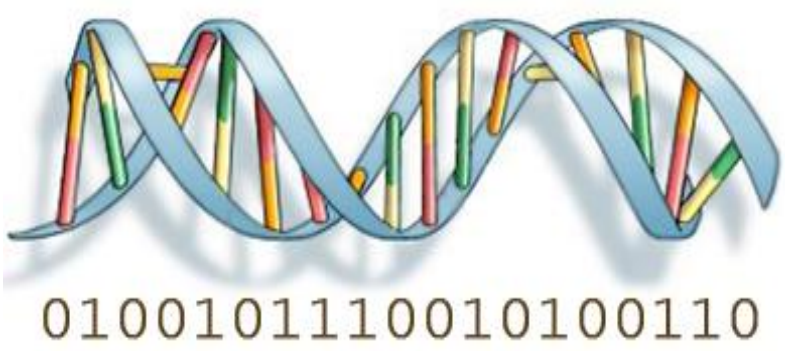
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΙΟ

Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών

Τομέας Μαθηματικών

ΓΕΝΕΤΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ ΣΤΟ ΠΡΟΒΛΗΜΑ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ



ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΠΑΛΤΖΟΓΛΟΥ Δ. ΕΛΕΝΗ

Επιβλέπων Καθηγητής: Φουσκάκης Δημήτριος

Αθήνα, Οκτώβριος 20

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

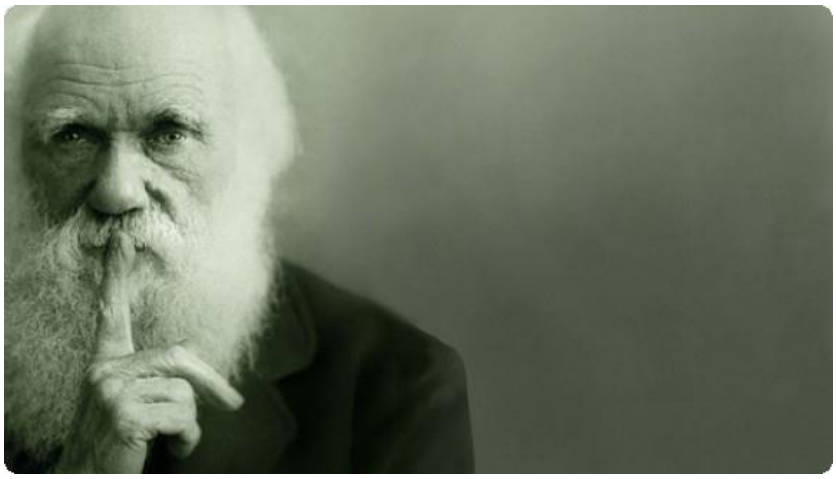
Σε πολλά προβλήματα βελτιστοποίησης παρουσιάζονται προβλήματα επίλυσης τόσο σημαντικά που είναι αδύνατο να επιλυθούν με τις κλασσικές μεθόδους. Υπάρχει λοιπόν, η αναγκαιότητα για μεθόδους επίλυσης οι οποίες να ξεπερνούν τα προβλήματα που εμφανίζονται και τέτοιες είναι οι ευριστικές μέθοδοι. Αυτές κάνουν επαναληπτικά μία περιορισμένη αναζήτηση στον χώρο λύσεων, για να βελτιώνουν μία υπάρχουσα λύση. Με τον τρόπο αυτό βελτιώνουν την λύση αυτή και μέσω των επαναλήψεων φτάνουν σε μία λύση που πιθανότατα είναι "πάρα πολύ καλή". Οι μέθοδοι αυτές υλοποιούνται σχετικά εύκολα αλγοριθμικά. Ένας τέτοιος αλγόριθμος είναι και ο γενετικός αλγόριθμος.

Ένας γενετικός αλγόριθμος μιμείται την αναπαραγωγή των πληθυσμών και την αλλαγή μέσω της διαδοχής των γενεών των ιδιοτήτων και δυνατοτήτων των ατόμων του αρχικού πληθυσμού. Για τους γενετικούς αλγορίθμους αυτό μεταφράζεται σε βελτίωση των λύσεων οι οποίες και αντιστοιχούν στα άτομα ενός πληθυσμού. Το πρόβλημα στο οποίο θα εφαρμοστεί ένας γενετικός αλγόριθμος για την επίλυσή του είναι το πρόβλημα της επιλογής μεταβλητών στο πολλαπλό γραμμικό μοντέλο γραμμικής παλινδρόμησης.

Η παρούσα εργασία αποτελείται από 6 κεφάλαια. Στο πρώτο δίνεται μια γενική ιδέα των γενετικών αλγορίθμων, στο δεύτερο κεφάλαιο αναλύονται τα μέρη ενός γενετικού αλγορίθμου και τα πλεονεκτήματα και μειονεκτήματα αυτών. Στο τρίτο κεφάλαιο παρουσιάζεται το πρόβλημα επιλογής μεταβλητών στο πολλαπλό γραμμικό μοντέλο, ενώ στο τέταρτο κεφάλαιο αναλύεται η μοντελοποίηση του γενετικού αλγορίθμου και η προσαρμογή του στο συγκεκριμένο πρόβλημα. Ακολούθως, στο πέμπτο κεφάλαιο ελέγχεται ο αλγόριθμος με προσομοιωμένα δεδομένα και στο έκτο κεφάλαιο επιλύεται μέσω του αλγορίθμου ένα πρόβλημα επιλογής μεταβλητών με πραγματικά δεδομένα στο οποίο εμπλέκονται 56 μεταβλητές. Τέλος, στο παράρτημα, υπάρχει ο κώδικας του αλγορίθμου στην **R.**

ΚΕΦΑΛΑΙΟ 1
ΕΙΣΑΓΩΓΗ ΣΤΟΝ ΓΕΝΕΤΙΚΟ ΑΛΓΟΡΙΘΜΟ

1.1 ΕΙΣΑΓΩΓΗ

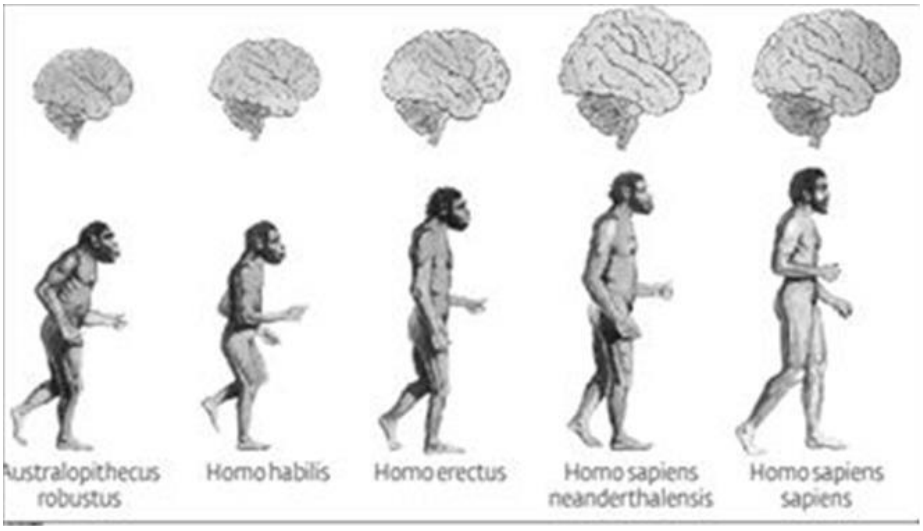


Ο γενετικός αλγόριθμος (ΓΑ) είναι μια τεχνική βελτιστοποίησης που βασίζεται στη θεωρία της εξέλιξης των ειδών και της φυσικής επιλογής, όπως αυτή διατυπώθηκε από το Δαρβίνο στα μέσα του 19ου αιώνα. Η θεωρία της *Εξέλιξης των Ειδών (Evolution of Species)*, προκάλεσε μεγάλη αναστάτωση, αφού ερχόταν σε σύγκρουση με τις επικρατούσες θρησκευτικές αντιλήψεις περί προέλευσης της ζωής. Με την πάροδο ενός και πλέον αιώνα, η αναστάτωση αυτή δεν έχει σταματήσει εντελώς, όμως η θεωρία έχει γίνει αποδεκτή από το σύνολο των επιστημόνων, γιατί κατάφερε να πείσει και να δώσει ικανοποιητικές απαντήσεις σε βασικά ερωτήματα.
Σκοπός της θεωρίας αυτής είναι να δώσει μια εξήγηση για το φαινόμενο της ζωής, τις ιδιότητες που αναπτύσσει κάθε είδος και το βοηθά να επιβιώσει ή αυτές που δημιουργεί ανάλογα με το τι απαιτεί η προσαρμογή του στο περιβάλλον ώστε να εξασφαλιστεί η συνέχεια, η «διαιώνιση». Τα κυριότερα σημεία της, που σχετίζονται και ερμηνεύουν τον τρόπο λειτουργίας των Γενετικών Αλγορίθμων, είναι τα εξής:
• Οι ζωντανοί οργανισμοί δε διαχωρίζονται σε ανώτερους και κατώτερους (εννοείται στο ίδιο βιολογικό είδος, λ.χ. των ανθρώπων). Σε κάθε βιολογικό είδος, μερικά άτομα αφήνουν περισσότερους απογόνους σε σύγκριση με τα υπόλοιπα και έτσι τα κληροδοτούμενα χαρακτηριστικά των αναπαραγωγικά επιτυχημένων ατόμων γίνονται περισσότερα στην επόμενη γενιά. Αναλόγως με τις δυσκολίες και τις αντιξοότητες που δημιουργούνται κατά τη διάρκεια της ζωής των οργανισμών, κάποιοι απ' αυτούς επιβιώνουν και πολλαπλασιάζονται. Έτσι λοιπόν, αλλάζουν και τα χαρακτηριστικά των οργανισμών προκειμένου να προσαρμοστούν κάθε φορά στις νέες συνθήκες και να επιβιώσουν.
• Οι αλλαγές, που συμβαίνουν στα χαρακτηριστικά των ατόμων είναι αλλαγή στα *χρωμοσώματά* τους (*chromosomes*), που είναι πολύπλοκα οργανικά μόρια τα οποία κωδικοποιούν τη δομή και τα χαρακτηριστικά τους. Τα χρωμοσώματα είναι ένα

μοναδικό κομμάτι που περιλαμβάνει πολλά γονίδια *(genes)* και άλλες ακολουθίες νουκλεοτιδίων. Το σύνολο της γενετικής πληροφορίας που είναι κωδικοποιημένο στα γονίδια ονομάζεται *γονότυπος (genotype)*. Η δημιουργία ενός νέου οργανισμού περιλαμβάνει την αποκωδικοποίηση των χρωμοσωμάτων. Όλα τα μορφολογικά, παραγωγικά, ηθολογικά κλπ χαρακτηριστικά που εκδηλώνει ένας οργανισμός σε μία δεδομένη στιγμή, δηλαδή το μέρος του γονοτύπου του οργανισμού το οποίο μπορούμε να παρατηρήσουμε, ονομάζεται *φαινότυπος* (*phenotype*).

• Κυρίαρχες λειτουργίες του φαινομένου της εξέλιξης είναι η *αναπαραγωγή (reproduction)* και η *μετάλλαξη (mutation).* Οι μεταλλάξεις συμβαίνουν με τυχαίο τρόπο, χωρίς αυτό να σημαίνει ότι δεν υπόκεινται στην επίδραση του περιβάλλοντος. Η μετάλλαξη είναι μία αλλαγή σε ένα γονίδιο. Κάποιες φορές κατά τον διπλασιασμό του DNA είναι δυνατόν να γίνουν λάθη, έτσι το DNA στο νέο κύτταρο δε θα είναι ακριβώς το ίδιο, ως όφειλε, αυτή η αλλαγή λέγεται μετάλλαξη. Η μετάλλαξη, μερικές φορές, μπορεί να προκαλέσει βελτιώσεις και, χωρίς αμφιβολία, μερικά λάθη που έγιναν αποτέλεσαν σημαντικό παράγοντα για την προοδευτική εξέλιξη της ζωής.

• Προϊόν της αναπαραγωγής είναι ένας νέος οργανισμός, τα χρωμοσώματα του οποίου αποτελούνται από γονίδια που προέρχονται τα μισά από τον πατέρα και τα μισά από τη μητέρα. Έτσι, για κάθε χαρακτηριστικό, το νέο άτομο έχει πάρει ένα γονίδιο από κάθε γονέα. Τα αλληλόμορφα γονίδια είναι γονίδια που δρουν για το ίδιο γνώρισμα αλλά με διαφορετικό τρόπο. Συνήθως από τα δύο αλληλόμορφα, το ένα επικρατεί του άλλου και καθορίζει το φαινότυπό του, αυτό το αλληλόμορφο ονομάζεται επικρατές (*dominant*) και το άλλο υπολειπόμενο (*recessive*). Όλος αυτός ο μηχανισμός της φυσικής επιλογής φάνηκε ιδιαίτερα ελκυστικός στον John Holland, πρωτοπόρο των Γενετικών Αλγορίθμων, στις αρχές της δεκαετίας του '70. Ο Holland φαντάστηκε ότι κάποιες ιδέες και λειτουργίες που εφαρμόζει η φύση στα συστήματά της θα μπορούσαν να έχουν αποτελέσματα, αν ενσωματώνονταν σε αλγόριθμους για υπολογιστές, ώστε να προκύψουν αποδοτικές τεχνικές επίλυσης δύσκολων προβλημάτων. Αποτέλεσμα αυτής της εργασίας του Holland ήταν οι Γενετικοί Αλγόριθμοι, μια καινούργια εξελισσόμενη και πολλά υποσχόμενη τεχνική αναζήτησης και βελτιστοποίησης, όπου η βασική τους ιδέα είναι η μίμηση των μηχανισμών της φύσης.



1.2 ΟΙ ΚΥΡΙΟΤΕΡΟΙ ΣΤΑΘΜΟΙ ΣΤΗΝ ΕΞΕΛΙΞΗ ΤΩΝ ΓΕΝΕΤΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ ΕΙΝΑΙ ΟΙ ΠΑΡΑΚΑΤΩ:

- Ο Bagley (1967) με τη διδακτορική του διατριβή ουσιαστικά «βαφτίζει» τους γενετικούς αλγορίθμους.
- Ο Rosenberg (1967) δημοσιεύει εργασία, στην οποία γίνεται λόγος για προσομοίωση πληθυσμών μονοκύτταρων οργανισμών σε υπολογιστικό περιβάλλον.
- Ο Holland (1975) εκδίδει το βιβλίο «Προσαρμογή στα Φυσικά και Τεχνητά συστήματα», στο οποίο αναπτύσσει τις ιδέες και τη θεωρία των ΓΑ. Το βιβλίο θεωρείται πλέον κλασικό για το χώρο. Θίγονται θέματα όπως η θεωρία των σχημάτων, η βέλτιστη κατανομή των ευκαιριών, σχέδια αναπαραγωγής, γενετικές λειτουργίες, η ευρωστία των ΓΑ και πλήθος άλλα.
- Ο De Jong (1975) με την εργασία που εκδίδει βοηθά την πειραματική αξιολόγηση των ΓΑ. Σύμφωνα με αυτήν, προτείνονται λειτουργίες που ελέγχουν ένα ΓΑ και την ικανότητά του να αντιμετωπίζει δύσκολα προβλήματα.
- Ο Grefenstette (1980) δημιουργεί το GENESIS, ένα σύστημα ανάπτυξης ΓΑ υλοποιημένο στη γλώσσα προγραμματισμού C, που έχει βοηθήσει σημαντικά στη διάδοση του γενετικού προγραμματισμού καθώς έγινε διαθέσιμο στο ευρύ κοινό.
- Γίνεται το 1ο Διεθνές Συνέδριο των Γ.Α. και των εφαρμογών τους, 1985. Ο χώρος αποκτά ένα μεγάλο συνέδριο που πλέον λαμβάνει χώρα κάθε δύο χρόνια και αντικατοπτρίζει το μεγάλο οργασμό που παρατηρείται σε επίπεδο τόσο θεωρίας, όσο και εφαρμογών.
- Πολυάριθμες εκδόσεις βιβλίων για Γ.Α., 1989-1999. Άλλη μια ένδειξη της τεράστιας ανάπτυξης του χώρου και της αποδοχής της νέας τεχνολογίας.

- Ανάπτυξη πακέτων λογισμικού για Γ.Α., 1990-1999. Πολλές εταιρίες δημιουργούν εμπορικά πακέτα που επιτρέπουν σε χρήστες να ενσωματώσουν στις εφαρμογές τους στοιχεία Γενετικού Προγραμματισμού (Genetic Programming). Ένα τέτοιο πακέτο είναι το EOS (Evolutionary Object System). Βασίζεται στη δημοφιλή γλώσσα αντικειμενοστραφούς προγραμματισμού C++ και παρέχει μεγάλες δυνατότητες προσαρμογών και επεκτάσεων.

Οι γενετικοί αλγόριθμοι ανήκουν στον κλάδο της επιστήμης υπολογιστών και αποτελούν μια μέθοδο αναζήτησης βέλτιστων λύσεων σε συστήματα που μπορούν να περιγραφούν με κάποιο μαθηματικό μοντέλο. Χρησιμοποιούνται ιδιαιτέρως σε προβλήματα, που έχουν πολλές παραμέτρους/μεταβλητές απόφασης και δεν έχει βρεθεί κάποια αναλυτική μέθοδος, που να μπορεί να βρει το βέλτιστο συνδυασμό τιμών των μεταβλητών αυτών.

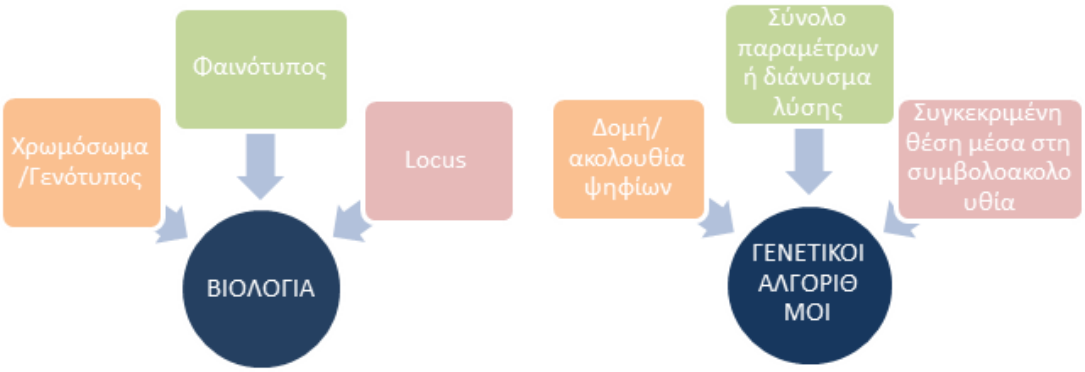
Ο ΓΑ, ουσιαστικά, επιτρέπει σε έναν πληθυσμό αποτελούμενο από πολλά άτομα να εξελιχθεί κάτω από συγκεκριμένους κανόνες, προκειμένου να ελαχιστοποιηθεί (μεγιστοποιηθεί) μια αντικειμενική συνάρτηση. Αυτή η μέθοδος που αναπτύχθηκε το 1975 από τον John Holland όπως αναφέρθηκε, δημοσιεύτηκε από το David Goldberg το 1989 με εφαρμογή στον έλεγχο της μεταφοράς πετρελαίου μέσω αγωγών.

Οι δύο πιο διαδεδομένες μορφές ΓΑ είναι ο δυαδικός ΓΑ (binary GA) και ο συνεχής ΓΑ (continuous GA). Η κύρια διαφορά τους είναι πως στο δυαδικό αλγόριθμο το χρωμόσωμα πρέπει να αποκωδικοποιηθεί από τη δυαδική του μορφή (γονίδιο) στο φαινότυπο του (αποκωδικοποιημένες τιμές), ενώ στο συνεχή ΓΑ το χρωμόσωμα δεν είναι κωδικοποιημένο.

1.3 Η ΟΡΟΛΟΓΙΑ ΤΩΝ ΓΕΝΕΤΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ

Το μεγαλύτερο μέρος της ορολογίας των γενετικών αλγορίθμων είναι δανεισμένο από την ορολογία της βιολογίας. Το γονίδιο αποτελεί τη βασική δομική μονάδα στην γενετική αλλά και στην μέθοδο βελτιστοποίησης με τη χρήση γενετικών αλγόριθμων. Τα γονίδια αποτελούν την κωδικοποιημένη παράσταση των παραμέτρων βελτιστοποίησης ενώ τα χρωμοσώματα είναι πολλά γονίδια μαζί. Πιο συγκεκριμένα, κάθε υποψήφια λύση παριστάνεται με μια *συμβολοσειρά* ενός πεπερασμένου αλφαβήτου. Κάθε στοιχείο της συμβολοσειράς είναι ένα γονίδιο και η ίδια η συμβολοσειρά είναι ένα χρωμόσωμα.

Στην φύση (DNA) το αλφάβητο έχει μήκος τέσσερα και αποτελείται από τα στοιχεία A (αδενίνη), G (γουανίνη), T (θυμίνη) και C (κυτοσίνη) ενώ στους γενετικούς αλγορίθμους χρησιμοποιείται συνήθως δυαδικό αλφάβητο ή μερικές φορές πιο σύνθετες μορφές αναπαράστασης όπως πραγματικοί αριθμοί.



Ο ΓΑ μπορεί να ξεκινάει, όπως κάθε άλλος αλγόριθμος βελτιστοποίησης, με τον καθορισμό των μεταβλητών απόφασης (decision variables), την αντικειμενική συνάρτηση (objective function) ή συνάρτηση καταλληλότητας (fitness function) και τερματίζεται ελέγχοντας αν συνέκλινε σε μία καλή ή τη βέλτιστη λύση, όμως διαφέρει σημαντικά στα ενδιάμεσα στάδια σε σχέση με άλλους αλγορίθμους βελτιστοποίησης, διότι χρησιμοποιεί γενετικούς τελεστές και τεχνικές για την αναζήτηση της καλύτερης λύσης.

