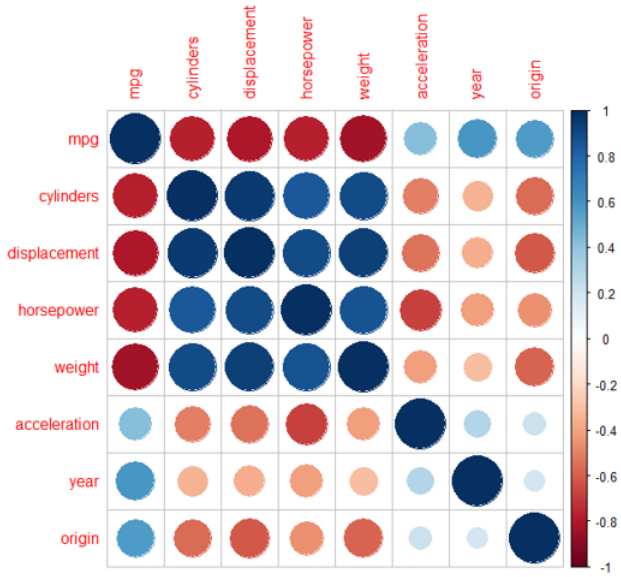
Σχήμα 5.14: Συγκεντρωτικός πίνακας βέλτιστων παραμετροποιήσεων για κάθε σύνολο δεδομένων

5.2.5 Σύνολο δεδομένων "Auto"

Εξετάζουμε την αποτελεσματικότητα του αλγορίθμου σε ένα ακόμη σύνολο δεδομένων το οποίο βρίσκεται στο πακέτο "MASS" της R και περιλαμβάνει δεδομένα για 392 αυτοκίνητα και 9 συνολικά μεταβλητές. Συγκεκριμένα, τα δεδομένα περιέχουν πληροφορίες σχετικά με τα μίλια ανά γαλόνι (mpg), τον αριθμό των κυλίνδρων (cylinders), τη μετατόπιση του κινητήρα (engine displacement), την ιπποδύναμη (horsepower), το βάρος του οχήματος (weight), τον χρόνο επιτάχυνσης (acceleration), το έτος παραγωγής του μοντέλου (year), την προέλευση του αυτοκινήτου (origin) και το όνομα του οχήματος (name). Από τις παραπάνω μεταβλητές, οι mpg, displacement, horsepower, weight, acceleration είναι συνεχείς, ενώ οι cylinders, year, origin είναι διακριτές. Το

πραγματικό dataset περιείχε 408 παρατηρήσεις, όμως 16 από αυτές αφαιρέθηκαν από το δείγμα διότι περιείχαν ελλιπείς τιμές (missing values). Στόχος είναι να εξάγουμε ένα μοντέλο που να μπορεί να εκτιμήσει κατάλληλα την αποδοτικότητα καυσίμου (mpg) από διαφορετικά μοντέλα αυτοκινήτων όταν είναι γνωστές οι τιμές των παραπάνω παραμέτρων. Με άλλα λόγια, η μεταβλητή mpg είναι η μεταβλητή απόκρισης (output) ενώ οι υπόλοιπες είναι οι επεξηγηματικές μεταβλητές (explanatory variables). Για την εύρεση του βέλτιστου γραμμικού μοντέλου που περιγράφει τη σχέση μεταξύ των μεταβλητών θα χρησιμοποιηθεί ο γενετικός αλγόριθμος και θα επαληθευτεί το αποτέλεσμα που θα δώσει από τον αλγόριθμο της εξαντλητικής αναζήτησης.

Προτού εκτελέσουμε τον γενετικό αλγόριθμο, αφαιρούμε από τον πίνακα των δεδομένων την μεταβλητή "mpg" την οποία θέλουμε να προβλέψουμε μέσω του μοντέλου παλινδρόμησης που θα κατασκευάσουμε και τη μεταβλητή "name" διότι δεν προσδίδει κάποια πληροφορία στο μοντέλο. Επομένως οι επεξηγηματικές μεταβλητές πλέον είναι 7, και συγκεκριμένα οι:
cylinders, displacement, horsepower, weight, acceleration, year και origin. Αρχικά ελέγχουμε αν υπάρχουν συσχετίσεις των παραπάνω μεταβλητών μέσω ενός κατάλληλου γραφήματος στην R, που φαίνεται στο Σχήμα 5.15.