Ένας ΓΑ πραγματοποιεί αναζήτηση σε πολλές κατευθύνσεις με το να διατηρεί έναν πληθυσμό από πιθανές λύσεις και να υποστηρίζει καταγραφή και ανταλλαγή πληροφοριών μεταξύ αυτών των κατευθύνσεων. Ο πληθυσμός υφίσταται μια προσομοιωμένη γενετική εξέλιξη. Σε κάθε γενιά, οι σχετικά «καλές» λύσεις αναπαράγονται, ενώ οι σχετικά «κακές» αφαιρούνται. Ο διαχωρισμός και η αξιολόγηση των διαφόρων λύσεων γίνεται με την βοήθεια μιας *αντικειμενικής συνάρτησης* ή *συνάρτησης ικανότητας (objective ή fitness function)*, η οποία παίζει το ρόλο του περιβάλλοντος μέσα στο οποίο εξελίσσεται ο πληθυσμός. Επίσης στη βιβλιογραφία αναφέρεται και ως συνάρτηση αξιολόγησης και συνάρτηση καταλληλότητας.

Στην αρχή του ΓΑ αρχικοποιείται ο πληθυσμός των N_{pop} χρωμοσωμάτων,που θα πρέπει να είναι άρτιος αριθμός. Ο πληθυσμός αναπαριστάται με έναν πίνακα, όπου κάθε γραμμή του είναι ένα διάνυσμα $1 \times N_{var}$ (χρωμόσωμα). Ο πίνακας $N_{pop} \times N_{var}$ (αρχικός πληθυσμός) δημιουργείται συνήθως από τυχαίες τιμές των μεταβλητών του χρωμοσώματος. Όμως, πρακτικά παρατηρείται ότι ο πληθυσμός επιλεγμένος με τέτοιο τρόπο δεν καλύπτει ομοιόμορφα όλο το χώρο αναζήτησης, οπότε υπάρχει το ενδεχόμενο να μειωθεί η απόδοση του αλγορίθμου. Για το λόγο αυτό ίσως πρέπει να υιοθετηθούν πιο μελετημένες στατιστικά μέθοδοι για την επιλογή του. Επίσης το μέγεθος του αρχικού πληθυσμού πρέπει να επιλεγεί με τέτοιο τρόπο ώστε να συμβιβάσει την αποδοτικότητα και την αποτελεσματικότητα του αλγορίθμου.

Στους γενετικούς αλγόριθμους η συνάρτηση καταλληλότητας αποδίδει μια τιμή σε κάθε άτομο του πληθυσμού. Αυτή η τιμή είναι ο δείκτης της ποιότητας της λύσης που αντιπροσωπεύει κάθε χρωμόσωμα. Η αξιολόγηση της συνάρτησης καταλληλότητας για κάθε άτομο θα πρέπει να είναι σχετικά γρήγορη, λόγω του αριθμού των φορών που θα γίνεται η κλήση της. Στην αντίθετη περίπτωση που η διαδικασία δηλαδή είναι αργή, χρησιμοποιούνται τεχνικές όπως η παράλληλη επεξεργασία, κάποια προσεγγιστική μέθοδος αποτίμησης ή κάποια μέθοδος που κάνει υπολογισμούς μόνο στα στοιχεία που τροποποιούνται κάθε φορά.

Μόλις ένας πληθυσμός δημιουργηθεί και εκτιμηθεί η καταλληλότητα του, επιλέγονται οι λύσεις οι οποίες πρόκειται να συνδυαστούν για τη δημιουργία της επόμενης γενιάς. Η διαδικασία της επιλογής αυτής θα πρέπει να επιτρέπει στις καλύτερες λύσεις να αναπαράγονται, επιτρέποντας ταυτόχρονα και στα άτομα με χαμηλότερη καταλληλότητα να συμμετέχουν στη διαδικασία αυτή, αλλά με χαμηλότερη πιθανότητα. Επίσης, κατά τη διαδικασία της επιλογής θα πρέπει να λαμβάνονται υπόψη οι διαφορετικές καταλληλότητες κάθε χρωμοσώματος.

ΚΕΦΑΛΑΙΟ 2
Ο ΓΕΝΕΤΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ ΣΕ ΠΡΟΒΛΗΜΑΤΑ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ

2.1 ΓΕΝΕΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ

Σε πολλά προβλήματα βελτιστοποίησης παρουσιάζονται σημαντικά προβλήματα επίλυσης. Αυτά μπορεί να οφείλονται στην δυσκολία υλοποίησης των αντίστοιχων μεθόδων επίλυσης ή ακόμα και στην αποδοτικότητα της λειτουργίας αυτών των μεθόδων, είτε λόγω της μοντελοποίησης των προβλημάτων είτε λόγω της αδυναμίας λειτουργίας των μεθόδων αυτών όταν η διάσταση του προβλήματος μεγαλώνει.

Σε αρκετά προβλήματα βελτιστοποίησης ο παραμετρικός χώρος είναι διακριτός και σκοπός είναι συνδυάζοντας τις τιμές του συγκεκριμένου χώρου να βρούμε τη μέγιστη ή την ελάχιστη τιμή μιας συνάρτησης. Τέτοια προβλήματα βρίσκουν πολλές εφαρμογές στη στατιστική όπως θα δούμε παρακάτω, για παράδειγμα στο πρόβλημα επιλογής επεξηγηματικών μεταβλητών στην παλινδρόμηση. Τα εν λόγω προβλήματα ονομάζονται συνδυαστικής βελτιστοποίησης (*combinatorial optimization*), όπου πολλές παράμετροι συνδυάζονται και αλληλεπιδρούν ενδεχομένως με διαφόρους τρόπους, επηρεάζοντας με την σειρά τους την τιμή κάποιας άλλης (εξαρτημένης) μεταβλητής μέσω μιας συνάρτησης *f*. Αν ζητείται η μεγιστοποίηση της τιμής της συνάρτησης *f* πάνω στο σύνολο των δυνατών τιμών των παραμέτρων $\theta_1,\theta_2,\ldots,\theta_v$ έχουμε προβλήματα πάρα πολύ μεγάλης πολυπλοκότητας. Για τέτοια προβλήματα μπορεί η επίλυσή τους να μην απαιτεί απλώς πολυωνυμικό χρόνο επίλυσης, ως προς το πλήθος π.χ. των παραμέτρων, αλλά εκθετικό χρόνο επίλυσης, κάτι που κάνει αδύνατη την εύρεση της βέλτιστης λύσης με κλασσικές μεθόδους επίλυσης, καθώς αυξάνεται η διάσταση του αντίστοιχου προβλήματος. Ένα τέτοιο γνωστό πρόβλημα είναι το πρόβλημα του περιοδεύοντος πωλητή, ο οποίος αναζητά την βέλτιστη διαδρομή/διέλευση από *p* πόλεις δεδομένων κερδών από τις διελεύσεις αλλά και κόστους μετάβασης από πόλη σε πόλη. Το πλήθος των διαφορετικών διελεύσεων από τις *p* πόλεις είναι *(p -1)!/2*.

Η αναγκαιότητα χρήσης άλλων μεθόδων και τεχνικών εύρεσης της βέλτιστης λύσης σε τέτοια προβλήματα, πέρα των κλασσικών μεθόδων, είναι δεδομένη. Η αρχή λειτουργίας τέτοιων μεθόδων είναι ότι, είναι επαναληπτικές μέθοδοι και αναζητούν λύσεις πάντα σε τοπικό επίπεδο (με αυθαίρετη συνήθως εκκίνηση), αλλά προσπαθώντας πάντα να εξερευνούν νέες περιοχές του χώρου λύσεων ενώ συγχρόνως εκμεταλλεύονται τις υπάρχουσες "καλές" λύσεις. Τέτοιες μέθοδοι είναι οι γενετικοί αλγόριθμοι, μέθοδοι *Monte Carlo, simulated annealing, tabu algorithms* κ.α.

Υπάρχουν προβλήματα βελτιστοποίησης όπου οι κλασσικές μέθοδοι εύρεσης λύσης, δεν μπορούν στην πράξη να επιτύχουν την επίλυσή τους. Σε προβλήματα πολύ μεγάλης πολυπλοκότητας και μεγάλης διάστασης του χώρου των δυνατών λύσεων, είναι πρακτικά αδύνατη η εύρεση του ολικού μέγιστου μιας συνάρτησης σε πραγματικά χρονικά όρια. Έτσι είναι αναγκαίο να χρησιμοποιηθούν οι λεγόμενοι ***ευριστικοί αλγόριθμοι*** οι οποίοι αναζητούν την βέλτιστη λύση τοπικά. Πάντα οι ευριστικοί αλγόριθμοι δουλεύουν επαναληπτικά, με την βελτίωση της λύσης σε κάθε

επανάληψη αλλά και την περιορισμένη αναζήτηση σε κάθε επανάληψη σε ένα υποσύνολο του χώρου των λύσεων κάθε φορά. Ένας Γενετικός Αλγόριθμος είναι ένας ευριστικός αλγόριθμος, όπου κάθε επανάληψή του είναι μία γενιά ενός πληθυσμού (υποσυνόλου δυνατών λύσεων), και η οποία γενιά ανανεώνεται σε κάθε επανάληψη. Η πρώτη γενιά μπορεί να είναι αυθαίρετη με την έννοια πως απλά μπορεί να είναι ένα υποσύνολο των εφικτών λύσεων του προβλήματος. Οι επαναλήψεις μπορεί να τερματιστούν μετά από συγκεκριμένο αριθμό επαναλήψεων ή μετά από την σύγκλιση του αλγορίθμου σε έναν "σημείο" του χώρου των λύσεων.

Στην ουσία ο αλγόριθμος σε κάθε βήμα που κάνει, έχοντας μία λύση του προβλήματος (όχι απαραίτητα την βέλτιστη), κάνει αναζήτηση γύρω από την λύση αυτή (τοπικά). Γύρω από την λύση εννοούμε σε μία μικρή γειτονιά της λύσης αυτής και βέβαια στον χώρο των εφικτών λύσεων. Αυτή η τοπική αναζήτηση (*local search*) επιτρέπει στον αλγόριθμο να βρει μία λύση καλύτερη της υπάρχουσας ειδικά όταν η υπάρχουσα λύση είναι κοντά σε ένα τοπικό μέγιστο. Όταν η λύση είναι κοντά σε ένα τοπικό μέγιστο το ψάξιμο αντιστοιχεί σε ανάβαση λόφου όπου στην κορυφή είναι το τοπικό μέγιστο (*hill climbing*). Η τοπική αναζήτηση είναι αναγκαία ώστε να μην φύγουμε μακριά από το τοπικό μέγιστο κάνοντας αναζήτηση σε τυχαία σημεία (μακριά από αυτό). Με αυτήν την τεχνική επιτυγχάνουμε τον στόχο που είναι να βελτιώνουμε την υπάρχουσα λύση. Εφαρμόζοντας αποκλειστικά ανάβαση λόφου (*hill climbing*) υπάρχει μεγάλη πιθανότητα η λύση που θα βρούμε να είναι τοπικού μεγίστου ή ελαχίστου ειδικά όταν η συνάρτηση f είναι πολυκόρυφη. Γι'αυτό το λόγο προσπαθούμε να εφαρμόσουμε έναν πιο προχωρημένο αλγόριθμο με σκοπό να αποφεύγουμε τα τοπικά ακρότατα.

Ο αρχικός πληθυσμός (η πρώτη γενιά του αλγόριθμου), όπως αναφέραμε είναι ένα υποσύνολο του χώρου των εφικτών λύσεων. Ο τρόπος που παράγονται αυτές οι λύσεις εξ αρχής μπορεί να είναι μία σύνθετη διαδικασία (αναλόγως και της φύσης του προβλήματος) ή και μία πολύ απλή διαδικασία δειγματοληψίας με επανάθεση. Για παράδειγμα, αν ο πληθυσμός χρωμοσωμάτων που απαιτεί ο ΓΑ είναι δυαδικά διανύσματα (όπως στο πρόβλημα επιλογής μεταβλητών στο πολλαπλό γραμμικό μοντέλο), ο αρχικός πληθυσμός παράγεται με δειγματοληψία μηδενικών και άσσων με επανάθεση και το δείγμα να έχει μέγεθος όσο απαιτείται για το μήκος του διανύσματος. Το μήκος του διανύσματος όπως και τι ακριβώς δηλώνει η παρουσία του 0 και του 1 σε ένα τέτοιο χρωμόσωμα, ορίζεται από την μοντελοποίηση του προβλήματος ως προς τον ΓΑ. Κάθε διάνυσμα είναι και ένα χρωμόσωμα δηλαδή ένα άτομο του αρχικού πληθυσμού. Γενικότερα, η συνήθης κωδικοποίηση που ακολουθείται στους ΓΑ για τα χρωμοσώματα είναι η δυαδική κωδικοποίηση, δηλαδή το χρωμόσωμα να είναι ένα διάνυσμα αποτελούμενο από μηδενικά και άσσους. Με τον τρόπο αυτό είναι εύκολη η υλοποίηση του αλγορίθμου καθώς και οι γενετικές διαδικασίες όπως θα δούμε παρακάτω. Ακόμα, η υλοποίηση της κωδικοποίησης με δυαδικά διανύσματα είναι πολύ εύκολη και γρήγορη και φυσικά μπορεί πάντα να βρεθεί εύκολα αντιστοίχιση ενός πεπερασμένου συνόλου πάνω σε δυαδικά διανύσματα κατάλληλου μήκους.

Υπάρχουν και προβλήματα όπου η κωδικοποίηση απαιτείται να είναι φυσικοί αριθμοί όπως στο πρόβλημα του **περιοδεύοντος πωλητή**. Στο πρόβλημα αυτό ένας πωλητής πρέπει να περάσει από ένα πεπερασμένο πλήθος πόλεων (σημείων πώλησης) στην διάρκεια της περιοδείας του, και φυσικά μόνο μία φορά από κάθε πόλη. Το κέρδος του ορίζεται από την προκαθορισμένη πώληση σε κάθε πόλη όπως και από το κόστος μετάβασης του συνόλου των διαδρομών που θα κάνει (π.χ. από τα συνολικά χλμ). Ζητούμενο είναι να ελαχιστοποιηθεί το κόστος μετάβασης των διαδρομών του πωλητή. Η μοντελοποίηση του προβλήματος είναι ο ορισμός των πόλεων με φυσικούς αριθμούς, οπότε αν έχει να περάσει ο πωλητής από p=9 πόλεις αυτές θα αριθμούνται από το 1 έως το 9. Η διαδρομή που θα ακολουθήσει ο πωλητής είναι μία αλληλουχία των αριθμών που αντιστοιχούν σε κάθε πόλη και βέβαια με την σειρά με την οποία θα περάσει από αυτές ο πωλητής. Για παράδειγμα μία "διαδρομή" του πωλητή μπορεί να είναι η 268745193, η οποία δηλώνει πως ο πωλητής ξεκινά από την πόλη 2 συνεχίζει στην πόλη 6, μετά με την σειρά στις 8,7,4,5,1,9 και τερματίζει στην πόλη 3. Κάθε αριθμός (πόλη) θα πρέπει να εμφανίζεται μόνο μία φορά σε κάθε "διαδρομή". Ένα τέτοιο πρόβλημα με την συγκεκριμένη μοντελοποίηση, αν και συνήθη, μπορεί να εμφανίζει διάφορα προβλήματα σε έναν ΓΑ. Τόσο στην διαδικασία της διασταύρωσης όσο και στην διαδικασία της μετάλλαξης, δεδομένου πως κάθε πόλη-αριθμός θα πρέπει να εμφανίζεται μόνο μία φορά σε κάθε "διαδρομή". Επίσης, μπορεί να υπάρχουν και φυσικοί περιορισμοί μετάβασης από πόλη σε πόλη και έτσι π.χ. να μην μπορεί ο αριθμός 5 να διαδεχθεί τον αριθμό 7. Θα αναφερθούμε στις παραγράφους 2.3 για τον τρόπο επίλυσης τέτοιων προβλημάτων.

Η αρχικοποίηση είναι το βήμα στο οποίο ορίζεται ο αρχικός πληθυσμός, πάνω στον οποίο θα εκτελεστούν οι λειτουργίες του ΓΑ. Ο πληθυσμός αυτός διαλέγεται με τυχαίο τρόπο ανάμεσα σε όλες τις δυνατές τιμές των μεταβλητών του προβλήματος, ενώ το μέγεθός του ορίζεται από το χρήστη (συνήθως, όμως, εξαρτάται από τους πόρους που αυτός έχει στη διάθεσή του). Σε μερικές υλοποιήσεις, η επιλογή των αρχικών σημείων γίνεται με ευρετικές μεθόδους, δίνοντας εξαρχής ένα πλεονέκτημα στην αναζήτηση.

Για να δούμε πόσο αποδοτικό είναι ένα χρωμόσωμα, υπολογίζουμε μία συνάρτηση η οποία ονομάζεται συνάρτηση καταλληλότητας (fitness) και χρησιμοποιείται για να ελαχιστοποιήσει ή μεγιστοποιήσει την πιθανότητα επιλογής κάθε χρωμοσώματος. Σε τέτοιου είδους προβλήματα μπορεί να χρησιμοποιηθεί η συνάρτηση f(x) αυτή καθεαυτή ή ένας μετασχηματισμός αυτής. Συγκεκριμένα για προβλήματα ολικής ελαχιστοποίησης η συνάρτηση καταλληλότητας μπορεί να υπολογιστεί από τον τύπο

$$(f_{max}-f(x))/(f_{max}-f_{min}),\ \text{αν}\ f_{max}\neq f_{min}$$

όπου μπορεί να πάρει τιμές στο διάστημα [0,1] και θεωρούμε f_{max}: τη μέγιστη τιμή της συνάρτησης των χρωμοσωμάτων, f_{min}: την ελάχιστη τιμή της συνάρτησης των χρωμοσωμάτων και f(x): την τιμή της συνάρτησης για το x χρωμόσωμα.

Άλλη συνάρτηση καταλληλότητας που μπορεί να χρησιμοποιηθεί είναι η παρακάτω

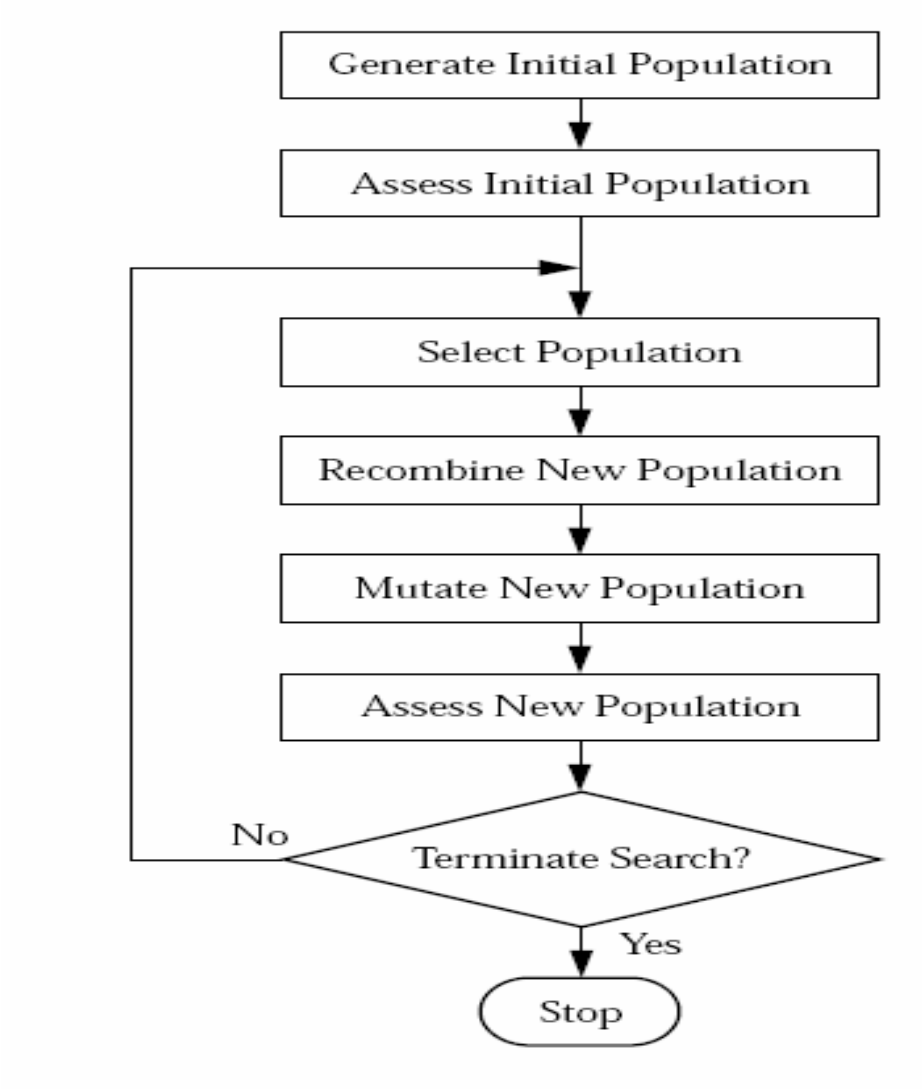
$$f_i/f_{average}$$

όπου f_i: είναι η τιμή του i χρωμοσώματος και $f_{average}$: η τιμή που εκφράζει το μέσο όρο των τιμών των χρωμοσωμάτων.

Αφού προκύψει η πρώτη γενιά, ο ΓΑ εισέρχεται στο επαναληπτικό μέρος του όπως φαίνεται και στο ***σχήμα 2.1***. Ο πληθυσμός πρέπει να αξιολογηθεί, δηλαδή να εκτιμηθεί η δυνατότητα επιβίωσης του κάθε ατόμου χωριστά. Για να συμβεί αυτό πρέπει να γίνει αποκωδικοποίηση χαρακτηριστικών και έπειτα υπολογισμός της απόδοσης των ατόμων. Ο παραλληλισμός με το φυσικό μοντέλο, ίσως βοηθά στην κατανόηση αυτής της διαδικασίας: Στη φύση τα χρωμοσώματα ενός οργανισμού έχουν στα γονίδιά τους κωδικοποιημένα τα χαρακτηριστικά τους. Το σύνολο αυτής της κωδικοποιημένης γενετικής πληροφορίας ονομάζεται, όπως είπαμε, γονότυπος. Ο γονότυπος δεν είναι αντιληπτός με τις φυσικές αισθήσεις των έμβιων όντων. Αντίθετα, αντιληπτή γίνεται η αλληλεπίδραση του με το περιβάλλον, που έχει ως αποτέλεσμα την ορατή εμφάνιση των χαρακτηριστικών αυτών.

Ανάλογος είναι ο ρόλος της αποκωδικοποίησης στο τεχνητό μοντέλο. Εδώ το ρόλο του γονότυπου παίζει η δομή της συμβολοσειράς με τα δυαδικά ψηφία ως αντίστοιχα των γονιδίων. Ο φαινότυπος αναφέρεται στην παρατηρήσιμη εμφάνιση μιας συμβολοσειράς, στο πώς φαίνεται στο περιβάλλον της. Περιβάλλον, όμως, θεωρείται η αντικειμενική συνάρτηση, άρα ο φαινότυπος μιας συμβολοσειράς αντιστοιχεί στην αποκωδικοποιημένη τιμή της, που ανήκει στο σύνολο ορισμού της αντικειμενικής συνάρτησης.

Σκοπός της λειτουργίας αξιολόγησης είναι να υπολογιστεί για κάθε άτομο του πληθυσμού η ικανότητα του για επιβίωση. Στη φύση οι ικανότητες των ατόμων δεν είναι προσδιορίσιμες με αυστηρό τρόπο. Είναι, όμως, καθορισμένες από το γενετικό υλικό των χρωμοσωμάτων τους. Συνεπώς, ο υπολογισμός της ικανότητας είναι θεμελιώδης λειτουργία για το Γ.Α. Η εφαρμογή της είναι πολύ απλή (τουλάχιστον για απλά προβλήματα): για κάθε συμβολοσειρά του τρέχοντος πληθυσμού υπολογίζεται η απόδοσή της από την ήδη γνωστή αντικειμενική συνάρτηση. Σε πιο σύνθετα προβλήματα, ο υπολογισμός ικανότητας μπορεί να ισοδυναμεί με την εκτέλεση μιας εργαστηριακής προσομοίωσης.



Σχήμα 2.1 Τα βασικά βήματα ενός γενετικού αλγόριθμου.

2.2 ΑΝΑΠΑΡΑΓΩΓΗ

Σειρά έχει τώρα η σημαντικότερη λειτουργία του ΓΑ, η αναπαραγωγή, όπου εδώ γίνεται ο κύριος όγκος της εργασίας του αλγορίθμου. Η δομή της αναπαραγωγικής διαδικασίας είναι σύνθετη. Περιλαμβάνει τα εξής μέρη: διασταύρωση και μετάλλαξη. Όμως πριν από την αναπαραγωγή, εκτελείται η διαδικασία της επιλογής.

Με την επιλογή, εφαρμόζεται στα πλαίσια του αλγορίθμου, ο νόμος της επιβίωσης του ικανότερου. Μέσω αυτής της διαδικασίας, καθορίζεται ποια άτομα από τον υπάρχοντα πληθυσμό θα έχουν την ευκαιρία να συμμετέχουν στην αναπαραγωγή και να κληροδοτήσουν στην επόμενη γενιά τα χαρακτηριστικά τους. Στόχος της λειτουργίας της επιλογής είναι να αφήνει τα ικανότερα άτομα να αυξάνονται εκθετικά

και να επικρατούν στο τέλος μετά την αναπαραγωγή αρκετών γενεών. Ένας ΓΑ χωρίς επιλογή στην αναπαραγωγική του διαδικασία θεωρείται ότι κάνει τυχαίο ψάξιμο.

Για την πιθανότητα επανεμφάνισης ενός χρωμοσώματος στην επόμενη γενιά, υπάρχει σχετικά το ***Θεώρημα Σχήματος*** (*schema theorem*). Σχήμα θεωρείται κάθε χρωμόσωμα με συγκεκριμένες θέσεις αυτού (γονίδια) να καταλαμβάνονται από δεδομένες τιμές. Για παράδειγμα για ένα δυαδικό χρωμόσωμα 8 θέσεων όπου η $5^{η}$ και $7^{η}$ θέση λαμβάνουν τιμές 1 και 0 αντιστοίχως, ενώ οι υπόλοιπες θέσεις αυτού μπορούν να πάρουν οποιαδήποτε τιμή, το αντίστοιχο σχήμα είναι το ****1*0*. Άλλα παραδείγματα σχημάτων μπορεί είναι 1***01** , 01**1*** κ.α. Για κάθε σχήμα ορίζεται η **τάξη** (*order*) του σχήματος η οποία δηλώνει το πλήθος των θέσεων (γονιδίων) που έχουν δεδομένες τιμές. Για το σχήμα ****1*0* η τάξη είναι 2, ενώ για τα 1***01** , 01**1*** η τάξη είναι 3. Επίσης ορίζεται το **μήκος** (*length*), το οποίο είναι το πλήθος των θέσεων (γονιδίων) μετά το πρώτο και έως το τελευταίο καθορισμένο γονίδιο. Για το ****1*0* το μήκος είναι 2, για το 1***01** είναι 5 και για το 01**1*** είναι 4. Αν ένα συγκεκριμένο σχήμα χρωμοσώματος επιλέγεται ως γονέας στην διαδικασία επιλογής, η πιθανότητα να υπάρχει και στην επόμενη γενιά το συγκεκριμένο σχήμα, μετά την διασταύρωση και μετά την μετάλλαξη, και συγκεκριμένα το πλήθος των χρωμοσωμάτων που θα περιέχουν το σχήμα αυτό στην επόμενη γενιά, δίνεται από το *Θεώρημα Σχήματος*. Το πλήθος των συγκεκριμένων σχημάτων που θα υπάρχουν στην επόμενη γενιά θα είναι ανάλογη του γινομένου του πλήθους των σχημάτων που υπάρχουν στην υπάρχουσα γενιά επί την πιθανότητα επιλογής του σχήματος αυτού επί μία πιθανότητα η οποία μειώνεται καθώς αυξάνεται το μήκος και η τάξη του σχήματος. Έτσι, το Θεώρημα Σχήματος λέει πως σχήματα μικρού μήκους και τάξης είναι πιο πιθανό να διατηρούνται στις επόμενες γενιές. Επομένως, τέτοια σχήματα θα είναι αυτά που παίζουν σημαντικότατο ρόλο στην εξέλιξη της κάθε γενιάς και τελικά την εύρεση καλύτερης λύσης.

Υπάρχουν διάφοροι τρόποι υλοποίησης της επιλογής στα πλαίσια ενός ΓΑ. Το μέγεθος του πληθυσμού από γενιά σε γενιά δεν αλλάζει (στη βασική μορφή του αλγορίθμου) έτσι λοιπόν η επιλογή οφείλει να δίνει μεγαλύτερες πιθανότητες αναπαραγωγής στα πιο ικανά άτομα. Ο τελεστής αναπαραγωγής μπορεί να εκφραστεί σε αλγοριθμική βάση, με πολλούς τρόπους. Ίσως ο ευκολότερος από αυτούς είναι η έκφραση μέσω μιας εξαναγκασμένης ρουλέτας, στην οποία κάθε συμβολοσειρά ενός πληθυσμού αντιπροσωπεύεται σε ένα μέρος της ρουλέτας, σε αναλογία με την απόδοσή της.

Η δομή ενός απλού γενετικού αλγορίθμου έχει σε γενικές γραμμές ως εξής:

Κατά την διάρκεια της επαναληπτικής εκτέλεσης *t*, ο ΓΑ διατηρεί ένα πληθυσμό από πιθανές λύσεις:

$P(t) = \{x_1^t, \ldots, x_n^t\}$.

Κάθε λύση x_i^t αξιολογείται και δίνει ένα μέτρο της καταλληλότητας και ορθότητάς της. Αφού ολοκληρωθεί η αξιολόγηση όλων των στοιχείων του πληθυσμού, δημιουργείται ένας νέος πληθυσμός (επαναληπτική εκτέλεση $t + 1$) που προκύπτει από την επιλογή των πιο κατάλληλων στοιχείων του πληθυσμού της προηγούμενης γενιάς.

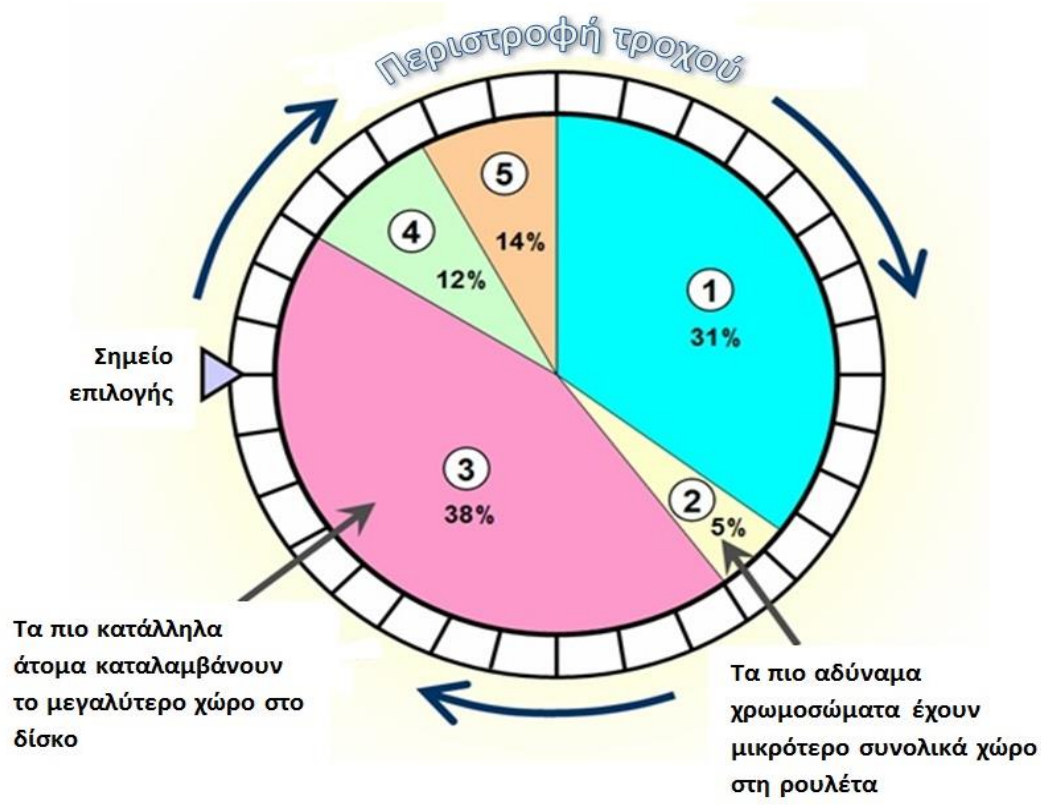
Μερικά μέλη από τον καινούριο αυτό πληθυσμό υφίστανται μετατροπές με τη βοήθεια των διαδικασιών της *μετάλλαξης (mutation)* και της *διασταύρωσης (crossover* ή *mating)* σχηματίζοντας νέες πιθανές λύσεις. Η διασταύρωση συνδυάζει τα στοιχεία δύο χρωμοσωμάτων γονέων για να δημιουργήσει δύο νέους απογόνους ανταλλάσσοντας αντίστοιχα κομμάτια από τους γονείς Για παράδειγμα, έστω ότι οι γονείς αναπαριστώνται με διανύσματα πέντε διαστάσεων $(a1,b1,c1,d1,e1)$ και $(a2,b2,c2,d2,e2)$, τότε οι απόγονοι (με *σημείο διασταύρωσης —crossover point* = 2) είναι οι $(a1,b1,c2,d2,e2)$ και $(a2,b2,c1,d1,e1)$. Διαισθητικά μπορούμε να πούμε ότι η διασταύρωση εξυπηρετεί την ανταλλαγή πληροφοριών μεταξύ διαφορετικών πιθανών λύσεων. Εδώ πρέπει να γίνει η εξής παρατήρηση. Αν οι μεταβλητές στα παραπάνω διανύσματα είναι δυαδικές, τότε κάθε διάνυσμα αναπαριστά την τιμή μιας μεταβλητής, δηλαδή ένα χρωμόσωμα. Στην περίπτωση που είναι πραγματικές, τότε καθεμία είναι ένα χρωμόσωμα, δηλαδή κάθε διάνυσμα αναπαριστά τις τιμές πολλών μεταβλητών, δηλαδή αποτελεί ένα γονότυπο. Για παράδειγμα, η βελτιστοποίηση μίας συνάρτησης πολλών μεταβλητών, απαιτεί την κωδικοποίηση της λύσης με ένα γονότυπο.

Η διαδικασία της μετάλλαξης αλλάζει αυθαίρετα ένα ή περισσότερα γονίδια ενός συγκεκριμένου χρωμοσώματος. Πραγματοποιείται με τυχαία αλλαγή γονιδίων και με πιθανότητα ίση με το *ρυθμό μετάλλαξης (mutation rate)***.** Διαισθητικά μπορούμε να πούμε ότι η μετάλλαξη εξυπηρετεί την εισαγωγή νέων πιθανών λύσεων, διαφορετικών από τις υπάρχουσες, στον ήδη υπάρχοντα πληθυσμό.

Γενικά υπάρχουν αρκετοί τρόποι ώστε να γίνει η διαδικασία της επιλογής. Κάποιοι από αυτούς είναι:

- <u>Ποσοστιαία επιλογή</u> *(μέθοδος του «τροχού της τύχης» (Roulette wheel method)):*

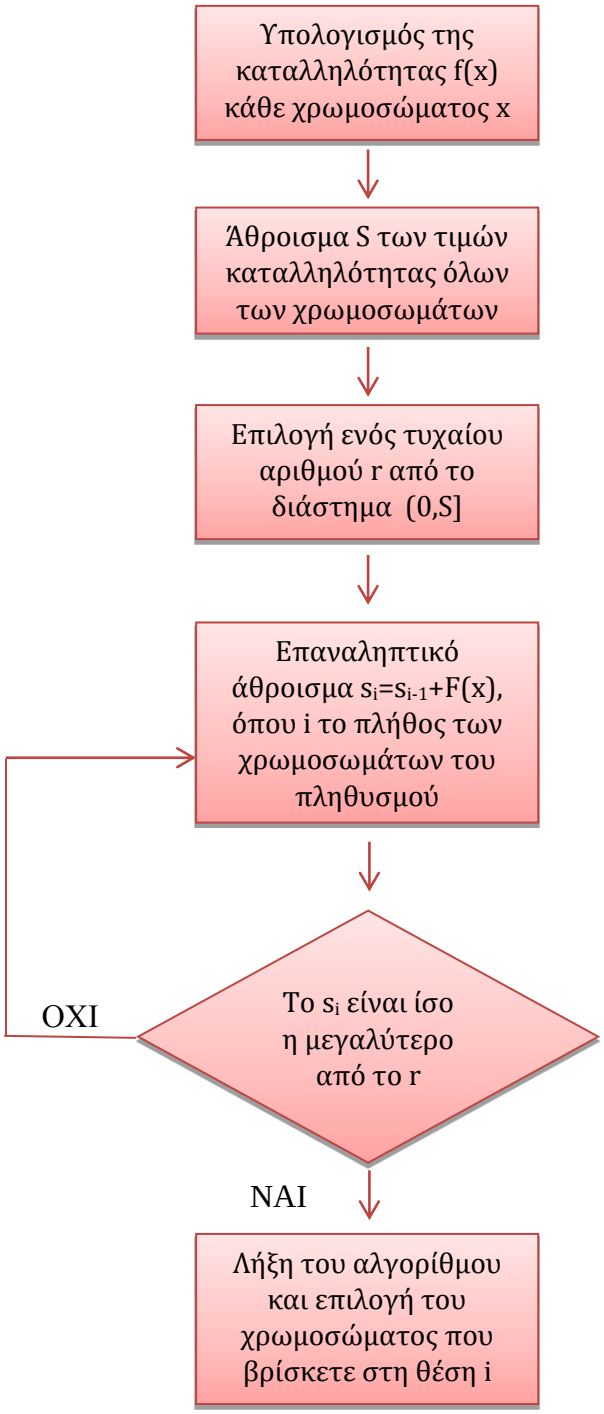
  Αυτή η μέθοδος είναι από τις πιο συνηθισμένες μεθόδους επιλογής. Εδώ ο προσδοκώμενος αριθμός των απογόνων ενός ατόμου i δίνεται από τη σχέση $p_i = \frac{f_i}{\bar{f}}$ όπου η $f: S \to \mathbb{R}^+$ είναι η συνάρτηση καταλληλότητας και η $\bar{f}$ η μέση τιμή της καταλληλότητας όλων των ατόμων. Αν φανταστούμε το συνολικό πληθυσμό να σχηματίζει το δίσκο της ρουλέτας και κάθε άτομο του πληθυσμού να αντιπροσωπεύεται από ένα χώρο ανάλογο με την καταλληλότητά του, τότε προκύπτει η αναλογία με τη ρουλέτα. Έτσι, καθώς «γυρνάμε επαναλαμβανόμενα τον τροχό», τα άτομα επιλέγονται χρησιμοποιώντας τυχαία δειγματοληψία με αντικατάσταση. Για να γίνει η ποσοστιαία επιλογή ανεξάρτητη από τις διαστάσεις των τιμών της καταλληλότητας, χρησιμοποιούνται συνήθως οι λεγόμενες τεχνικές δημιουργίας παραθύρων. Άλλες παραλλαγές της μεθόδου του «τροχού της τύχης» στοχεύουν στη μείωση της κυριαρχικής θέσης στην αναπαραγωγή ενός ή μιας ομάδας από άκρως κατάλληλα άτομα («super individuals») χρησιμοποιώντας στοχαστικές τεχνικές δειγματοληψίας.



Σχήμα 2.2 μέθοδος ποσοστιαίας επιλογής – "τροχός της τύχης"

Η μέθοδος roulette wheel υλοποιείται βάσει του παρακάτω αλγορίθμου (***Σχήμα 2.3***):

1. ***[Υπολογισμός Καταλληλότητας]*** Υπολογισμός της συνάρτησης καταλληλότητας κάθε χρωμοσώματος του πληθυσμού.
2. ***[Άθροισμα]*** Άθροισμα ***S*** της καταλληλότητας όλων των χρωμοσωμάτων του πληθυσμού.
3. ***[Επιλογή]*** Παραγωγή ενός τυχαίου αριθμού ***r*** στο διάστημα ***(0,S)*** .
4. ***[Βρόγχος]*** Πρόσθεση των συναρτήσεων καταλληλότητας των χρωμοσωμάτων από το ***0*** έως το ***s***. Όταν το ***s*** γίνει μεγαλύτερο από το ***r***, επιστροφή του χρωμοσώματος που βρίσκεται στη θέση στην οποία το ***s*** ξεπέρασε το ***r.***



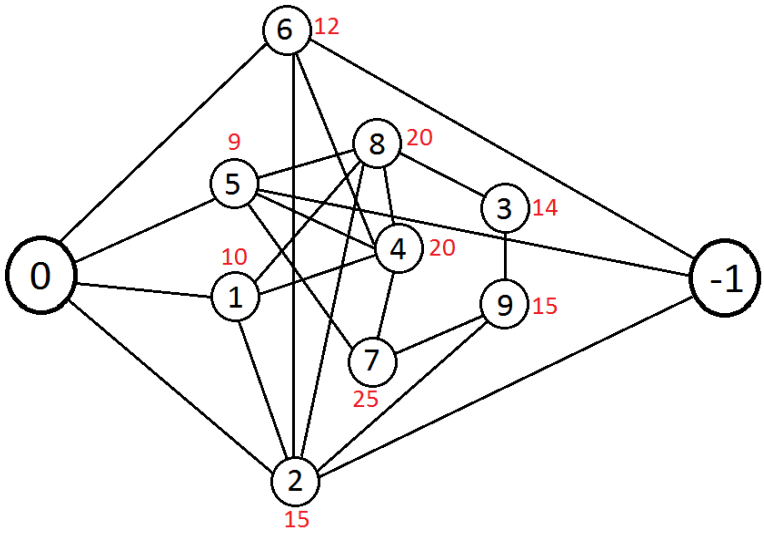
Σχήμα 2.3: Σχηματική Αναπαράσταση Τεχνικής της Ρουλέτας (roulette wheel)

Η συγκεκριμένη μέθοδος επιλογής διαφέρει από τις υπόλοιπες επειδή δίνει σε κάθε μέλος του πληθυσμού την ευκαιρία να γίνει γονέας. Το μειονέκτημά της είναι ότι εμφανίζεται να είναι ευπαθής στην συνάρτηση καταλληλότητας. Η διαδικασία της επιλογής ολοκληρώνεται με την επανατοποθέτηση των χρωμοσωμάτων στον πληθυσμό.

Έστω για παράδειγμα πως έχουμε ένα πρόβλημα όπου το ζητούμενο είναι να περάσουμε από τους κόμβους του σχήματος ***2.4*** με αφετηρία τον κόμβο "0" και τέρμα τον κόμβο "-1", με τις ακμές να δηλώνουν της δυνατές μεταβάσεις. Περιορισμός είναι να περάσουμε μία το πολύ φορά από έναν κόμβο. Τα κέρδη διάβασης από κάθε κόμβο δίνονται από μία συνάρτηση P(…). Τα κέρδη αυτά για τους κόμβους 1, 2, 3, 4,

5, 6, 7, 8 και 9 είναι αντιστοίχως 10, 15, 14, 20, 9, 12, 25, 20 και 15. Έστω τώρα ότι έχουμε έναν πληθυσμό που υπάρχουν πέντε υποψήφιες εφικτές λύσεις, που όπως είναι γνωστό καλούνται χρωμοσώματα :

X1: [0 6 2 1 8 5 -1]
X2: [0 5 4 8 2 -1]
X3: [0 1 2 9 7 4 6 -1]
X4: [0 5 4 8 2 -1]
X5: [0 2 9 3 8 4 7 5 -1]



Σχήμα 2.4: Σχεδιάγραμμα διαδρομής με αρχή τον κόμβο 0 και τερματισμό στον κόμβο -1. Σε κάθε κόμβο είναι σημειωμένο το αντίστοιχο κέρδος του με κόκκινο.