Σχήμα 5.15: Πίνακας συσχετίσεων μεταξύ των μεταβλητών ανά μία

Από το Σχήμα 5.15 παρατηρούμε ότι υπάρχουν έντονες θετικές συσχετίσεις μεταξύ ορισμένων επεξηγηματικών μεταβλητών. Ενδεικτικά, αναφέρουμε ότι οι

μεταβλητές cylinders και displacement,weight και cylinders, weight και displacement παρουσιάζουν έντονη θετική συσχέτιση.
Λόγω των εν λόγω συσχετίσεων, θα διαλέξουμε τιμές για τις παραμέτρους που δέχεται ως είσοδο ο γενετικός αλγόριθμος τέτοιες ώστε να αποδίδει με υψηλό ποσοστό επιτυχίας για σύνολα δεδομένων στα οποία οι επεξηγηματικές μεταβλητές συσχετίζονται μεταξύ τους, και άρα θα επιλέξουμε από τον πίνακα 5.11. Επιλέγουμε την παραμετροποίηση που εμφάνισε το μεγαλύτερο ποσοστό επιτυχίας (100%) για το τρίτο σύνολο δεδομένων υπό την ύπαρξη πολυσυγγραμμικότητας, δηλαδή:

- pop_size=50
- maxiter=25
- p_cross=1
- type_cross=2
- p_old=0.20
- elitism=FALSE
- p_mut=0.1

Τα αποτελέσματα του γενετικού αλγορίθμου για την αναζήτηση του βέλτιστου γραμμικού μοντέλου παλινδρόμησης συνοψίζονται στο Σχήμα 5.16.

Όπως φαίνεται από τα αποτελέσματα του Σχήματος 5.16 το μοντέλο που προτείνει ο γενετικός ως βέλτιστο είναι το μοντέλο με την 4η, 6η και 7η μεταβλητή, δηλαδή τις weight, year και origin.
Επαληθεύουμε τα αποτελέσματα αυτά και μέσω της εκτέλεσης του αλγορίθμου της εξαντλητικής αναζήτησης και το αποτέλεσμα που προκύπτει παρουσιάζεται στο Σχήμα 5.17.

```
[1] "best evaluations through all iterations"
 [1] 2090.218 2090.130 2090.130 2085.548 2085.548 2085.548 2085.548 2085.548 2085.548
[10] 2085.548 2085.548 2085.548 2085.548 2085.548 2085.548 2085.548 2085.548 2085.548
[19] 2085.548 2085.548 2085.548 2085.548 2085.548 2085.548 2085.548
[1] "final population"
                     BIC_gen
 [1,] 0 0 0 1 0 1 1 2085.548
 [2,] 1 1 0 1 0 1 1 2094.346
 [3,] 1 1 1 1 0 1 1 2095.397
 [4,] 1 1 1 0 1 1 1 2184.294
 [5,] 1 1 0 0 0 1 1 2206.258
 [6,] 0 1 0 0 1 1 1 2208.846
 [7,] 1 1 0 0 1 1 1 2210.717
 [8,] 1 0 0 0 0 1 1 2225.146
 [9,] 0 1 0 1 0 0 1 2273.183
[10,] 1 1 0 0 0 0 1 2331.415
[11,] 0 0 0 1 1 1 0 2726.383
[12,] 0 0 0 1 0 0 0 2726.383
[13,] 1 0 0 0 0 1 0 2726.383
[14,] 0 1 0 1 1 0 0 2726.383
[15,] 1 1 0 0 1 1 0 2726.383
[16,] 0 1 0 0 1 1 0 2726.383
[17,] 0 1 0 0 0 1 0 2726.383
[18,] 0 0 1 0 0 0 0 2726.383
[19,] 0 0 0 0 0 1 0 2726.383
[20,] 1 1 0 0 0 1 0 2726.383
```

Σχήμα 5.16: Αποτελέσματα γενετικού αλγορίθμου για το σύνολο δεδομένων “Auto”

```
[1] "best model"
    var1 var2 var3 var4 var5 var6 var7
105    0    0    0    1    0    1    1
```

Σχήμα 5.17: Βέλτιστο γραμμικό μοντέλο παλινδρόμησης σύμφωνα με τον αλγόριθμο πλήρους αναζήτησης