Τα ψηφία 0 και -1 είναι τα σημεία έναρξης και τερματισμού αντίστοιχα, ενώ η ενδιάμεση αλυσίδα αριθμών του χρωμοσώματος είναι οι κόμβοι από τους οποίους θα περάσουμε, π.χ. για το χρωμόσωμα X1 η αλυσίδα 6-2-1-8-5 δηλώνει πως θα περάσουμε με την σειρά τους κόμβους 6,2,1,8 και 5. Ας δούμε αναλυτικά τα βήματα που ακολουθούνται για την κατανόηση της επιλογής κάθε χρωμοσώματος:

Βήμα 1: Υπολογίζεται η καταλληλότητα κάθε υποψήφιας λύσης.

$$f(X1) = P(6) + P(2) + P(1) + P(8) + P(5) = 12 + 15 + 10 + 20 + 9 = 66$$

Ομοίως $f(X2)=64$, $f(X3)=97$, $f(X4)=64$ και $f(X5)=118$

Βήμα 2: Αθροίζονται οι καταλληλότητες όλων των χρωμοσωμάτων του πληθυσμού.

$S = f(X1) + f(X2) + f(X3) + f(X4) + f(X5) = 66 + 64 + 97 + 64 + 118 = 409$

Βήμα 3: Παράγεται ένας τυχαίος αριθμός r στο διάστημα $(0,S)$.

Έστω ότι ο τυχαίος αριθμός που παράγεται, με τη βοήθεια ομοιόμορφης κατανομής, στο διάστημα (0,409) είναι ο $r = 220$.

Βήμα 4: Προστίθενται οι καταλληλότητες όλων των χρωμοσωμάτων έως ότου το άθροισμά τους s υπερβεί τον αριθμό r. Όταν το s γίνει μεγαλύτερο από το $\boldsymbol{r}$, επιστρέφεται το χρωμόσωμα που βρίσκεται στη θέση στην οποία το s ξεπέρασε το r.

s = f (X1) = 66
s = f (X1) + f (X2) = 66 + 64 = 130
s = f (X1) + f (X2) + f (X3) = 66 + 64 + 97 = 227 > r

Οπότε το πρώτο υποψήφιο προς διασταύρωση χρωμόσωμα που θα επιλέξει η ρουλέτα θα είναι το X3: [0 1 2 9 7 4 6 -1].

Στη συνέχεια επαναλαμβάνεται η διαδικασία που περιγράφηκε προηγουμένως (συγκεκριμένα τα βήματα 3 έως 4) για να επιλεχθούν και τα υπόλοιπα υποψήφια προς διασταύρωση χρωμοσώματα με μη μηδενική την πιθανότητα να επιλεχθεί κάποιο χρωμόσωμα περισσότερες από μία φορές ή ακόμα και καμία ως γονέας. Δηλαδή, ο πληθυσμός των γονέων μπορεί να έχει κάποιο χρωμόσωμα περισσότερες από μία φορές ως γονέα, ή ακόμα και να μην περιέχει καθόλου κάποια από τα χρωμοσώματα του πληθυσμού που προϋπήρχε.

- *Επιλογή γραμμικής τάξεως(linear-rank selection):*

  Σύμφωνα με την επιλογή γραμμικής τάξεως τα άτομα του πληθυσμού ταξινομούνται ανάλογα με την καταλληλότητα τους και τα αντίγραφά τους εκχωρούνται με τέτοιο τρόπο ώστε το καλύτερο άτομο λαμβάνει ένα προκαθορισμένο πολλαπλάσιο του αριθμού των αντιγράφων που λαμβάνει το χειρότερο. Η γραμμική επιλογή μπορεί να μειώνει έμμεσα την κυριαρχία των ατόμων που έχουν πολύ καλύτερη τιμή καταλληλότητας από τα υπόλοιπα, αλλά στρεβλώνει τη διαφορά μεταξύ ατόμων με κοντινές τιμές καταλληλότητας. Αυξάνεται με αυτόν τον τρόπο η διαφορά στην πιθανότητα επιλογής διαφορετικών ατόμων, όταν έχουμε στάσιμους πληθυσμούς. Ακόμη και αν έχει χρησιμοποιηθεί με κάποια επιτυχία η επιλογή γραμμικής τάξεως, αγνοεί τις πληροφορίες για τις διαφορές στην τιμή της καταλληλότητας των διαφόρων ατόμων και έτσι παραβιάζει το θεώρημα σχήματος (*schema theorem*).

- *Τυχαία επιλογή από ένα σύνολο γονέων (Tournament selection)*

  Υπάρχουν διάφορες εκδοχές αυτής της μεθόδου. Η πιο κοινή από αυτές είναι η (*n-tournament selection)* **ν-τυχαία επιλογή**, όπου n άτομα επιλέγονται τυχαία με επανάθεση από τον πληθυσμό και από αυτά, εκείνο με την καλύτερη τιμή καταλληλότητας κρατιέται ως ένας γονέας. Για κάθε γονέα επιλέγεται ξανά ένα αντίστοιχο μέρος n ατόμων/χρωμοσωμάτων του πληθυσμού και κρατείται εκείνο με την καλύτερη τιμή καταλληλότητας. Σε αυτή την εκδοχή υπάρχει η δυνατότητα παραλλαγής, όπου στην επιλογή του ενδιάμεσου πληθυσμού των n ατόμων μπορεί να διαφοροποιηθεί το μέγεθος n σε κάθε βήμα και έτσι να κλιμακωθεί η επιλογή των γονέων με κατάλληλες τιμές της παραμέτρου n. Η συγκεκριμένη μέθοδος ακολουθήθηκε για την επιλογή γονέων στην παρούσα εργασία.

2.3 ΔΙΑΣΤΑΥΡΩΣΗ

Ο πληθυσμός που έχουμε μέχρι τώρα από τη διαδικασία της επιλογής πρέπει να περάσει από τη φάση του «ζευγαρώματος» για να πραγματοποιηθεί ένα είδος γονιμοποίησης, όπως συμβαίνει και στη φύση. Η νέα, λοιπόν, ομάδα ατόμων που προέκυψε από την επιλογή σχηματίζει με τυχαίο τρόπο ομάδες των δύο. Το ποιος θα ζευγαρώσει με ποιον, από τα άτομα του προσωρινού πληθυσμού, υπάρχει περίπτωση να επηρεάζει την ταχύτητα σύγκλισης του αλγορίθμου. Προς το παρόν όμως αυτό αποτελεί αντικείμενο μελέτης και στη βιβλιογραφία σε όλες τις εφαρμογές το ζευγάρωμα γίνεται με τυχαίο τρόπο.

Σε κάθε ομάδα, τα δύο μέλη παίρνουν μέρος σε μια απλή λειτουργία ανταλλαγής γενετικού υλικού που ονομάζεται διασταύρωση. Η διασταύρωση είναι μια απαραίτητη λειτουργία που συμβάλει αποφασιστικά στην επίδοση ενός ΓΑ και γι' αυτόν ακριβώς το λόγο έχουν επινοηθεί πολλοί τρόποι υλοποίησης του. Μερικοί μπορούν να εφαρμοστούν σε κάθε τύπο προβλήματος, ενώ άλλοι είναι πιο κατάλληλοι και εξειδικευμένοι για ειδικές περιπτώσεις. Στόχος της διασταύρωσης είναι η νέα γενιά που θα προκύψει μετά την εφαρμογή της να περιλαμβάνει άτομα που θα διαφέρουν από τους γονείς τους και θα φέρουν συνδυασμό των καλύτερων χαρακτηριστικών τους. Ερευνητές που ασχολούνται χρόνια με τους ΓΑ υποστηρίζουν ότι, αν αφαιρεθεί η διασταύρωση από έναν ΓΑ, τότε μειώνεται σημαντικά η απόδοσή του, αλλά αυτή δεν είναι μια άποψη ευρέως αποδεκτή.

Ένα ενδεικτικό της χρησιμότητας της διασταύρωσης είναι η ανακατεύθυνση του ψαξίματος σε «παρθένες» περιοχές του χώρου αναζήτησης. Έτσι διευρύνεται το πεδίο δράσης του αλγορίθμου και αυξάνουν οι πιθανότητες επιτυχίας του. Επίσης, τα νέα άτομα περιλαμβάνουν συνδυασμούς χαρακτηριστικών των γονέων τους και με αυτό τον τρόπο μπορούν να προκύψουν επιτυχημένοι συνδυασμοί υψηλής ικανότητας. Στην περίπτωση που η διασταύρωση δώσει χειρότερα παιδιά από τους γονείς, αυτά δεν θα έχουν μεγάλη πιθανότητα πολλαπλασιασμού στον επόμενο αναπαραγωγικό κύκλο, λόγω μικρής απόδοσης. Στην πράξη, η διασταύρωση χρησιμοποιείται με παραμετροποιημένη μορφή, δηλαδή λαμβάνει χώρα με πιθανότητα, την λεγόμενη *πιθανότητα διασταύρωσης (crossover probability) pc*, που καθορίζεται από το σχεδιαστή του ΓΑ. Συνήθως, αυτή η πιθανότητα ποικίλει από πρόβλημα σε πρόβλημα, ενώ είναι δυνατό και να αλλάζει κατά τον χρόνο τρεξίματος άρα επηρεάζει και τη σύγκλισή του. Η τιμή pc=1, σημαίνει συνεχή εφαρμογή του τελεστή διασταύρωσης, το ψάξιμο γίνεται με μικρό βήμα, άρα η αναζήτηση γίνεται σε όλο το χώρο και ο αλγόριθμος θα συγκλίνει στο βέλτιστο, αλλά πολύ αργά. Για μικρές τιμές της pc, το ψάξιμο κάνει άλματα, άρα ο αλγόριθμος είναι πιθανόν να συγκλίνει πιο γρήγορα. Υπάρχει βέβαια πάντα ο φόβος, χρησιμοποιώντας μεγάλο βήμα, ο αλγόριθμος να ξεπεράσει το βέλτιστο και έτσι να αποκλίνει, γι 'αυτό επιλέγουμε συνήθως μεγάλο βήμα στην αρχή του ψαξίματος, και στη συνέχεια, όταν ο αλγόριθμος προσεγγίσει την τιμή του βέλτιστου, χρησιμοποιούμε μικρό βήμα αναζήτησης. Με αυτό τον τρόπο, μπορούμε να αυξήσουμε την ταχύτητα αναζήτησης, χωρίς να κινδυνεύουμε να αποκλίνει ο αλγόριθμος. Οι γονείς διασταυρώνονται με

πιθανότητα Pc (0≤Pc≤1). Η πιθανότητα Pc καθορίζει πόσο συχνά εκτελείται η διασταύρωση. Για παράδειγμα, έστω ότι η πιθανότητα διασταύρωσης (Pc) δύο γονέων ορίζεται ίση με 0.3. Με τη βοήθεια ομοιόμορφης κατανομής παράγεται ένας τυχαίος αριθμός rc στο διάστημα [0,1]. Αν ο αριθμός rc είναι μικρότερος ή ίσος από 0.3 τότε πραγματοποιείται διασταύρωση, διαφορετικά ως απόγονοι καλούνται τα ακριβή αντίγραφα των γονέων. Εάν η πιθανότητα διασταύρωσης (Pc) είναι ίση με 1 τότε όλοι οι απόγονοι δημιουργούνται από κομμάτια και των δύο γονέων, ενώ εάν η πιθανότητα διασταύρωσης (Pc) είναι ίση με 0 τότε όλοι οι απόγονοι αποτελούν ακριβή αντίγραφα των γονέων τους.

Υπάρχουν αρκετές τεχνικές διασταύρωσης. Παρουσιάζουμε μερικές εδώ:

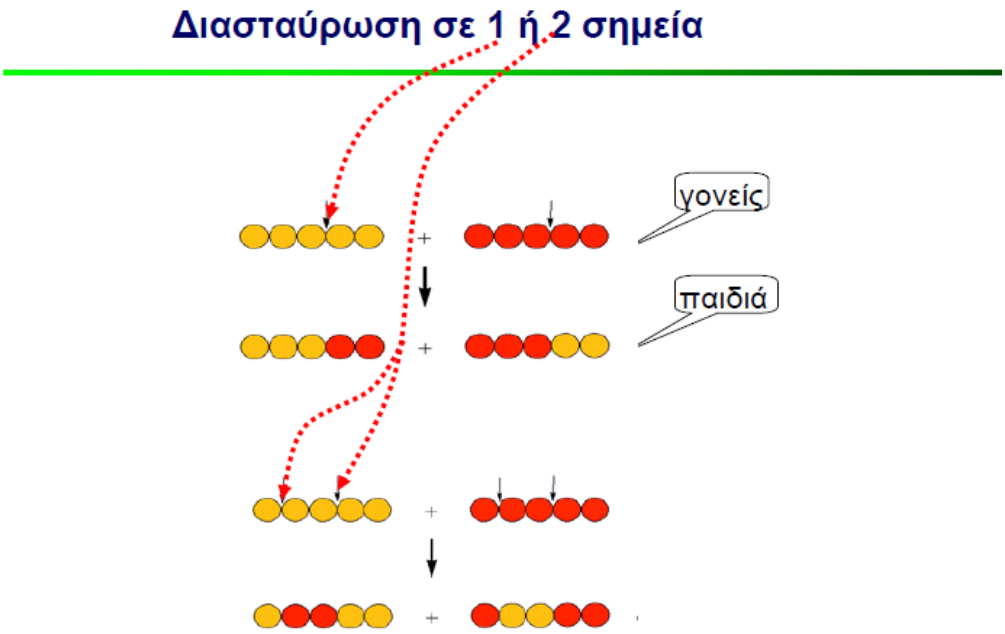
- *Διασταύρωση ενός σημείου (single-point crossover):*
  κατά τη διαδικασία αυτή σε κάθε ένα από τα χρωμοσώματα-γονείς δημιουργείται μια τομή σε τυχαίο σημείο έτσι ώστε να δημιουργηθούν δυο τμήματα κεφαλής και δυο ουράς. Έπειτα τα τμήματα της ουράς εναλλάσσονται ώστε να δημιουργηθούν δυο καινούργια άτομα.

- *Διασταύρωση πολλαπλών σημείων (multiple-point crossover):*
  Η διαδικασία αυτή είναι μια φυσική επέκταση της διασταύρωσης ενός σημείου. Σε μια διασταύρωση N-σημείων υπάρχουν N τυχαία σημεία τομής στα χρωμοσώματα και γίνονται εναλλαγές μεταξύ των τμημάτων που προκύπτουν. Σύμφωνα με κάποιους ερευνητές η διασταύρωση πολλαπλών σημείων επειδή παίρνει δείγματα ομοιόμορφα από ολόκληρο το μήκος του χρωμοσώματος, είναι πιο κατάλληλη στο να συνδυάζει τα καλά χαρακτηριστικά που υπάρχουν στα άτομα. Την ίδια στιγμή η διασταύρωση αυτή μπορεί να γίνει αποδιοργανωτική καθώς αυξάνεται ο αριθμός των σημείων διασταύρωσης, δηλαδή η δημιουργία μεγαλύτερων δομικών στοιχείων γίνεται όλο και πιο δύσκολη. Μειώνοντας τον αριθμό των σημείων διασταύρωσης κατά τη διάρκεια της εκτέλεσης του γενετικού αλγορίθμου μπορεί να είναι μια καλή συμβιβαστική λύση.

Στη διασταύρωση δύο σημείων επιλέγονται δύο σημεία σε κάθε γονέα και όλα τα στοιχεία που βρίσκονται ανάμεσα στα δύο σημεία ανταλλάσσονται, με αποτέλεσμα να δημιουργούνται δύο απόγονοι. Αν, τυχαία, οι δύο γονείς είναι ακριβώς οι ίδιοι τότε οι δύο απόγονοι θα ταυτίζονται με τους γονείς). Τα σημεία είναι κοινά και για τους δύο γονείς και θα πρέπει να είναι μικρότερα ή ίσα από το μήκος του μικρότερου χρωμοσώματος-γονέα.

Για παράδειγμα, έστω ότι οι γονείς είναι τα χρωμοσώματα X_1 και X_2:
X1 = [0 2 8 10 7 6 4 13 -1]
X2 = [0 9 3 20 12 5 1 19 21 11 -1]
Το μήκος του X_1 είναι 9 ενώ το μήκος του X_2 είναι 11. Τα δύο σημεία που θα επιλεγούν θα πρέπει να ανήκουν στο διάστημα [1,9). Το 9 δεν συμπεριλαμβάνεται στο διάστημα διότι θα προκαλέσει τη δημιουργία μη εφικτού απογόνου ενώ το 1

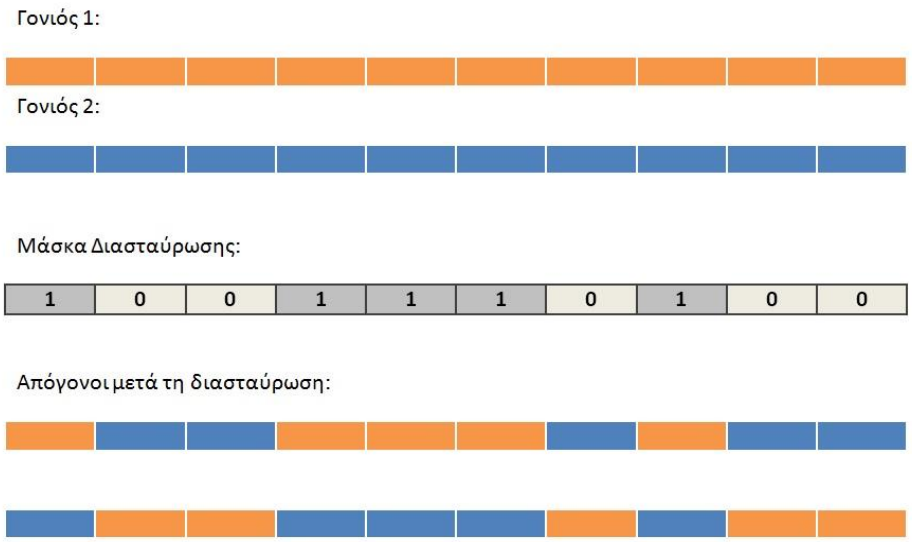
συμπεριλαμβάνεται στο διάστημα διότι δημιουργεί εφικτούς απογόνους. Έστω ότι τα σημεία που επιλέγονται είναι τα α=3 και β=7. Και οι δύο γονείς διαχωρίζονται σε αυτές τις θέσεις και ανταλλάσουν τους πελάτες που βρίσκονται ανάμεσα στη θέση 3 και στη θέση 7.
1ος Γονέας: X_1 = [0 2 8 | 10 7 6 4 | 13 -1]
2ος Γονέας: X_2 = [0 9 3 | 20 12 5 1| 19 21 11 -1]
με αποτέλεσμα να δημιουργούν τους απογόνους Y_1 και Y_2.
1ος Απόγονος: Y1 = [0 2 8 | 20 12 5 1| 13 -1]
2ος Απόγονος: Y2 = [0 9 3 | 10 7 6 4 | 19 21 11 -1]



Σχήμα 2.4: Διασταύρωση ενός και δύο σημείων αντιστοίχως για δύο χρωμοσώματα.

- *Ομοιόμορφη διασταύρωση(Uniform Crossover):*
  Δεδομένων δύο γονέων, κάθε γονίδιο στους απογόνους δημιουργείται από την αντιγραφή του αντίστοιχου γονίδιου από έναν από τους γονείς. Η επιλογή του γονέα γίνεται μέσω μιας μάσκας διασταύρωσης η οποία δημιουργείται τυχαία: Σε κάθε θέση, το γονίδιο των απογόνων έχει ληφθεί από τον πρώτο γονέα εάν υπάρχει 1 στην μάσκα σε αυτό το σημείο, και αντίστροφα (αν υπάρχει ένα 0 στην μάσκα σε αυτό το σημείο) το γονίδιο λαμβάνεται από τον δεύτερο γονέα. Στην ομοιόμορφη διασταύρωση οι δύο γονείς μεταδίδουν στον απόγονο που δημιουργείται κάθε γονίδιό τους με πιθανότητα 0.5. Αν οι δύο γονείς έχουν διαφορετικό μήκος τότε στον απόγονο ή μεταδίδεται το γονίδιο του ενός γονέα με το μεγαλύτερο μήκος ή τίποτα. Για παράδειγμα, έστω ότι οι γονείς είναι τα χρωμοσώματα X1 και X2:
  1ος Γονέας: X1 = [0 2 8 10 7 6 4 13 -1]
  2ος Γονέας: X2 = [0 9 3 20 12 5 1 19 21 11 -1]
  Τότε ο απόγονος που θα δημιουργηθεί ενδέχεται να είναι ο εξής:
  Απόγονος: Y = [0 2 3 20 7 5 4 13 11 -1]

Και στις δύο περιπτώσεις διασταύρωσης, οι απόγονοι - πριν ελεγχθούν για την εφικτότητά τους ως προς τους χρονικούς περιορισμούς – επιδιορθώνονται από τυχόν διπλούς πελάτες. Δηλαδή, αν ένας απόγονος περιέχει έναν πελάτη δύο φορές, τότε διαγράφεται ο πελάτης από τη μία θέση (τυχαία) και στη συνέχεια ο απόγονος ελέγχεται ως προς την εφικτότητά του.



Σχήμα 2.5: Ομοιόμορφη διασταύρωση για δύο χρωμοσώματα, με την βοήθεια της μάσκας διασταύρωσης, η οποία προκύπτει από την επιλογή ή όχι του ένα σε κάθε συντεταγμένη βάσει της ομοιόμορφης κατανομής.

Η επιλογή του κατάλληλου τελεστή διασταύρωσης εξαρτάται πολύ από την αναπαράσταση του χώρου αναζήτησης του προβλήματος. Τα ακολουθιακά προβλήματα, όπως τα προβλήματα αναζήτησης πορείας για παράδειγμα, συχνά απαιτούν τη χρήση διαφορετικών τελεστών από αυτούς που περιγράψαμε παραπάνω καθώς ενδέχεται οι απόγονοι που θα δημιουργηθούν να είναι έξω από το σύνολο των επιτρεπόμενων λύσεων.

Για το πρόβλημα του περιοδεύοντος πωλητή που έχουμε αναφέρει στην αρχή του κεφαλαίου παρουσιάζει τέτοια προβλήματα. Το κάθε χρωμόσωμα είναι μία διάταξη των αριθμών 1,2,3,…,p όπου p το πλήθος των πόλεων για το πρόβλημα., και λόγω της ιδιαιτερότητάς του (να εμφανίζεται κάθε αριθμός μόνο μία φορά σε κάθε χρωμόσωμα), παρουσιάζονται προβλήματα κατά την διασταύρωση όπως φαίνεται στα παρακάτω.

Έστω οι δύο χρωμοσώματα-γονείς:

752631948 και 912386754 και επιλογή τύπου διασταύρωσης ενός σημείου. Αν το σημείο αυτό είναι το $3^{ο}$ (μεταξύ $3^{ου}$ και $4^{ου}$ στοιχείου) τότε η διασταύρωση θα χώριζε τα χρωμοσώματα σε δύο μέρη:

752 - 631948 και 912 - 386754 για τα δύο χρωμοσώματα, οπότε τα παιδιά που θα προέκυπταν θα ήταν 752 - 386754 και 912 - 631948. Αλλά και τα δύο δεν είναι εφικτές λύσεις για το πρόβλημα καθώς υπάρχουν και στα δύο πόλεις που επαναλαμβάνονται στο χρωμόσωμα: **75**2386**75**4 και **91**263**19**48. Υπάρχει όμως ένας άλλος τύπος διασταύρωσης όπου επιλέγονται κάποιες θέσεις πόλεων (κατά την σειρά διέλευσης του πωλητή από αυτές, οπότε και συγκεκριμένες θέσεις στο χρωμόσωμα), π.χ. η $4^{η}$ η $6^{η}$ και $7^{η}$ θέση του χρωμοσώματος. Αυτές αποτελούν και τις θέσεις του $1^{ου}$ χρωμοσώματος γονέα που θα αλλάξουν στοιχεία σύμφωνα με τα στοιχεία του $2^{ου}$ χρωμοσώματος-γονέα, ενώ κάποιες άλλες 3 θέσεις του $2^{ου}$ χρωμοσώματος-γονέα θα αλλάξουν στοιχεία σύμφωνα με τα στοιχεία του $1^{ου}$ χρωμοσώματος-γονέα. Συγκεκριμένα για τα δύο χρωμοσώματα που έχουμε πάρει 752631948 και 912386754 ως γονείς, και για την επιλογή για τις θέσεις $4^{η}$,$6^{η}$ και $7^{η}$ του πρώτου χρωμοσώματος, έχουμε τα στοιχεία 752**6**3**19**48. Τα στοιχεία αυτά (το 6, 1 και 9) θα πάρουν την θέση των ίδιων στοιχείων στο δεύτερο χρωμόσωμα με την σειρά που υπάρχουν στο πρώτο. Δηλαδή για το δεύτερο χρωμόσωμα θα έχουμε την αλλαγή **91**238**6**754 σε **61**238**9**754 το οποίο θα αποτελεί και το πρώτο παιδί. Αντιστοίχως, για τις θέσεις $4^{η}$,$6^{η}$ και $7^{η}$ του δεύτερου χρωμοσώματος, έχουμε τα στοιχεία 912**3**8**67**54. Με τον ίδιο τρόπο τα στοιχεία αυτά (το 3, 6 και 7) θα πάρουν την θέση των ίδιων στοιχείων στο πρώτο χρωμόσωμα με την σειρά που υπάρχουν στο δεύτερο. Δηλαδή για το πρώτο χρωμόσωμα θα έχουμε την αλλαγή **7**52**63**1948 σε **3**52**67**1948, το οποίο θα αποτελεί και το δεύτερο παιδί. Έτσι, μπορεί να έχουμε μία αναδιάταξη των στοιχείων των δύο γονέων με τέτοιον τρόπο ώστε να προκύπτουν νέες εφικτές λύσεις. Φυσικά, υπάρχουν και άλλοι τρόποι διασταύρωσης και στο πρόβλημα το περιοδεύοντος πωλητή αλλά και σε κάθε πρόβλημα βελτιστοποίησης μπορεί να έχουμε τύπους διασταύρωσης προσαρμοσμένους στην μοντελοποίηση του προβλήματος.

2.4 ΜΕΤΑΛΛΑΞΗ

Τελευταία στον κύκλο αναπαραγωγικής διαδικασίας και, ίσως, λιγότερο σημαντική, αλλά πάντως χρήσιμη, είναι η μετάλλαξη. Είναι μια λειτουργία που όταν συμβαίνει αραιά στη φύση δρα βελτιωτικά για τους οργανισμούς και γενικά για την εξέλιξη της ζωής. Ανάλογος είναι ο ρόλος της και στα τεχνικά περιβάλλοντα. Ο πιο συνήθης τελεστής μετάλλαξης για προβλήματα δυαδικής κωδικοποίησης είναι η μετάλλαξη σε επίπεδο "μπιτ". Σε κάποιες περιπτώσεις γίνεται με τη δημιουργία ενός ψηφίου και την εισαγωγή του στο χρωμόσωμα και σε άλλες με την αναστροφή των ήδη υπαρχόντων ψηφίων. Καθώς αντιγράφονται δυαδικά ψηφία από τον γονέα στον απόγονο, επιλέγεται τυχαία με μικρή πιθανότητα, τη λεγόμενη *πιθανότητα μετάλλαξης (mutation probability)* p_m, ένα ψηφίο και αντιστρέφεται (από 0 σε 1 ή το αντίστροφο). Είναι πολύ σημαντικό η πιθανότητα να πραγματοποιηθεί η μετάλλαξη να είναι αρκετά μικρή (περίπου μία μετάλλαξη σε κάθε χίλια ψηφία που αντιγράφονται), γιατί σε αντίθετη περίπτωση ο ΓΑ εκφυλίζεται σε τυχαίο ψάξιμο. Ο τελεστής της μετάλλαξης επιτρέπει την τυχαία μετάβαση σε τελείως διαφορετικές περιοχές του χώρου αναζήτησης. Οι τυχαίες μεταλλάξεις που συμβαίνουν στον πληθυσμό

αλλάζουν ένα συγκεκριμένο ποσοστό των μεταβλητών των χρωμοσωμάτων του. Με την διαδικασία αυτή εισάγεται πληροφορία και χαρακτηριστικά που δεν υπήρχαν πριν στον πληθυσμό και αποφεύγεται με αυτόν τον τρόπο ο εγκλωβισμός του σε τοπικά ακρότατα της αντικειμενικής συνάρτησης. Η πιθανότητα μετάλλαξης καθορίζει πόσο συχνά μεταλλάσσονται κομμάτια των χρωμοσωμάτων-απογόνων. Για παράδειγμα, έστω ότι η πιθανότητα μετάλλαξης (Pm) των απογόνων ορίζεται ίση με 0.3. Με τη βοήθεια ομοιόμορφης κατανομής παράγεται ένας τυχαίος αριθμός rc στο διάστημα [0,1]. Αν ο αριθμός rc είναι μικρότερος ή ίσος από 0.3 τότε πραγματοποιείται μετάλλαξη στον απόγονο, διαφορετικά ο απόγονος προστίθεται στη δεξαμενή με τους άλλους μεταλλαγμένους απογόνους χωρίς εκείνος να μεταλλαχθεί. Εάν η πιθανότητα μετάλλαξης (Pm) είναι ίση με 1 τότε όλοι οι απόγονοι μεταλλάσσονται, ενώ εάν η πιθανότητα μετάλλαξης (Pm) είναι ίση με 0 τότε δεν μεταλλάσσεται κανένας απόγονος.

Σε πληθυσμούς με διαφορετική κωδικοποίηση, όπως η αναπαράσταση με ακέραιους αριθμούς, παίρνει τη μορφή της αντικατάστασης ενός ψηφίου με ένα άλλο, τυχαία επιλεγμένο από ένα κατάλληλο εύρος τιμών, με πιθανότητα p_m. Όμως, για συνδυαστικά προβλήματα βελτιστοποίησης, κάποιος τέτοιος τελεστής μπορεί να προκαλέσει προβλήματα με τη «νομιμότητα» των χρωμοσωμάτων, για παράδειγμα πολλαπλές επαναλήψεις μιας τιμής μέσα σε κάποιο άτομο μπορεί να μην είναι επιτρεπτή για κάποια προβλήματα. Εναλλακτικά έχουν χρησιμοποιηθεί τεχνικές μετάλλαξης βασισμένες στην εναλλαγή ζευγών ή την ολίσθηση θέσεων.

Επιπλέον, προσαρμοστικά συστήματα μετάλλαξης (adaptive mutation schemes) παρόμοια με μετάλλαξη στο πλαίσιο των στρατηγικών εξέλιξης αξίζει να αναφερθούν. Τα προσαρμοστικά συστήματα μετάλλαξης μεταβάλλουν τόσο το ποσοστό, όσο και τον τρόπο με τον οποίο γίνεται η μετάλλαξη κατά την διάρκεια της εκτέλεσης του ΓΑ. Για παράδειγμα, σε κάποια προβλήματα η μετάλλαξη ορίζεται με ένα τρόπο ώστε στην αρχή να γίνεται η εξερεύνηση ομοιόμορφα σε ολόκληρο το χώρο αναζήτησης και έπειτα η αναζήτηση να γίνεται πιο τοπική, ώστε να γίνονται βελτιώσεις στις υποψήφιες λύσεις.

Αν και υπάρχει κάποια σύγχυση για το ρόλο της μετάλλαξης, τόσο φυσικής όσο και τεχνητής, το σίγουρο είναι πως είναι απαραίτητη. Η μετάλλαξη λειτουργεί ως ασφαλιστική δικλείδα για τις περιπτώσεις, κατά τις οποίες η επιλογή και η διασταύρωση, ενδεχομένως, χάσουν κάποιες πολύτιμες γενετικές πληροφορίες. Όταν συμβαίνει, επιφέρει ποικιλία στον πληθυσμό, ανακατευθύνει την αναζήτηση και εξασφαλίζει ότι κανένα σημείο του χώρου αναζήτησης δεν αποκλείεται από τη διαδικασία του ψαξίματος.

Στη μετάλλαξη εφαρμόζονται διαδοχικά τέσσερις υποτελεστές: η ***πρόσθεση*** (add), η ***διαγραφή*** (omit), η ***αντικατάσταση*** (replace) και η ***ανταλλαγή*** (swap).

Στην ***πρόσθεση*** επιλέγεται ένα σημείο που δεν είναι στον απόγονο και προστίθεται σε αυτόν σε μία τυχαία επιλεγόμενη θέση. Έστω ότι ο απόγονος είναι το χρωμόσωμα:
X = [1, 3, 11 18, 16, 15, 13, 2, 5, 6, 12, 8, 10, 20, 21]

Επιλέγεται ένας πελάτης που δεν είναι στον απόγονο, έστω ο πελάτης **14** και προστίθεται σε μία τυχαία επιλεγμένη θέση. Ο μεταλλαγμένος απόγονος που προκύπτει είναι ο:
X' = [1, 3, 11 18, 16, 15, 13, 2, 5, **14**, 6, 12, 8, 10, 20, 21]

Στη ***διαγραφή***, επιλέγεται ένας τυχαίος πελάτης από τον απόγονο και διαγράφεται.
Έστω ότι ο απόγονος είναι το χρωμόσωμα:
X = [1, 3, 11, **18**, 16, 15, 13, 2, 5, 6, 12, 8, 10, 20, 21]
Επιλέγεται ένας πελάτης που είναι στον απόγονο, π.χ. το **18** και διαγράφεται από τον απόγονο. Ο μεταλλαγμένος απόγονος που προκύπτει είναι ο:
X' = [1, 3, 11, 16, 15, 13, 2, 5, 6, 12, 8, 10, 20, 21]

Στην ***αντικατάσταση***, εντοπίζεται ένας πελάτης που δεν είναι στον απόγονο. Έπειτα επιλέγεται ένας τυχαίος πελάτης από τον απόγονο και αντικαθίσταται με τον τυχαία επιλεγόμενο πελάτη που δεν είναι στον απόγονο. Έστω ο απόγονος:
X= [1, 3, 11, 18, 16, **15**, 13, 2, 5, 6, 12, 8, 10, 20, 21]
Επιλέγεται ένας πελάτης που δεν είναι στον απόγονο, π.χ. ο πελάτης **4**. Στη συνέχεια, επιλέγεται ένας πελάτης από τον απόγονο, π.χ. ο πελάτης **15** και αντικαθίσταται από τον τυχαία επιλεγμένο πελάτη που δεν είναι στον απόγονο, δηλαδή τον πελάτη **4**. Ο μεταλλαγμένος απόγονος που προκύπτει είναι ο:
X' = [1, 3, 11, 18, 16, **4**, 13, 2, 5, 6, 12, 8, 10, 20, 21]

Τέλος, στην ***ανταλλαγή***, επιλέγονται δύο τυχαίοι πελάτες στον απόγονο και ανταλλάσσονται αμοιβαία. Έστω ο απόγονος:
X= [1, 3, 11, 18, 16, **15**, 13, 2, 5, 6, **12**, 8, 10, 20, 21]
Επιλέγονται δύο τυχαίοι πελάτες στον απόγονο, π.χ. οι πελάτες **15** και **12** και ανταλλάσσονται αμοιβαία. Ο μεταλλαγμένος απόγονος που προκύπτει είναι ο:
X' = [1, 3, 11, 18, 16, **12**, 13, 2, 5, 6, **15**, 8, 10, 20, 21]

2.5. ΠΑΡΑΜΕΤΡΟΙ ΓΕΝΕΤΙΚΟΥ ΑΛΓΟΡΙΘΜΟΥ

- Πιθανότητα Διασταύρωσης (crossover rate)

Η πιθανότητα διασταύρωσης Pc συσχετίζεται με την 'επιθετικότητα' της έρευνας στον χώρο των πιθανών λύσεων. Όσο πιο υψηλή είναι η πιθανότητα διασταύρωσης τόσο πιο πολλοί απόγονοι δημιουργούνται και υπάρχει κίνδυνος να χαθούν αρκετά καλά χρωμοσώματα από τον τρέχοντα πληθυσμό. Η χαμηλή πιθανότητα διασταύρωσης συντηρεί τα καλά χρωμοσώματα από τη μία γενιά στην άλλη, με μία πιο συντηρητική διερεύνηση μεταξύ των υποψήφιων λύσεων.

- Πιθανότητα Μετάλλαξης (mutation rate)

Η πιθανότητα μετάλλαξης Pm μπορεί να είναι μεταβλητή ή σταθερή κατά την διάρκεια εκτέλεσης του αλγορίθμου. Στα πρώτα στάδια του γενετικού αλγορίθμου, ο μηχανισμός της διασταύρωσης είναι ο υπεύθυνος για την εξερεύνηση του συνόλου των λύσεων (exploration). Καθώς ο αλγόριθμος συγκλίνει, ο τελεστής διασταύρωσης γίνεται λιγότερο παραγωγικός, διότι τα χρωμοσώματα του πληθυσμού παρουσιάζουν

πολλές ομοιότητες μεταξύ τους και επομένως οποιοσδήποτε συνδυασμός τους δεν επιφέρει σημαντικές βελτιώσεις στους απογόνους που δημιουργούνται. Συνεπώς, κατά τη σύγκλιση του αλγορίθμου, η αύξηση της πιθανότητας μετάλλαξης επιτρέπει στον μηχανισμό της μετάλλαξης να αναζητά τοπικά, γύρω από μία αποδεδειγμένα καλή λύση (exploitation), νέους απογόνους και έτσι η εύρεση ακόμα καλύτερης λύσης είναι ακόμα δυνατή. Ο μηχανισμός της μετάλλαξης ουσιαστικά επιφέρει μικρές αλλαγές στα ήδη καλά χρωμοσώματα του πληθυσμού που δημιουργήθηκαν από τη διασταύρωση. Αυτός ο δυναμικός τρόπος χρήσης της πιθανότητας μετάλλαξης είναι αρκετά σύνθετος. Ακόμα, λόγω αρκετών άλλων παραμέτρων που μπορεί να εμπλέκονται στην λειτουργία του αλγορίθμου, όπως είναι το μέγεθος του πληθυσμού, το πλήθος των επαναλήψεων, ο τύπος διασταύρωσης, ο ελιτισμός κ.α., η εκδοχή της επιλογής σταθερής τιμής της πιθανότητας μετάλλαξης είναι ικανοποιητική για την αποδοτικότητα του αλγορίθμου. Δεδομένου πως μπορεί να γίνει και δοκιμαστικά έλεγχος της απόδοσης του αλγορίθμου για κάποιες τιμές της πιθανότητας μετάλλαξης, η μεταβλητή τιμή αυτής δεν είναι αναγκαία. Αξίζει να σημειωθεί όμως, πως μεγάλη τιμή της πιθανότητας μετάλλαξης αλλοιώνει τα υπάρχοντα χρωμοσώματα σε μεγάλο βαθμό και έτσι χάνονται κάποιες καλές λύσεις, ενώ πολύ μικρή τιμή της πιθανότητας μετάλλαξης δεν αφήνει τον αλγόριθμο να αναζητήσει νέες πιθανόν καλύτερες λύσεις.

- Μέγεθος Πληθυσμού (population size)

Το μέγεθος του πληθυσμού θα πρέπει να είναι τέτοιο ώστε ο γενετικός αλγόριθμος και να είναι γρήγορος αλλά και να εξερευνά ένα ικανοποιητικό τμήμα του χώρου αναζήτησης λύσεων. Εάν υπάρχουν υπερβολικά λίγα χρωμοσώματα στον πληθυσμό, ο γενετικός αλγόριθμος έχει λίγες πιθανότητες να εκτελέσει διασταύρωση και συνεπώς διερευνάται μόνο ένα μικρό τμήμα του χώρου αναζήτησης. Από την άλλη, εάν υπάρχουν υπερβολικά πολλά χρωμοσώματα στον πληθυσμό, ο αλγόριθμος επιβραδύνεται σημαντικά και γίνεται χρονοβόρος. Με δοκιμές διαφόρων μεγεθών για τον πληθυσμό μπορεί να βρεθεί το κατάλληλο μέγεθος αυτού, ώστε ο αλγόριθμος να είναι αποδοτικός αλλά συγχρόνως να μην καθυστερεί να συγκλίνει. Να σημειωθεί πως δεν τίθεται μόνο θέμα μνήμης για τον υπολογιστή όταν ορίζεται μεγάλο μέγεθος πληθυσμού, καθώς μπορεί ο αλγόριθμος να καθυστερεί σε διάφορους υπολογισμούς, συγκρίσεις ή ακόμα και εκτελέσεις δευτερευόντων ρουτινών οι οποίες ενδεχομένως καλούνται από τον αλγόριθμο (όπως είναι και ο υπολογισμός της συνάρτησης αξιολόγησης). Έτσι μπορεί να μην είναι γραμμικός (ως προς το μέγεθος του πληθυσμού) ο χρόνος εκτέλεσης του αλγορίθμου.

2.6 ΤΕΧΝΙΚΕΣ ΑΝΤΙΚΑΤΑΣΤΑΣΗΣ

Αφού έχει δημιουργηθεί μια νέα γενιά των απογόνων από τη διασταύρωση και τη μετάλλαξη, τίθεται το ερώτημα ποια από τις νέες υποψήφιες λύσεις πρέπει να γίνουν μέλη της επόμενης γενιάς. Στο πλαίσιο των στρατηγικών εξέλιξης το γεγονός αυτό προσδιορίζει τη διάρκεια ζωής των ατόμων και ουσιαστικά επηρεάζει τη

συμπεριφορά της σύγκλισης του αλγορίθμου. Οι ακόλουθες τεχνικές μπορούν να χρησιμοποιηθούν ως πιθανοί μηχανισμοί αντικατάστασης στους γενετικούς αλγόριθμους:

- *Αντικατάσταση γενιάς:*
  Ολόκληρος ο πληθυσμός αντικαθιστάται από τους απογόνους του. Με αυτόν τον τρόπο όλος ο πληθυσμός της προηγούμενης γενιάς χάνεται με αποτέλεσμα τυχόν καλές λύσεις να μην διατηρούνται στην επόμενη γενιά και ο αλγόριθμος δεν συγκλίνει σταθερά σε μία λύση. Έχει παρατηρηθεί πως από γενιά σε γενιά η λύση μπορεί να είναι είτε καλύτερη είτε χειρότερη και αυτό είναι καθαρά τυχαίο, ενώ μπορεί μία λύση που ήταν καλύτερη στην γενιά της, να χάνεται στην επόμενη αλλά να επανέρχεται σε μεταγενέστερη γενιά.
- *Ελιτισμός:*
  Όταν ο ΓΑ παρουσιάζει ελιτισμό, η καλύτερη λύση που έχει βρεθεί (ή οι N_{pop} καλύτερες λύσεις) δε χάνεται κατά την εξέλιξη του πληθυσμού από μια γενιά σε άλλη, παρά μόνο αν βρεθεί καλύτερη λύση. Συγκεκριμένα, συγκρίνονται οι απόγονοι με τους γονείς και διατηρούνται ως νέα χρωμοσώματα στον νέο πληθυσμό το καλύτερο ζευγάρι αυτών. Υπάρχει περίπτωση κάποιες φορές να εφαρμοστεί μετάλλαξη και στα χρωμοσώματα που διατηρούνται λόγω ελιτισμού, ώστε να αποφευχθεί η πρόωρη σύγκλιση. Αυτός ο μηχανισμός αντικατάστασης ονομάζεται «αδύναμος ελιτισμός».
- *Διαγραφή m τελευταίων :*
  Τα m πιο αδύναμα χρωμοσώματα αντικαθιστώνται από m απογόνους οι οποίοι παράγονται με τυχαίο τρόπο, όπως ακριβώς και ο αρχικός πληθυσμός. Με αυτόν τον τρόπο γίνεται ανανέωση του πληθυσμού χωρίς να χάνονται τα καλύτερα μέλη αυτού και συγχρόνως τα νέα άτομα που εισάγονται στον πληθυσμό δεν βασίζονται στα προϋπάρχοντα, οπότε μπορεί να γίνεται καλύτερη αναζήτηση στον χώρο λύσεων. Εάν ισχύει ότι $m \ll N_{pop}$ τότε έχουμε την περίπτωση ενός συστήματος αντικατάστασης σταθερής κατάστασης.
- *Διαγραφή m ατόμων:*
  Σε αντίθεση με την τεχνική της διαγραφής m τελευταίων, σε αυτήν δεν αντικαθιστώνται τα m πιο αδύναμα χρωμοσώματα, αλλά m αυθαίρετα επιλεγμένα άτομα από την παλιά γενιά, το οποίο από τη μία πλευρά μειώνει την ταχύτητα σύγκλισης του αλγόριθμου αλλά από την άλλη πλευρά βοηθά να αποφευχθεί η πρόωρη σύγκλιση του σε τοπικά ακρότατα.