Κεφάλαιο 6

ΠΡΟΕΚΤΑΣΕΙΣ

Όπως σχολιάστηκε και στο 5ο Κεφάλαιο, η εφαρμογή του γενετικού αλγορίθμου προϋποθέτει την απόδοση κατάλληλων τιμών στις παραμέτρους του. Οι γενετικοί αλγόριθμοι, όπως και οι περισσότεροι αλγόριθμοι στοχαστικής βελτιστοποίησης, απαιτούν την αρχικοποίηση αρκετών παραμέτρων ώστε να μπορούν να υλοποιηθούν. Στην παρούσα διπλωματική εργασία, ο γενετικός αλγόριθμος απαιτεί την αρχικοποίηση επτά παραμέτρων για να εκτελεστεί, και όπως αναφέρθηκε αυτές είναι το μέγεθος του πληθυσμού, ο αριθμός των γενεών-επαναλήψεων, η πιθανότητα διασταύρωσης, ο τύπος διασταύρωσης, η εφαρμογή ή μη του ελιτισμού, το ποσοστό παλαιών ατόμων που κρατείται και η πιθανότητα μετάλλαξης.
Ωστόσο, παρά το γεγονός ότι ο γενετικός αλγόριθμος βασίζεται στην ιδέα της φυσικής εξέλιξης των ειδών και η βασική δομή του είναι αυτή που περιγράψαμε, είναι αρκετά ευέλικτος ως προς τον σχεδιασμό του, που σημαίνει ότι ο εκάστοτε προγραμματιστής μπορεί να προσθέσει και άλλες διαδικασίες ή παραμέτρους που να επιλύουν αποτελεσματικά το πρόβλημα που καλείται να βελτιστοποιήσει. Η προσθήκη επιπλέον παραμέτρων στον γενετικό αλγόριθμο συνεπάγεται και την αναγκαιότητα για εύρεση της τιμής εκείνης που επιφέρει τα επιθυμητά αποτελέσματα στον αλγόριθμο, δηλαδή καλύτερες λύσεις και γρηγορότερη σύγκλιση. Η εύρεση των κατάλληλων παραμετροποιήσεων (tuning) αποδεικνύεται σε αυτήν την περίπτωση μια χρονοβόρα διαδικασία η οποία έχει καθοριστική σημασία για την επιτυχία του αλγορίθμου. Για το σκοπό αυτό, έχουν αναπτυχθεί αρκετά πακέτα και αλγόριθμοι που εξασφαλίζουν την εύρεση του βέλτιστου συνδυασμού παραμέτρων για το εκάστοτε πρόβλημα.
Για την σημασία των παραμέτρων στους γενετικούς αλγορίθμους, εστιάζουμε στα εξής τρία σημεία:

- Διαφορετικοί σχηματισμοί παραμέτρων ενδέχεται να επιφέρουν μεγάλη αλλαγή στην απόδοση του αλγορίθμου
- Ένας αποδεδειγμένα καλός σχεδιασμός παραμέτρων για ένα συγκεκριμένο πρόβλημα μπορεί να μην είναι κατάλληλος για κάποιο άλλο πρόβλημα
- Υπάρχουν αλληλοεξαρτήσεις μεταξύ των παραμέτρων του γενετικού αλγορίθμου, γεγονός που καθιστά ακατάλληλη τον προσδιορισμό των παραμέτρων χωριστά και ανεξάρτητα τον έναν από τον άλλον

Πρόσφατα έχει αποδειχθεί ότι η μέθοδος REVAC (Relevance Estimation and Value Calibration) δύναται να βρει αποδεδειγμένα καλές τιμές των παραμέτρων για τους εξελικτικούς αλγορίθμους. Η διαδικασία επιλογής τιμών για τις παραμέτρους αποτελεί στάδιο της επίλυσης του προβλήματος το οποίο μπορεί να αντιμετωπιστεί είτε πριν το "τρέξιμο" του εξελικτικού αλγορίθμου (parameter tuning) είτε κατά τη διάρκεια υλοποίησης του αλγορίθμου (parameter control). Ωστόσο, θα δοθεί έμφαση στον πρώτο τρόπο κατά τον οποίο ο προσδιορισμός των παραμέτρων γίνεται ως αρχικό στάδιο και οι τιμές αυτές παραμένουν αμετάβλητες κατά τη διάρκεια του εξελικτικού αλγορίθμου.
Η μέθοδος REVAC βασίζεται στη θεωρία πληροφορίας (information theory) για να μετρήσει την καταλληλότητα των παραμέτρων ενός εξελικτικού αλγορίθμου. Συγκεκριμένα, αντί να υπολογίζεται η απόδοση του εξελικτικού αλγορίθμου για διαφορετικούς συνδυασμούς των τιμών των παραμέτρων, η μέθοδος αυτή εκτιμά την αναμενόμενη απόδοσή του όταν οι τιμές των παραμέτρων διαλέγονται από μια κατανομή πυκνότητας πιθανότητας με μεγιστοποιημένη την εντροπία του Shannon. Η εντροπία του Shannon είναι ένα μέγεθος που μας επιτρέπει να ποσοτικοποιήσουμε την πληροφορία και συνδέεται με ορισμένες παραδοχές τις οποίες δε θα αναλύσουμε εδώ. Για έναν εξελικτικό αλγόριθμο, η μεγιστοποίηση της εντροπίας του Shannon αποτελεί ένα μέτρο της καταλληλότητας των παραμέτρων του, και με όρους από τη θεωρία της πληροφορίας, παρέχει ένα μέτρο που μας βοηθά να εκτιμήσουμε πόση πληροφορία χρειάζεται ώστε να φτάσουμε σε ένα συγκεκριμένο επίπεδο απόδοσης στον αλγόριθμο *(Vajapeyam, S. (2014). Understanding Shannon's Entropy metric for Information. University of Wisconsin-Madison).* Ταυτόχρονα, μας παρέχει μια ένδειξη σχετικά με το πώς αυτή η πληροφορία διανέμεται στις παραμέτρους του εξελικτικού αλγορίθμου. Υπό την έννοια αυτή, η μέθοδος REVAC επιδιώκει να μεγιστοποιήσει την εντροπία της κατανομής των παραμέτρων, για μια συγκεκριμένη απόδοση του αλγορίθμου *(Eiben et al. (1999). Parameter control in evolutionary algorithms. IEEE Transactions on Evolutionary Computation, 124–141*).
Μια ενδιαφέρουσα πρόταση λοιπόν προς επέκταση της διπλωματικής εργασίας και περαιτέρω αξιοποίηση του αλγορίθμου που κατασκευάστηκε θα ήταν η