ΣΥΓΚΛΙΣΗ

Ο ΓΑ τερματίζεται είτε όταν ο αριθμός των γενεών έχει φτάσει το μέγιστο αριθμό γενεών που έχει θέσει ο χρήστης, ή όταν η καλύτερη λύση δεν έχει αλλάξει για έναν προκαθορισμένο αριθμό γενεών.

2.7 ΤΕΧΝΙΚΕΣ ΚΩΔΙΚΟΠΟΙΗΣΗΣ ΤΩΝ ΓΕΝΕΤΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ

Οι τεχνικές κωδικοποίησης χρησιμοποιούνται για να μετασχηματίσουν πιθανές λύσεις του προβλήματος σε χρωμοσώματα και εξαρτώνται κάθε φορά από το συγκεκριμένο πρόβλημα. Κάποιες από τις συνήθης κωδικοποιήσεις που χρησιμοποιούνται είναι οι ακόλουθες:

- Δυαδική κωδικοποίηση
  Είναι η πιο συνηθισμένη μορφή κωδικοποίησης , κατά την οποία τα δεδομένα του προβλήματος μετασχηματίζονται σε ακολουθίες από 0 και 1. Η δυαδική κωδικοποίηση μπορεί να δώσει ένα μεγάλο αριθμό χρωμοσωμάτων με σχετικά μικρό αριθμό ψηφίων σε αυτά.
- Κωδικοποίηση αντιμετάθεσης
  Η κωδικοποίηση αυτή είναι κατάλληλη σε προβλήματα ταξινόμησης ή ακολουθιακά. Τα χρωμοσώματα αποτελούνται από ακέραιους τοποθετημένους σε ακολουθία. Το πρόβλημα του περιπλανώμενου πωλητή είναι ένα κλασσικό παράδειγμα χρήσης της κωδικοποίησης αυτής.
- Κωδικοποίηση τιμής
  Η κωδικοποίηση αυτή χρησιμοποιεί ακέραιους, πραγματικούς ακόμη και χαρακτήρες για να σχηματιστούν τα χρωμοσώματα.
- Κωδικοποίηση δέντρου
  Είναι η καλύτερη κωδικοποίηση για την αποτίμηση εκφράσεων και προγραμμάτων, όπως ο γενετικός προγραμματισμός. Σε αυτήν κάθε χρωμόσωμα είναι ένα δέντρο ορισμένων αντικειμένων, συναρτήσεων ή εντολών που συναντάμε στις γλώσσες προγραμματισμού. Η γλώσσα προγραμματισμού LISP χρησιμοποιείται για αυτό το σκοπό, καθώς οι εφαρμογές της μπορούν εύκολα να αναπαρασταθούν και να διαχειριστούν από τους τελεστές της διασταύρωσης και της μετάλλαξης.

  Δεν υπάρχουν σαφής κανόνες και οδηγίες για το τι τρόπος κωδικοποίησης θα ακολουθηθεί σε κάθε πρόβλημα καθώς αυτό εξαρτάται από τις απαιτήσεις κάθε εφαρμογής και πόσο αποδοτική μπορεί να είναι η κάθε μέθοδος στην συγκεκριμένη περίπτωση.

2.8 ΓΕΝΕΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ ΚΑΙ ΑΛΛΕΣ ΜΟΡΦΕΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ

Η αρχή πίσω από την υλοποίηση ενός ΓΑ είναι απλή: μιμείται τη γενετική και τη φυσική επιλογή μέσω ενός προγράμματος. Οι παράμετροι του προβλήματος κωδικοποιούνται σαν μια γραμμική δομή δεδομένων παρόμοια με μια αλυσίδα DNA,

ένα διάνυσμα ή συμβολοσειρά. Αρκετές φορές όταν το πρόβλημα είναι πολυδιάστατο μπορεί να γίνει και χρήση πίνακα.

Σε αντίθεση με τις άλλες τεχνικές βελτιστοποίησης η αντικειμενική συνάρτηση μπορεί να είναι οτιδήποτε μπορεί να υπολογιστεί από ένα κομπιούτερ. Οπότε δεν υπάρχουν σαφής μαθηματικοί περιορισμοί για τις ιδιότητες τις οποίες πρέπει να έχει η αντικειμενική συνάρτηση.

Τα κύρια κριτήρια που χρησιμοποιούνται για την ταξινόμηση των αλγορίθμων βελτιστοποίησης είναι οι εξής : να είναι συνεχείς/ διακριτοί , με περιορισμούς / χωρίς περιορισμούς και ακολουθιακοί/ παράλληλοι . Υπάρχει σαφής διαφορά μεταξύ διακριτών και συνεχών προβλημάτων και το πώς αυτά επιλύονται. Ως εκ τούτου, είναι εποικοδομητικό να παρατηρήσουμε ότι οι συνεχείς μέθοδοι χρησιμοποιούνται μερικές φορές για την επίλυση εγγενώς διακριτών προβλημάτων και αντίστροφα. Οι παράλληλοι αλγόριθμοι που χρησιμοποιούνται συνήθως για την επιτάχυνση της επεξεργασίας καθώς, σε αρκετές περιπτώσεις είναι πιο αποτελεσματικό για να τρέχουν σε αρκετούς επεξεργαστές παράλληλα και όχι διαδοχικά . Οι περιπτώσεις αυτές περιλαμβάνουν μεταξύ άλλων αυτές στις οποίες υπάρχει μεγάλη πιθανότητα κάθε μεμονωμένη εκτέλεση αναζήτησης να κολλήσει σε ένα τοπικό ακρότατο.

Ανεξάρτητα από την παραπάνω ταξινόμηση, οι μέθοδοι βελτιστοποίησης μπορεί επίσης να διακριθούν σε ντετερμινιστικές και μη ντετερμινιστικές μεθόδους. Επιπλέον οι αλγόριθμοι βελτιστοποίησης μπορούν να ταξινομηθούν ως τοπικοί ή ολικοί. Χρησιμοποιώντας τους όρους της ενέργειας και εντροπίας, η τοπική αναζήτηση αντιστοιχεί στην εντροπία ενώ η ολική βελτιστοποίηση εξαρτάται ουσιαστικά από την καταλληλότητα, δηλαδή το ενεργειακό τοπίο.

Οι κύριες διαφορές των γενετικών αλγόριθμων από τις προαναφερθείσες τεχνικές βελτιστοποίησης είναι οι εξής:

1. Οι ΓΑ λειτουργούν με κωδικοποιημένες εκδοχές των παραμέτρων του προβλήματος, παρά με τις παραμέτρους αυτές καθ’ αυτές. Δηλαδή οι ΓΑ δουλεύουν με ένα κωδικοποιημένο σετ λύσεων και όχι με την ίδια τη λύση του προβλήματος.

2. Σχεδόν όλες οι συμβατικές τεχνικές βελτιστοποίησης αναζητούν τη λύση ψάχνοντας από ένα μόνο σημείο , ενώ οι ΓΑ πάντα λειτουργούν σε ολόκληρο τον πληθυσμό σημείων. Αυτό παίζει σημαντικό ρόλο για την ευρωστία των γενετικών αλγορίθμων, καθώς βελτιώνει τις πιθανότητες για την εύρεση του ολικού μεγίστου/ελάχιστου και επίσης βοηθά στην αποφυγή του εγκλωβισμού της διαδικασίας σε τοπικά ακρότατα.

3. Οι ΓΑ χρησιμοποιούν την συνάρτηση καταλληλότητας για τον υπολογισμό της λύσης και όχι παράγωγους. Ως αποτέλεσμα, μπορούν να εφαρμοστούν σε οποιοδήποτε είδος πρόβλημα βελτιστοποίησης, συνεχούς ή διακριτού. Το βασικό σημείο στο οποίο πρέπει να δοθεί σημασία είναι να προσδιορισθεί και να καθορισθεί μια κατάλληλη διαδικασία κωδικοποίησης του προβλήματος .

4. Οι ΓΑ λειτουργούν χρησιμοποιώντας πιθανοτικές μεταβάσεις, ενώ στις συμβατικές μεθόδους για συνεχή βελτιστοποίηση οι μεταβάσεις γίνονται ντετερμινιστικά, δηλαδή, οι ΓΑ δεν χρησιμοποιούν ντετερμινιστικούς κανόνες .

2.9 ΠΛΕΟΝΕΚΤΗΜΑΤΑ-ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΩΝ Γ.Α.

ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΓΕΝΕΤΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ

1) Μπορούν να επιλύουν δύσκολα προβλήματα γρήγορα και αξιόπιστα. Ένας από τους σημαντικούς λόγους χρήσης των Γ.Α. είναι η μεγάλη αποδοτικότητά τους. Τόσο η θεωρία, όσο και η πράξη έχουν δείξει ότι προβλήματα που έχουν πολλές δύσκολα προσδιορισμένες λύσεις μπορούν να αντιμετωπιστούν καλύτερα από Γ.Α. είναι δε αξιοσημείωτο ότι συναρτήσεις που παρουσιάζουν μεγάλες διακυμάνσεις και καθιστούν ανεπαρκείς άλλες μεθόδους στην εύρεση των ακροτάτων τους, για τους Γ.Α. δεν αποτελούν σημεία δυσχέρειας.
2) Μπορούν εύκολα να συνεργαστούν με τα υπάρχοντα μοντέλα και συστήματα. Οι Γ.Α. προσφέρουν το σημαντικό πλεονέκτημα της χρήσης τους με προσθετικό τρόπο στα μοντέλα που χρησιμοποιούνται σήμερα, μη απαιτώντας την επανασχεδίασή τους. Μπορούν εύκολα να συνεργαστούν με τον υπάρχοντα κώδικα, χωρίς μεγάλο κόπο. Αυτό συμβαίνει, διότι χρησιμοποιούν μόνο πληροφορίες της διαδικασίας ή συνάρτησης που πρόκειται να βελτιστοποιήσουν, δίχως να ενδιαφέρει άμεσα ο ρόλος της μέσα στο σύστημα ή η όλη δομή του συστήματος.
3) Είναι εύκολα επεκτάσιμοι και εξελίξιμοι. Οι Γ.Α. δεν αντιστέκονται σε αλλαγές, επεκτάσεις και μετεξελίξεις, ανάλογα με την κρίση του σχεδιαστή. Σε πολλές εφαρμογές, έχουν αναφερθεί λειτουργίες των Γ.Α. που δεν είναι δανεισμένες από τη φύση ή που έχουν υποστεί σημαντικές αλλαγές, πάντα προς όφελος της απόδοσης. Παραλλαγές στο βασικό σχήμα δεν είναι απλά αναγκαίες, αλλά σε ορισμένες περιπτώσεις επιβάλλονται.
4) Μπορούν να συμμετέχουν σε υβριδικές μορφές με άλλες μεθόδους. Αν και η ισχύς των Γ.Α. είναι μεγάλη, σε μερικές ειδικές περιπτώσεις προβλημάτων, όπου άλλες μέθοδοι συμβαίνει να έχουν πολύ υψηλή αποδοτικότητα, λόγω εξειδίκευσης, υπάρχει η δυνατότητα χρησιμοποίησης ενός υβριδικού σχήματος Γ.Α. με άλλη μέθοδο. Αυτό είναι αποτέλεσμα της μεγάλης ευελιξίας των Γ.Α.
5) Εφαρμόζονται σε πολύ περισσότερα πεδία από κάθε άλλη μέθοδο. Το χαρακτηριστικό που τους εξασφαλίζει αυτό το πλεονέκτημα είναι η ελευθερία επιλογής των κριτηρίων που καθορίζουν την επιλογή μέσα στο τεχνικό περιβάλλον. Έτσι, Γ.Α. μπορούν να χρησιμοποιηθούν στην οικονομία, στο

σχεδιασμό μηχανών, στην επίλυση μαθηματικών εξισώσεων, στην εκπαίδευση Νευρωνικών Δικτύων και σε πολλούς άλλους τομείς.

6) Δεν απαιτούν περιορισμούς στις συναρτήσεις που επεξεργάζονται. Ο κύριος λόγος που καθιστά τις παραδοσιακές μεθόδους δύσκαμπτες και ακατάλληλες για πολλά προβλήματα είναι η απαίτησή τους για ύπαρξη περιορισμών, όπως ύπαρξη παραγώγων, συνέχεια, όχι «θορυβώδεις» συναρτήσεις κ.τ.λ. Τέτοιου είδους ιδιότητες είναι αδιάφορες για τους Γ.Α. πράγμα που τους κάνει κατάλληλους για μεγάλο φάσμα προβλημάτων.
7) Δεν ενδιαφέρει η σημασία της υπό εξέτασης πληροφορίας. Η μόνη «επικοινωνία» του Γ.Α. με το περιβάλλον του είναι η αντικειμενική συνάρτηση. Αυτό εγγυάται την επιτυχία του ανεξάρτητα από την σημασία του προβλήματος. Βέβαια, δεν σημαίνει ότι δεν υπάρχουν άλυτα προβλήματα για τους Γ.Α. Όπου όμως δεν τα καταφέρνουν, η αιτία είναι η φύση του χώρου που ερευνούν και όχι το πληροφοριακό περιεχόμενο του προβλήματος.
8) Έχουν από τη φύση τους το στοιχείο του παραλληλισμού. Οι Γ.Α. σε κάθε τους βήμα επεξεργάζονται μεγάλες ποσότητες πληροφορίας, αφού κάθε άτομο θεωρείται αντιπρόσωπος πολλών άλλων. Έχει υπολογιστεί ότι η αναλογία αυτή είναι της τάξεως του n^3 δηλαδή 10 άτομα αντιπροσωπεύουν περίπου 1000. Είναι, λοιπόν, προφανές ότι μπορούν να καλύψουν με αποδοτικό ψάξιμο μεγάλους χώρους σε μικρούς χρόνους.
9) Είναι μία μέθοδος που κάνει ταυτόχρονα εξερεύνηση του χώρου αναζήτησης και εκμετάλλευση της ήδη επεξεργασμένης πληροφορίας. Ο συνδυασμός αυτός σπάνια συναντάται σε οποιαδήποτε άλλη μέθοδο. Με το τυχαίο ψάξιμο γίνεται καλή εξερεύνηση του χώρου, αλλά δεν γίνεται εκμετάλλευση της πληροφορίας. Αντίθετα, με την ανάβαση-λόφου (*hill-climbing*) γίνεται εκμετάλλευση της πληροφορίας, αλλά όχι καλή εξερεύνηση. Συνήθως τα δύο αυτά χαρακτηριστικά είναι ανταγωνιστικά και το επιθυμητό είναι να συνυπάρχουν και τα δύο προς όφελος της όλης διαδικασίας. Οι Γ.Α. επιτυγχάνουν το βέλτιστο συνδυασμό εξερεύνησης και εκμετάλλευσης, πράγμα που τους κάνει ιδιαίτερα αποδοτικούς και ελκυστικούς.
10) Επιδέχονται παράλληλη υλοποίηση. Οι Γ.Α. μπορούν να εκμεταλλευτούν τα πλεονεκτήματα των παράλληλων μηχανών, αφού λόγω της φύσης τους, εύκολα μπορούν να δεχτούν παράλληλη υλοποίηση. Το χαρακτηριστικό αυτό αυξάνει ακόμη περισσότερο την απόδοσή τους, ενώ σπάνια συναντάται σε ανταγωνιστικές μεθόδους.

ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΩΝ ΓΕΝΕΤΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ

Παρά τη μεγάλη τους χρησιμότητα σε πολλές εφαρμογές της καθημερινής ζωής, οι ΓΑ έχουν και κάποια σοβαρά μειονεκτήματα, τα οποία θα μπορούσαν να σταθούν εμπόδιο στην ολοένα και μεγαλύτερη εξάπλωση αυτής της τεχνολογίας. Τα σημαντικότερα από αυτά είναι τα εξής:

1) Προβλήματα εξοικείωσης με τη Γενετική. Για την κατανόηση των ΓΑ δεν απαιτούνται γνώσεις γενετικής ή Βιολογίας. Οι ΓΑ μιμούνται με αφαιρετικό τρόπο κάποιες διαδικασίες που παρατηρούνται στη φύση, χωρίς να ενδιαφέρει σε μεγάλο βαθμό λεπτομέρειας η λειτουργία τους και χωρίς να είναι απαραίτητο το γνωστικό υπόβαθρο που έχουν οι βιολόγοι για να μελετήσουν αυτά τα φαινόμενα. Οι όροι είναι δανεισμένοι από τη Βιολογία με σκοπό την καλύτερη εισαγωγή και κατανόηση του θέματος κι όχι την παραπομπή του μελετητή στα άγνωστα πεδία μίας ξένης επιστήμης. Επιπλέον, η εξέλιξη των ΓΑ δεν είναι συνυφασμένη με την αντίστοιχη εξέλιξη των θεωριών της Βιολογία. Το αρχικό μοντέλο είναι δανεισμένο από εκεί, όμως η εφαρμογή στα Τεχνικά Συστήματα έγινε με πλήθος διαφοροποιήσεων και προσαρμόσεων, με στόχο πάντα τη βελτίωση της απόδοσης.

2) Το πρόβλημα του χρόνου. Ο μεγάλος αριθμός αξιολογήσεων καταλληλότητας όπως κι η τυχόν πολυπλοκότητα της συνάρτησης αποτίμησης, ακριβώς επειδή αυτή υπολογίζεται για κάθε χρωμόσωμα κάθε γενιάς, συνεπάγεται σημαντικό υπολογιστικό χρόνο.

3) Η αδυναμία παροχής εγγυήσεων εύρεσης της βέλτιστης λύσης. Αυτό σημαίνει ότι τις περισσότερες φορές δεν υπάρχει βεβαιότητα ότι ο ΓΑ έδωσε την καλύτερη δυνατή λύση στο εκάστοτε πρόβλημα.

2.10 ΕΦΑΡΜΟΓΕΣ ΤΩΝ ΓΕΝΕΤΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ

Οι γενετικοί αλγόριθμοι έχουν αποδειχθεί ικανοί να επιλύουν πολλά αρκετά μεγάλα και πολύπλοκα προβλήματα, στα οποία άλλες μέθοδοι αντιμετωπίζουν δυσκολίες. Παραδείγματα είναι μεγάλης κλίμακας συνδυαστικά προβλήματα βελτιστοποίησης και εκτιμήσεις πραγματικών παραμέτρων μέσα σε αρκετά περίπλοκους χώρους αναζήτησης που παρουσιάζουν πολλά τοπικά ακρότατα. Αυτός είναι και ο λόγος που όλο περισσότεροι επιστήμονες και μηχανικοί τους χρησιμοποιούν για την επίλυση προβλημάτων.

Παρουσιάζονται στη συνέχεια μερικές αντιπροσωπευτικές εφαρμογές των γενετικών αλγορίθμων.

I. Εύρεση μέγιστης τιμής αριθμητικών συναρτήσεων

Πρόκειται για την πιο καλά μελετημένη εφαρμογή των γενετικών αλγορίθμων. Η εύρεση του μέγιστου μιας συνάρτησης δεν είναι καθόλου εύκολη υπόθεση για συναρτήσεις πολλών μεταβλητών, οι οποίες εμφανίζουν ασυνέχειες, θόρυβο, κλπ. Το πλεονέκτημα που εμφανίζει η εφαρμογή τους σε αυτά τα προβλήματα είναι ότι η συνάρτηση καταλληλότητας είναι δεδομένη.

II. Επεξεργασία εικόνων

Οι γενετικοί αλγόριθμοι χρησιμοποιούνται για την αναγνώριση προτύπων, όπως ακμές, επιφάνειες, ακόμη και αντικείμενα, σε ψηφιοποιημένες εικόνες. Το αποτέλεσμα αυτής της επεξεργασίας μπορεί να αποτελέσει τη βάση για την ψηφιακή όραση.

III. Σχεδίαση

Οι γενετικοί αλγόριθμοι μπορούν να χρησιμοποιηθούν στη σχεδίαση κατασκευών και εξαρτημάτων, όπως π.χ. γέφυρες, μηχανολογικά εξαρτήματα, όπου ζητούμενο μπορεί να είναι τόσο η εύρεση μιας λύσης, όσο και η βελτιστοποίηση της. Οι αλγόριθμοι μπορούν να δοκιμάσουν συνδυασμούς και ιδέες που ο ανθρώπινος νους δε θα δοκίμαζε ποτέ, δίνοντας ενίοτε πρωτότυπα αποτελέσματα.

IV. Μηχανική μάθηση

Στα συστήματα μηχανικής μάθησης οι γενετικοί αλγόριθμοι μπορούν να χρησιμοποιηθούν για την ανακάλυψη κανόνων if…then... Η πιο γνωστή εφαρμογή είναι αυτή των *συστημάτων κατηγοριοποίησης (classified systems),* ωστόσο οι γενετικοί αλγόριθμοι έχουν χρησιμοποιηθεί και σε παιχνίδια, επίλυση λαβυρίνθων, καθώς και για πολιτικές και οικονομικές αναλύσεις.

V. Συνδυαστική βελτιστοποίηση

Πρόκειται για το κλασσικό πρόβλημα κατανομής πόρων σε δραστηριότητες, με σκοπό τη μεγιστοποίηση του οφέλους ή την ελάττωση του κόστους. Τα προβλήματα αυτής της κατηγορίας παρουσιάζουν συνδυαστική έκρηξη του χώρου αναζήτησης, ως προς το μέγεθος του προβλήματος, με αποτέλεσμα ο έλεγχος όλων των υποψήφιων λύσεων να είναι αδύνατος. Το πιο γνωστό πρόβλημα αυτής της κατηγορίας είναι αυτό του *πλανόδιου πωλητή,* όπου στόχος είναι η εύρεση της συντομότερης διαδρομής για την επίσκεψη ενός συνόλου πόλεων.

Οι γενετικοί αλγόριθμοι μπορούν να δώσουν σε αυτό το πρόβλημα πολλές λύσεις κοντά στη βέλτιστη. Ένα άλλο πρόβλημα είναι η *αποθήκευση κιβωτίων (bin packing)* και αφορά την εύρεση του βέλτιστου τρόπου αποθήκευσης ενός αριθμού κιβωτίων σε περιορισμένο χώρο και έχει μεγάλη πρακτική σημασία στη βιομηχανία.

Τέλος στην κατηγορία αυτών των εφαρμογών εντάσσονται και τα προβλήματα *καταμερισμού – χρονοπρογραμματισμού εργασιών (Job shop & Flow shop scheduling).*

Γίνεται φανερό λοιπόν ότι οι γενετικοί αλγόριθμοι έχουν εφαρμοστεί σε διάφορα προβλήματα της Τεχνητής Νοημοσύνης και ιδιαίτερα σε προβλήματα βελτιστοποίησης. Σε ορισμένα προβλήματα τα αποτελέσματα ήταν πολύ καλά ενώ σε άλλα αρκετά απογοητευτικά.

Συνοπτικά οι γενετικοί αλγόριθμοι χρησιμοποιούνται στα παρακάτω:

- Επεξεργασία εικόνων
- Ψηφιακά συστήματα VLSI
- Πρόβλεψη των τρισδιάστατων δομών των πρωτεϊνών
- Τεχνολογίες laser
- Ιατρική
- Ανάλυση χρονοσειρών
- Τροχιές διαστημικών σκαφών
- Αεροναυπηγική
- Τεχνολογία υγρών κρυστάλλων
- Φυσική στερεάς κατάστασης
- Ρομποτική
- Δίκτυα ύδρευσης
- Έλεγχο
- Τεχνολογία λογισμικού
- Σχεδιασμός συστημάτων τεχνικής νοημοσύνης.

ΚΕΦΑΛΑΙΟ 3
ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΣΤΟ ΠΟΛΛΑΠΛΟ ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ

3.1 ΕΙΣΑΓΩΓΗ- ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Κατά αντιστοιχία με το απλό γραμμικό μοντέλο υπάρχει και το πολλαπλό γραμμικό μοντέλο, όπου μία τυχαία μεταβλητή (η οποία θα ονομάζεται και εξαρτημένη μεταβλητή για το γραμμικό μοντέλο) εξαρτάται γραμμικά από κάποιες ανεξάρτητες τυχαίες μεταβλητές. Θα ορίσουμε τον τύπο του πολλαπλού γραμμικού μοντέλου παρακάτω και θα θέσουμε το ερώτημα πόσες και ποιες θα είναι οι μεταβλητές που εμπλέκονται σε αυτό. Στην παρούσα εργασία θα μας απασχολήσει το πρόβλημα επιλογής μεταβλητών στο πολλαπλό γραμμικό μοντέλο και θα επιλυθεί με την βοήθεια γενετικού αλγορίθμου.

3.2 ΠΟΛΛΑΠΛΟ ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ

Το πολλαπλό γραμμικό μοντέλο χρησιμοποιείται όταν εμπλέκονται περισσότερες από δύο τυχαίες μεταβλητές, εκ των οποίων η μία θεωρείται εξαρτημένη, ποσοτική τυχαία μεταβλητή και συμβολίζεται με Υ, και όλες οι άλλες τυχαίες μεταβλητές, έστω p στο πλήθος, $\boldsymbol{X}=\left(\mathrm{X}_1,\mathrm{X}_2,...,\mathrm{X}_p\right)$, είναι επίσης ποσοτικές τυχαίες μεταβλητές και θεωρούνται πως επηρεάζουν (γραμμικά) την εξαρτημένη τυχαία μεταβλητή. Η σχέση που δηλώνει την γραμμική σχέση της εξαρτημένης τυχαίας μεταβλητής με τις ανεξάρτητες τυχαίες μεταβλητές του πολλαπλού γραμμικού μοντέλου είναι η:

$$\boldsymbol{Y}=\boldsymbol{a}+\boldsymbol{b}_1\boldsymbol{X}_1+\boldsymbol{b}_2\boldsymbol{X}_2+...+\boldsymbol{b}_p\boldsymbol{X}_p+\boldsymbol{\varepsilon},\quad \boldsymbol{\varepsilon}\sim\boldsymbol{N}(0,\boldsymbol{\sigma}^2)$$

με την αναμενόμενη τιμή της εξαρτημένης τυχαίας μεταβλητής Υ να είναι :

$$\boldsymbol{E}\left[\boldsymbol{Y}\mid\boldsymbol{X}_1,\boldsymbol{X}_2,...,\boldsymbol{X}_p\right]=\boldsymbol{a}+\boldsymbol{b}_1\boldsymbol{X}_1+\boldsymbol{b}_2\boldsymbol{X}_2+...+\boldsymbol{b}_p\boldsymbol{X}_p.$$

Αντιστοίχως, για το τυχαίο δείγμα n παρατηρήσεων των p+1 τυχαίων μεταβλητών $\left(\mathrm{Y}_1,\mathrm{X}_{11},\mathrm{X}_{21},...,\mathrm{X}_{p1}\right),\left(\mathrm{Y}_2,\mathrm{X}_{12},\mathrm{X}_{22},...,\mathrm{X}_{p2}\right),...,\left(\mathrm{Y}_3,\mathrm{X}_{13},\mathrm{X}_{23},...,\mathrm{X}_{p3}\right)$ ισχύει η σχέση:

$$\boldsymbol{Y}_i=\boldsymbol{a}+\boldsymbol{b}_1\boldsymbol{X}_{1i}+\boldsymbol{b}_2\boldsymbol{X}_{2i}+...+\boldsymbol{b}_p\boldsymbol{X}_{pi}+\boldsymbol{\varepsilon}_i,\quad \boldsymbol{\varepsilon}_i\sim\boldsymbol{N}(0,\boldsymbol{\sigma}^2),\ i=1,2,...,n$$

όπου $\boldsymbol{\varepsilon}_i$ είναι ανεξάρτητες τυχαίες μεταβλητές και n είναι το πλήθος των παρατηρήσεων.

Οι ποσότητες $\boldsymbol{a}$ και $\boldsymbol{b}_1,\boldsymbol{b}_2,...,\boldsymbol{b}_p$ είναι πραγματικοί αριθμοί, συντελεστές των τυχαίων μεταβλητών για το γραμμικό μοντέλο. Η $\boldsymbol{b}_j,\ \boldsymbol{j}=1,...,\boldsymbol{p}$ δηλώνει κατά πόσο μεταβάλλεται η αναμενόμενη τιμή της τυχαίας μεταβλητής Υ κατά την αύξηση μίας μονάδας της τυχαίας μεταβλητής $\boldsymbol{X}_j$, όταν οι τιμές των άλλων τυχαίων μεταβλητών παραμένουν σταθερές. Η σταθερά $\boldsymbol{a}$ εκφράζει την μέση τιμή της τυχαίας μεταβλητής Υ όταν όλες οι τυχαίες ανεξάρτητες μεταβλητές $\mathrm{X}_1,\mathrm{X}_2,...,\mathrm{X}_p$ λαμβάνουν την τιμή

μηδέν. Γενικότερα οι τιμές των παραμέτρων $\boldsymbol{a}$ και $\boldsymbol{b_1, b_2, ..., b_p}$ του πολλαπλού γραμμικού μοντέλου είναι άγνωστες, και μπορούν να εκτιμηθούν από το διατεταγμένο δείγμα παρατηρήσεων, εφαρμόζοντας την μέθοδο των ελαχίστων τετραγώνων, και οι εκτιμήσεις συμβολίζονται με $\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2, ..., \hat{\boldsymbol{b}}_p$. Η αντίστοιχη τιμή που προκύπτει ως εκτίμηση της εξαρτημένης μεταβλητής Y είναι: $\hat{\boldsymbol{y}}_i = \hat{\boldsymbol{a}} + \hat{\boldsymbol{b}}_1 \boldsymbol{x}_{1i} + \hat{\boldsymbol{b}}_2 \boldsymbol{x}_{2i} + ... + \hat{\boldsymbol{b}}_p \boldsymbol{x}_{pi}$ ενώ η διαφορά $\boldsymbol{y} - \hat{\boldsymbol{y}}_i = \hat{\boldsymbol{\varepsilon}}_i$ είναι τα κατάλοιπα για το γραμμικό μοντέλο των οποίων το άθροισμα των τετραγώνων τους, για όλες τις παρατηρήσεις, συμβολίζεται με SSE.

3.3 ΠΡΟΒΛΗΜΑ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

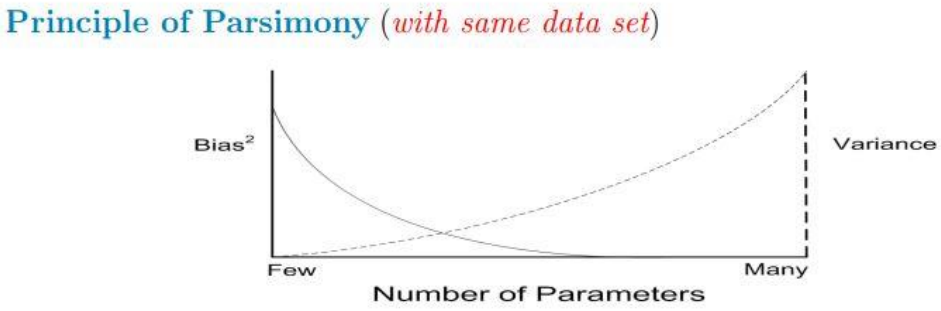
Όταν προσπαθούμε να ορίσουμε ένα πρόβλημα παλινδρόμησης για πολλαπλό γραμμικό μοντέλο, ουσιαστικά προσπαθούμε να βρούμε την ακριβή γραμμική σχέση εξάρτησης της μεταβλητής Y με τις p ανεξάρτητες μεταβλητές. Η σχέση αυτή προφανώς διαφαίνεται μέσα από τις παρατηρήσεις που διαθέτουμε για τις p+1 τυχαίες μεταβλητές. Το γραμμικό μοντέλο όμως που ερευνούμε κάθε φορά, δεδομένων των παρατηρήσεων, δίνει αποτελέσματα για την τιμή της παραμέτρου b_i, η οποία δηλώνει και την επίδραση της ανεξάρτητης μεταβλητής X_i, καθώς και αν η τιμή αυτή είναι στατιστικά σημαντική. Επίσης, για κάθε γραμμικό μοντέλο διαθέτουμε κριτήρια τα οποία μετρούν κατά κάποιον τρόπο αν είναι κάποιο γραμμικό μοντέλο καλύτερο από κάποιο άλλο. Ένα σημαντικότατο ζήτημα είναι, λοιπόν, όταν διαθέτουμε ένα συγκεκριμένο πλήθος ανεξάρτητων μεταβλητών, αν θα πρέπει να συμπεριλάβουμε όλες αυτές τις μεταβλητές στο γραμμικό μοντέλο ή κάποιο υποσύνολο αυτών. Και αν πρέπει να συμπεριλάβουμε κάποιο γνήσιο υποσύνολο των ανεξάρτητων μεταβλητών, τίθεται το ερώτημα: ποιες από αυτές και πόσες πρέπει να συμπεριληφθούν στο γραμμικό μοντέλο;

Είναι πολύ μεγάλο το πλήθος των δυνατών γραμμικών μοντέλων που μπορεί να προκύψει αν αναζητήσουμε όλους τους συνδυασμούς των τυχαίων επεξηγηματικών μεταβλητών που μπορούν να συμπεριληφθούν σε κάθε γραμμικό μοντέλο. Συγκεκριμένα, μπορεί κάποια τυχαία επεξηγηματική μεταβλητή να συμπεριληφθεί σε ένα γραμμικό μοντέλο ή να μην συμπεριληφθεί. Αυτό σημαίνει πως για το πλήθος των γραμμικών μοντέλων χωρίς την συγκεκριμένη τυχαία μεταβλητή έχουμε άλλα τόσα γραμμικά μοντέλα αν συμπεριλάβουμε και αυτήν. Γενικά για κάθε τυχαία μεταβλητή που έχουμε καθώς αυτή μπορεί να συμπεριλαμβάνεται ή όχι στα γραμμικά μοντέλα, έχουμε διπλάσιο πληθυσμό γραμμικών μοντέλων που προκύπτουν. Για p τυχαίες μεταβλητές το πλήθος των γραμμικών μοντέλων που προκύπτουν είναι 2^p και αυτό σημαίνει πως για προβλήματα γραμμικής παλινδρόμησης όπου υπάρχει μεγάλο πλήθος τυχαίων μεταβλητών είναι πολύ μεγάλος ο αριθμός των γραμμικών μοντέλων που υπάρχουν, καθώς το πλήθος αυτών μεγαλώνει εκθετικά σε σχέση με το πλήθος των τυχαίων μεταβλητών. Ενδεικτικά σημειώνουμε πως για 10 μεταβλητές έχουμε 2^{10}=1024 γραμμικά μοντέλα, ή για 30 μεταβλητές έχουμε 2^{30}=1.073.741.824 γραμμικά μοντέλα, ή όπως τα δεδομένα που θα χρησιμοποιήσουμε στο τελευταίο κεφάλαιο όπου έχουμε **56 μεταβλητές** προκύπτουν $\mathbf{2^{56} \approx 7{,}2 \cdot 10^{16}}$ γραμμικά μοντέλα.

Γενικά υπάρχουν διάφορα προβλήματα που μπορεί να προκύψουν από το μεγάλο πλήθος τυχαίων μεταβλητών σε ένα γραμμικό μοντέλο, όπως είναι η πολυσυγραμμικότητα, όπου μπορεί για κάποιο υποσύνολο των τυχαίων μεταβλητών X_i να υπάρχει μία εξ αυτών η οποία να εξαρτάται από δύο ή περισσότερες άλλες τυχαίες μεταβλητές. Αυτό είναι πιθανό στην πράξη καθώς ποτέ σχεδόν δεν ισχύουν απολύτως οι προϋποθέσεις του γραμμικού μοντέλου. Οπότε είναι καλό να αποφεύγεται να εισέρχονται πολλές τυχαίες μεταβλητές X_i σε ένα γραμμικό μοντέλο.

Ένα άλλο ζήτημα είναι η εξήγηση του γραμμικού μοντέλου που ερευνούμε, όπου κάθε μία από τις επεξηγηματικές μεταβλητές μέσω του συντελεστή b_i που προκύπτει από το γραμμικό μοντέλο συμβάλει στην επίδραση πάνω στην εξαρτημένη μεταβλητή Y. Λογικό είναι να αναζητούμε την εξάρτηση της Y από όσο το δυνατόν λιγότερες μεταβλητές X_i αλλά βέβαια χωρίς να χάνεται και η εξάρτηση-εξήγησή της από τις ανεξάρτητες μεταβλητές. Επίσης, 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 τα αποτελέσματα της παλινδρόμησης, ειδικά αν δεν έχουμε πολλές παρατηρήσεις. Αντιθέτως, η ύπαρξη μεγάλου πλήθους ανεξάρτητων μεταβλητών σε ένα γραμμικό μοντέλο μπορεί να δημιουργήσει πρόβλημα στην ακρίβεια των εκτιμήσεων, δηλαδή να αυξήσει την διασπορά των εκτιμήσεων των παραμέτρων, κάτι που δεν θέλουμε καθόλου.

Το πρόβλημα του πλήθους των ανεξάρτητων τυχαίων μεταβλητών σε ένα γραμμικό μοντέλο φαίνεται χαρακτηριστικά στο σχήμα 3.1 και είναι προφανώς από τα σημαντικότερα προβλήματα στον σχεδιασμό ενός γραμμικού μοντέλου.



Σχήμα 3.1 **Αρχή της φειδωλότητας** (*Principle of Parsimony*). Η αύξηση του πλήθους των παραμέτρων σε ένα γραμμικό μοντέλο δεν είναι πάντα καλή καθώς αυξάνεται η διασπορά των εκτιμήσεων, όπως και η μείωση του πλήθους των παραμέτρων μπορεί να αυξήσει την μεροληψία των εκτιμήσεων.

Αναζητούμε, λοιπόν, σε ένα γραμμικό μοντέλο από την μία περισσότερες ανεξάρτητες μεταβλητές να εμπλέκονται στο γραμμικό μοντέλο για να είναι περισσότερο "σωστή" η επεξήγηση-εξάρτηση της Y από τις X_i, και από την άλλη, όσο το δυνατόν μικρότερο αριθμό ανεξάρτητων μεταβλητών, ώστε να έχουμε ακριβέστερη εξήγηση της μεταβλητής Y από τις ανεξάρτητες μεταβλητές. Αυτή είναι η ***Αρχή της Φειδωλότητας*** (*Principle of Parsimony*) η οποία δηλώνει πως θέλουμε μικρό αριθμό ανεξάρτητων μεταβλητών να εμπλέκονται στο γραμμικό μοντέλο οι οποίες όμως να εξηγούν όσο το δυνατών καλύτερα την εξαρτημένη μεταβλητή Y.

Για το αν ένα γραμμικό μοντέλο, όπως αυτό προκύπτει από τις παρατηρήσεις και τις εκτιμήσεις των παραμέτρων b_i, εξηγεί καλά ή όχι την εξάρτηση της μεταβλητής Y από τις μεταβλητές X_i, δεν υπάρχει μεμονωμένη απάντηση. Αλλά υπάρχουν διάφορα κριτήρια τα οποία δίνουν ένα μέτρο για την απόδοση του γραμμικού μοντέλου, οπότε συγκρίνοντας δύο ή περισσότερα γραμμικά μοντέλα μέσω των κριτηρίων αυτών μπορούμε να αποφανθούμε αν κάποιο είναι καλύτερο ή όχι. Κάποια από τα μέτρα που υπάρχουν δίνονται στην επόμενη ενότητα.

3.4 ΚΡΙΤΗΡΙΑ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ ΠΑΛΛΙΝΔΡΟΜΙΣΗΣ

Όταν ελέγχονται διάφορα γραμμικά μοντέλα με σκοπό την καλύτερη εξήγηση της εξαρτημένης μεταβλητής από τις ανεξάρτητες τυχαίες μεταβλητές, τίθεται το ερώτημα πόσες και ποιές ανεξάρτητες τυχαίες μεταβλητές πρέπει να συμπεριληφθούν στο γραμμικό μοντέλο, και πως θα γίνει η διάκριση μεταξύ δύο γραμμικών μοντέλων (με διαφορετικές ανεξάρτητες τυχαίες μεταβλητές), για το ποιο είναι καλύτερο γραμμικό μοντέλο. Δηλαδή, από τις διαθέσιμες ανεξάρτητες τυχαίες μεταβλητές ποιες μεταβλητές πρέπει να υπάρχουν στο γραμμικό μοντέλο, ώστε αυτό να είναι το βέλτιστο δυνατό. Από την μία, περισσότερες τυχαίες μεταβλητές ενδεχομένως να βελτιώνουν την αξιοπιστία των εκτιμήσεων των παραμέτρων του μοντέλου, ενώ από την άλλη λόγω της αρχής της φειδωλότητας θα θέλαμε "λίγες" επεξηγηματικές μεταβλητές.

Υπάρχουν διάφορα κριτήρια για τον χαρακτηρισμό και την διάκριση διάφορων γραμμικών μοντέλων για το ποιο είναι καλύτερο. Κάποια από αυτά είναι :

- Ο συντελεστής προσδιορισμού R^2
- Ο προσαρμοσμένος συντελεστής προσδιορισμού R^2_{Adj}
- SSE
- MSE
- AIC
- BIC

Η αύξηση του πλήθους των επεξηγηματικών τυχαίων μεταβλητών στο γραμμικό μοντέλο επηρεάζει τις τιμές των κριτηρίων αυτών με μικρότερο ή μεγαλύτερο τρόπο. Κάποια δηλώνουν καλύτερο γραμμικό μοντέλο όταν η τιμή τους είναι μεγαλύτερη (π.χ. R^2, R^2_{Adj}), ενώ κάποια δηλώνουν καλύτερο γραμμικό μοντέλο όταν η τιμή τους είναι μικρότερη (SSE, MSE, AIC, BIC). Για παράδειγμα ο προσαρμοσμένος συντελεστής προσδιορισμού R^2_{Adj} δηλώνει το ποσοστό της εξαρτημένης μεταβλητής που εξηγείται από το γραμμικό μοντέλο. Οπότε όσο πιο κοντά στο ένα λαμβάνει τιμή, τόσο καλύτερο είναι και το αντίστοιχο γραμμικό μοντέλο. Ενώ το SSE όσο πιο μικρό είναι τόσο καλύτερο είναι το μοντέλο, καθώς ισούται με το άθροισμα των τετραγώνων των καταλοίπων.