εφαρμογή της μεθόδου REVAC για τον εντοπισμό του βέλτιστου διανύσματος παραμέτρων (best utility).

Αναφορές

Α) Ελληνικές

Καμπουρλάζος, Β. & Παπακώστας Γ. (2015). *Εισαγωγή στην Υπολογιστική Νοημοσύνη*, Κεφάλαιο 3. Αποθετήριο Κάλλιπος.

Καρώνη, Χ. & Οικονόμου, Π. (2010). *Στατιστικά Μοντέλα Παλινδρόμησης*. Εκδόσεις Συμεών. Αθήνα.

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

Β) Διεθνείς

Deepa, S.N. & Sivanandam, S.N. (2008). *Introduction to Genetic Algorithms*. Springer-Verlag Berlin Heidelberg.

Draper, D. & Fouskakis, D. (2002).*Stochastic Optimization: a Review.*. International Statistical Review 70, 3, 315–49.

Eberhart, R. C. & Kennedy, J.(2001). *Swarm Intelligence*, Morgan Kaufmann.

Eiben et al. (1999). *Parameter control in evolutionary algorithms*. IEEE Transactions on Evolutionary Computation, 124–141.

Hart, W.E., Krasnogor, N.& Smith, J.E. (2005). *Memetic Evolutionary Algorithms. Recent Advances in Memetic Algorithms. Studies in Fuzziness and Soft Computing*, vol 166. Springer-Verlag Berlin Heidelberg.

Heuvel, E., Romeijn, J.W. & Wit, E. (2012). *All models are wrong...': an introduction to model uncertainty*. Statistica Neerlandica, 66(3), 217-236.

Jeong, I.-S.et al. (2016). *A Feature Selection Approach Based on Simulated Annealing for Detecting Various Denial of Service Attacks.* River Publishers.

Rudolf, M. (2016). *Parameter tuning for numerical optimization algorithms.* Czech Technical University in Prague. Dept of Computer Science and Engineering. Prague.

Vajapeyam, S. (2014). *Understanding Shannon's Entropy metric for Information.* University of Wisconsin-Madison.

Βιβλιογραφία

[1] Bäck, T., Fogel, D. B. & Michalewicz, Z. (1997). *Handbook of Evolutionary Computation*. Oxford University Press, Oxford.

[2] Chambers, L. (2001). *The Practical Handbook of Genetic Algorithms.*, Chapman & Hall/CRC.

[3] Deepa, S.N &. Sivanandam, S.N. (2008). *Introduction to Genetic Algorithms*, Springer-Verlag Berlin Heidelberg.

[4] Fouskakis, D. (2001). *Stochastic Optimisation Methods for Cost-Effective Quality Assessment in Health*. University of Bath.

[5] Goldberg, D., Kendall, G. & Sastry, K. (2020). *Genetic Algorithms.*, Springer-Verlag Berlin Heidelberg

[6] Hayes-Roth, F. (1975). *Review of 'Adaptation in Natural and Artificial Systems by John H. Holland'*. The University of Michigan Press.