Αν $\hat{y}_i$ είναι η προβλεπόμενη τιμή της μεταβλητής απόκρισης με βάση το πολλαπλό γραμμικό μοντέλο και $\bar{y}$ η μέση τιμή με βάση n παρατηρήσεις που διαθέτουμε, τότε ορίζουμε :

$$\boldsymbol{SSE} = \sum \hat{\boldsymbol{\varepsilon}}_i^2 = \sum (\bar{\boldsymbol{y}} - \hat{\boldsymbol{y}}_i)^2$$

$$\boldsymbol{SST} = \sum (\boldsymbol{y}_i - \bar{\boldsymbol{y}})^2$$

$$\boldsymbol{SSR} = \sum (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2$$

Επιπλέον ισχύει η σχέση : $\boldsymbol{SST} = \boldsymbol{SSE} + \boldsymbol{SSR}$

Ακόμα, το μέσο τετραγωνικό σφάλμα MSE ορίζεται ως:

$$\boldsymbol{MSE} = \frac{1}{\boldsymbol{n}} \sum (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2 = \frac{\boldsymbol{SSR}}{\boldsymbol{n}}$$

Ο συντελεστής προσδιορισμού είναι

$$\boldsymbol{R}^2 = \frac{\boldsymbol{SSR}}{\boldsymbol{SST}} = \frac{\boldsymbol{SST} - \boldsymbol{SSE}}{\boldsymbol{SST}} \quad ,$$

ενώ ο προσαρμοσμένος συντελεστής προσδιορισμού είναι

$$\boldsymbol{R}^2_{\boldsymbol{Adj}} = \boldsymbol{R}^2 - (1 - \boldsymbol{R}^2) \frac{\boldsymbol{p}}{\boldsymbol{n} - \boldsymbol{p} - 1}$$

όπου n το πλήθος των παρατηρήσεων και p το πλήθος των επεξηγηματικών μεταβλητών και διορθώνει τον R^2 ως προς το πλήθος των μεταβλητών.

Το MSE είναι το μέσο τετραγωνικό σφάλμα και είναι

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

Το κριτήριο ***AIC***, *Akaike's information criterion*, ορίσθηκε το 1971 από τον Akaike :

$$\boldsymbol{AIC} = \boldsymbol{n} \cdot \ln\left(\frac{\boldsymbol{SSE}}{\boldsymbol{n}} \right) + 2(\boldsymbol{p} + 1)$$

όπου ***n*** το πλήθος των παρατηρήσεων και ***p*** το πλήθος των ανεξάρτητων μεταβλητών στο γραμμικό μοντέλο. Καλύτερο γραμμικό μοντέλο θεωρείται εκείνο που έχει μικρότερη τιμή του κριτηρίου ***AIC.*** Το κριτήριο *AIC* ποινικοποιεί την ύπαρξη μεγάλου πλήθους ανεξάρτητων μεταβλητών, ή για την ακρίβεια την διάσταση του παραμετρικού χώρου του μοντέλου (p+1) με συντελεστή 2. Καθώς το πλήθος των επεξηγηματικών μεταβλητών αυξάνεται (αύξηση του p), αυξάνεται και η τιμή *AIC* ενώ για μεγάλο πλήθος παρατηρήσεων μειώνεται συγκριτικά η επίδραση του SSE στην τιμή του κριτηρίου.

Το κριτήριο ***BIC***, *Bayesian information criterion*, ορίσθηκε το 1978 από τον Schwartz ως

$$\boldsymbol{BIC} = \boldsymbol{n} \cdot \ln\left(\frac{\boldsymbol{SSE}}{\boldsymbol{n}} \right) + \ln(\boldsymbol{n}) \cdot (\boldsymbol{p} + 1)$$

όπου ***n*** το πλήθος των παρατηρήσεων και ***p*** το πλήθος των ανεξάρτητων μεταβλητών στο γραμμικό μοντέλο και σε σχέση με το κριτήριο AIC είναι αυστηρότερο ως προς την ποινικοποίηση της αύξησης του πλήθους *p* των ανεξάρτητων τυχαίων μεταβλητών X_i που συμπεριλαμβάνονται στο γραμμικό μοντέλο. Συγκεκριμένα, για την διάσταση του παραμετρικού χώρου (p+1) του γραμμικού μοντέλου, ο συντελεστής είναι *ln(n)* εν αντιθέσει του συντελεστή *2* του κριτηρίου *AIC*. Για τα δύο κριτήρια ισχύει ότι, όταν το πλήθος των παρατηρήσεων είναι μικρότερο του 8 το κριτήριο *AIC* είναι αυστηρότερο ενώ αν το πλήθος των μεταβλητών είναι μεγαλύτερο του 7 το κριτήριο *BIC* είναι αυστηρότερο (καθώς ln(n)>2 σημαίνει $n>e^2$ δηλαδή n>7.38). Καλύτερο γραμμικό μοντέλο θεωρείται και στην περίπτωση του *BIC,* εκείνο που έχει μικρότερη τιμή και καθώς έχει μεγαλύτερη επιρροή ο αριθμός των τυχαίων μεταβλητών που συμπεριλαμβάνονται στο γραμμικό μοντέλο, είναι και το κριτήριο

που επιλέχθηκε για την σύγκριση και καθορισμό του βέλτιστου γραμμικού μοντέλου στην διαδικασία εύρεσης αυτού από τον γενετικό αλγόριθμο.

ΚΕΦΑΛΑΙΟ 4

Ο ΓΕΝΕΤΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ ΣΤΟ ΠΡΟΒΛΗΜΑ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

4.1 ΕΙΣΑΓΩΓΗ- Ο ΑΛΓΟΡΙΘΜΟΣ

Γενετικός αλγόριθμος για την εύρεση επιλογής μεταβλητών.
Αναπτύσσεται στα παρακάτω ο σχεδιασμός και η δομή του γενετικού αλγόριθμου που αναπτύξαμε για την εύρεση του καλύτερου γραμμικού μοντέλου όσων αφορά τις επεξηγηματικές μεταβλητές που θα πρέπει να συμπεριληφθούν στο γραμμικό μοντέλο. Συγκεκριμένα, σε ένα πολλαπλό γραμμικό μοντέλο υπάρχουν **n** παρατηρήσεις από **m** τυχαίες μεταβλητές, οι οποίες θεωρούνται και ανεξάρτητες μεταξύ τους (τις m ανεξάρτητες μεταβλητές Xi, i=1,…,m) και **n** παρατηρήσεις από μία άλλη τυχαία μεταβλητή (την εξαρτημένη μεταβλητή Y). Γίνεται έλεγχος για την γραμμική σχέση της εξαρτημένης μεταβλητής με κάποια ή κάποιες από τις ανεξάρτητες μεταβλητές (ακόμα και όλες). Υπάρχουν διάφορες μέθοδοι για την εύρεση του καλύτερου γραμμικού μοντέλου εκ των παραπάνω, και διάφορα κριτήρια για το πώς ένα γραμμικό μοντέλο θεωρείται καλύτερο από ένα άλλο. Εμείς θα χρησιμοποιήσουμε τον γενετικό αλγόριθμο για την εύρεση του καλύτερου γραμμικού μοντέλου και ως κριτήριο θα έχουμε την τιμή του BIC για κάθε ένα από τα γραμμικά μοντέλα. Η τιμή BIC όσο μικρότερη είναι τόσο καλύτερο είναι το αντίστοιχο γραμμικό μοντέλο. Επίσης, κάθε γραμμικό μοντέλο κωδικοποιείται ως ένα δυαδικό διάνυσμα, συγκεκριμένα ένα διάνυσμα m μεταβλητών όσες και οι τυχαίες μεταβλητές Xi, με τιμές 0 και 1. Όταν η i-οστή συντεταγμένη έχει τιμή 0 το αντίστοιχο γραμμικό μοντέλο δεν περιλαμβάνει την i-οστή τυχαία μεταβλητή Xi , ενώ όταν η i-οστή συντεταγμένη έχει τιμή 1 το αντίστοιχο γραμμικό μοντέλο συμπεριλαμβάνει την i-οστή τυχαία μεταβλητή Xi. Το δυαδικό αυτό διάνυσμα θα αποκαλείται χρωμόσωμα. Δηλαδή, τα βασικά στοιχεία του γενετικού αλγορίθμου είναι το χρωμόσωμα, και το περιβάλλον “μέσα” στο οποίο γίνεται η έρευνα/βελτιστοποίηση είναι ο χώρος των τιμών *BIC* των πιθανών μοντέλων για διάφορες επεξηγηματικές μεταβλητές. Η συνάρτηση *BIC* αποτελεί και την συνάρτηση αξιολόγησης για την εύρεση της βέλτιστης λύσης ή αλλιώς την αντικειμενική συνάρτηση. Επίσης, ως πληθυσμός ορίζεται το πλήθος των χρωμοσωμάτων (αντιστοίχως γραμμικών μοντέλων) που υπάρχουν στο περιβάλλον, και αποτελούν μία γενιά. Κάθε φορά ο αλγόριθμος (ανα)παράγει νέα χρωμοσώματα/άτομα του πληθυσμού, τα οποία και θα αποτελούν την νέα γενιά έως ότου καταλήξει στο “βέλτιστο μοντέλο”, πάντα κατά τον αλγόριθμο.

4.2 ΕΛΕΓΧΟΣ - ΔΟΜΗ ΑΛΓΟΡΙΘΜΟΥ - ΣΧΕΔΙΑΣΜΟΣ

Έλεγχος

Για τον έλεγχο της σωστής λειτουργίας του γενετικού αλγορίθμου για την εύρεση βέλτιστου γραμμικού μοντέλου, προσομοιώσαμε δεδομένα τα οποία βασίζονται στην κανονική κατανομή.
Συγκεκριμένα προσομοιώσαμε m=15 τυχαίες μεταβλητές από n=50 παρατηρήσεις η καθεμία που βασίζονται στην κανονική κατανομή με μέση τιμή μηδέν και τυπική απόκλιση ένα. Αυτές οι 15 τυχαίες μεταβλητές παίζουν το ρόλο των ανεξάρτητων τυχαίων μεταβλητών για το γραμμικό μοντέλο Xi, i=1,2,…,m.
Οι αντίστοιχες εντολές του κώδικα για την παραγωγή των εν λόγω τιμών είναι :

```
Nrow=50
X<-matrix(rep(0,Nrow*Ncol),ncol=Ncol)
for (i in 1:Ncol){
   X[,i]<-rnorm(Nrow,0,1)
}
```

Όλες τις παρατηρήσεις τις καταχωρούμε σε έναν πίνακα X όπου κάθε γραμμή αποτελεί ένα χρωμόσωμα (δηλαδή ένα διαφορετικό μοντέλο), ενώ κάθε στήλη αποτελεί τις 50 τιμές/παρατηρήσεις κάθε i-οστής τυχαίας μεταβλητής.
Επίσης προσομοιώσαμε και 50 παρατηρήσεις σε μια μεταβλητή Y η οποία θα είναι και η εξαρτημένη μεταβλητή. Η μεταβλητή Y βασίζεται μόνο σε τρεις από τις 15 τυχαίες μεταβλητές. Συγκεκριμένα στην πρώτη, στην έβδομη και στην δωδέκατη μεταβλητή σύμφωνα με τον παρακάτω τύπο :

$$\mathbf{Y_j=3+8*X_{1,j}+4*X_{7,j}-4*X_{12,j}+e_j}\ ,\ j=1,2,\ldots,50,\ e_j \sim N(0,s^2)$$

Οι αντίστοιχες εντολές του κώδικα για την παραγωγή των τιμών είναι :

```
Y<-rnorm(Nrow,3+8*X[,1]+4*X[,7]-4*X[,12] ,2)
```

Παρατήρηση

Εδώ να σημειώσουμε ότι η τυπική απόκλιση με τιμή δύο, μπορεί να παρουσιάσει προβλήματα στην αποδοτικότητα του γενετικού αλγορίθμου, για αυτόν το λόγο θα μπορούσαμε να μειώσουμε την τυπική απόκλιση στο ενάμιση ή ακόμα και στο ένα, για να μην έχουμε τέτοια προβλήματα. Αλλά καλό είναι να έχουμε αρχικά τυπική απόκλιση λίγο μεγαλύτερη, όπως το δύο, για να ελέγξουμε καλύτερα την αποδοτικότητα του αλγορίθμου. Συγκεκριμένα, μεγάλη διασπορά συνεπάγεται μεγαλύτερα σφάλματα για τις αντίστοιχες εκτιμήσεις και αυτό με την σειρά του συνεπάγεται μεγαλύτερη πιθανότητα ο αλγόριθμος να μην συγκλίνει στο βέλτιστο

μοντέλο. Επομένως, στο μοντέλο μας χρησιμοποιούμε μία τυπική απόκλιση ούτε πολύ μεγάλη ούτε πολύ μικρή για να μπορούμε να ελέγξουμε την αποδοτικότητα του αλγορίθμου.
Σκοπός μας είναι να ελέγξουμε τον γενετικό αλγόριθμο και αυτό θα φανεί από το καλύτερο γραμμικό μοντέλο στο οποίο θα καταλήξει ο αλγόριθμος. Δηλαδή, αν το καλύτερο μοντέλο περιλαμβάνει μόνο την πρώτη, την έβδομη και την δωδέκατη τυχαία μεταβλητή (αντίστοιχα συντεταγμένη του δυαδικού χρωμοσώματος), τότε θα ξέρουμε πως η εύρεση του καλύτερου γραμμικού μοντέλου είναι σωστή και ο αλγόριθμος αποδοτικός, δηλαδή πως δουλεύει σωστά.
Η τεχνική αυτή επαλήθευσης της απόδοσης του αλγορίθμου είναι αναγκαία γιατί μέσω αυτής θα γνωρίζουμε πως το εργαλείο που φτιάξαμε είναι πρώτον σωστό και κατά δεύτερον θα μπορούμε να ελέγξουμε τις τιμές των παραμέτρων για τις οποίες ο αλγόριθμος δουλεύει καλύτερα. Οι παράμετροι όπως θα αναφέρουμε αναλυτικά και παρακάτω είναι ο πληθυσμός, οι επαναλήψεις, ο τύπος διασταύρωσης, η μετάλλαξη κ.α.

Δομή Αλγορίθμου – Σχεδιασμός

Ο γενετικός αλγόριθμος έχει διαχωριστεί σε τρία μέρη. Τα δύο πρώτα μέρη είναι δύο μικρότερες συναρτήσεις, οι οποίες καλούνται πολλές φορές στην τρίτη συνάρτηση μέσα στις επαναλήψεις. Η τρίτη συνάρτηση είναι και το κυρίως μέρος του αλγορίθμου, με την παραγωγή του αρχικού πληθυσμού, τον υπολογισμό των τιμών BIC, όπως και τις επαναλήψεις που αφορούν την παραγωγή κάθε φορά μίας νέας γενιάς μέσω των διασταυρώσεων των μελών της προηγούμενης γενιάς.
Ο σκοπός που έγιναν οι δύο πρώτες μικρότερες συναρτήσεις είναι:
Πρώτον, γιατί είναι μικρές συναρτήσεις που επαναλαμβάνονται πολλές φορές μέσα στον αλγόριθμο. Για αυτόν το λόγο είναι καλύτερο να υπάρχουν έξω από το κυρίως σώμα του αλγορίθμου σαν αυτόνομες συναρτήσεις και να καλούνται εύκολα και γρήγορα μέσα στο κυρίως σώμα του. Με αυτόν τον τρόπο μειώνεται ο όγκος του αλγορίθμου καθώς και η πολυπλοκότητα του όταν το διαβάζει κάποιος.
Δεύτερον, είναι πάντα πολύ εύκολο να γίνουν μικρές αλλαγές στις συναρτήσεις αυτές και να μην επηρεάζουν καθόλου τον κώδικα στο κυρίως σώμα του αλγορίθμου δηλαδή την τρίτη συνάρτηση. Για παράδειγμα, έχουμε την δυνατότητα να αλλάξουμε την συνάρτηση BIC, η οποία είναι και η αντικειμενική συνάρτηση, και να βάλουμε στην θέση της μια άλλη συνάρτηση-κριτήριο των αντίστοιχων γραμμικών μοντέλων χωρίς να πειράξουμε καθόλου το κυρίως σώμα του αλγορίθμου.
Τέλος, η κάθε μία από τις συναρτήσεις είναι φτιαγμένη έτσι ώστε να λειτουργεί για συγκεκριμένους τύπους δεδομένων της **R** (π.χ. διανύσματα και πίνακες), αλλά με μια γενικότερη ελευθερία στις διαστάσεις αυτών. Για παράδειγμα, ένα δυαδικό χρωμόσωμα μπορεί να έχει μήκος **n** το οποίο δεν είναι δεδομένο εξ αρχής, αλλά η συνάρτηση να αναγνωρίζει αυτό το μήκος. Η συνάρτηση θα μπορεί να εφαρμόσει τις εντολές της πάνω σε αυτό το δυαδικό χρωμόσωμα οποιουδήποτε μήκους. Αυτός ο γενικότερος κατά μία έννοια σχεδιασμός των αλγορίθμων χρησιμοποιείται ευρέως

στον προγραμματισμό και έτσι είναι δυνατή η εφαρμογή των αλγορίθμων σε δεδομένα τα οποία πρέπει να έχουν μια συγκεκριμένη δομή, αλλά δεν περιορίζονται σε συγκεκριμένες διαστάσεις.
Η **πρώτη** από τις δύο μικρότερες συναρτήσεις αφορά τον υπολογισμό του BIC για κάθε χρωμόσωμα του πληθυσμού. Η συνάρτηση αυτή, η οποία ονομάζεται ***eval.function*** δίνεται από τις παρακάτω εντολές:

```
eval.fun<-function(chromos){
        Nrow=dim(X)[1]
        BIC<-AIC(lm(Y~X[,chromos==1 ]), k = log(Nrow) )
        return(BIC)
}
```

Η συνάρτηση αυτή δέχεται στο όρισμα της ένα μεμονωμένο δυαδικό χρωμόσωμα (*chromos*). Λόγω του ότι θέλουμε να έχει η συνάρτηση όση μεγαλύτερη ελευθερία γίνεται, διαβάζει το πλήθος των παρατηρήσεων των ανεξάρτητων μεταβλητών με την εντολή (*Nrow=dim(X)[1]*). Η εντολή αυτή διαβάζει την πρώτη διάσταση από τις δύο του πίνακα X (δηλαδή τις γραμμές του X), και η οποία αντιστοιχεί στο πλήθος των παρατηρήσεων για το γραμμικό σύστημα. που χρειάζεται για τον υπολογισμό της τιμής του BIC. Υπολογίζει την τιμή BIC για το γραμμικό μοντέλο (*BIC*) μέσω της εντολής της **R** *AIC(...)*, και επιστρέφει στο τέλος της συνάρτησης την τιμή αυτή με την εντολή *return(BIC).*

Να σημειώσουμε πως η μεμονωμένη συνάρτηση που αποδίδει την τιμή του BIC (*eval.function*) σε κάθε χρωμόσωμα, καλείται από το κυρίως μέρος του αλγορίθμου για τα χρωμοσώματα του αρχικού πληθυσμού αλλά και σε κάθε επανάληψη μέσα στον αλγόριθμο για τον υπολογισμό των αντίστοιχων τιμών BIC κάθε νέου χρωμοσώματος.

Η **δεύτερη** συνάρτηση, η οποία καλείται στο κυρίως σώμα του αλγορίθμου, αφορά την μετάλλαξη σε κάθε χρωμόσωμα, να υπενθυμίσουμε ότι ένα χρωμόσωμα όταν αφορά ένα γραμμικό μοντέλο μπορεί να είναι δυαδικό με τιμές 0 και 1 σε κάθε συντεταγμένη του οποίου το πλήθος των συντεταγμένων ισούται με το πλήθος των ανεξάρτητων τυχαίων μεταβλητών και όπου το χρωμόσωμα έχει τιμή ένα σε μια συγκεκριμένη συντεταγμένη η αντίστοιχη τυχαία μεταβλητή συμπεριλαμβάνεται στο γραμμικό μοντέλο. Για παράδειγμα ένα δυαδικό χρωμόσωμα που έχει σε όλες τις συντεταγμένες του μηδέν εκτός από την τρίτη συντεταγμένη όπου έχει ένα, θα σημαίνει πως το αντίστοιχο γραμμικό μοντέλο συμπεριλαμβάνει μόνο την τρίτη ανεξάρτητη τυχαία μεταβλητή. Επομένως, όταν μία συντεταγμένη του χρωμοσώματος έχει την τιμή ένα το χρωμόσωμα αντιστοιχεί σε ένα γραμμικό μοντέλο το οποίο περιλαμβάνει την συγκεκριμένη ανεξάρτητη τυχαία μεταβλητή. Η μετάλλαξη για ένα χρωμόσωμα σημαίνει πως κάποιες συντεταγμένες αλλάζουν τιμή

από 1 σε 0 και από 0 σε 1, βγάζοντας ή βάζοντας μία τυχαία μεταβλητή στο γραμμικό μοντέλο. Αυτή η λειτουργία είναι σημαντικότατη για τον γενετικό αλγόριθμο, καθώς

ερευνά νέα μοντέλα με μικρές (ή και μεγαλύτερες) αλλαγές και έτσι να επιτύχει κάποιο μοντέλο που θα είναι καλύτερο από τα προηγούμενα.

Η συνάρτηση αυτή ονομάζεται ***Mutation*** και δίνεται από τις παρακάτω εντολές:

```
Mutation<-function(chro,p.m){
    for (i in 1:length(chro) ){
      if (runif(1)<=p.m){
            chro[i]<-1-chro[i]
      }
    }
return(chro)
}
```

Η συνάρτηση *Mutation* δέχεται στο όρισμα της ένα χρωμόσωμα (*chro*) και την αντίστοιχη τιμή της πιθανότητας μετάλλαξης (*p.m*). Για κάθε συντεταγμένη του χρωμοσώματος η συνάρτηση αυτή αλλάζει την υπάρχουσα τιμή της συντεταγμένης από 1 σε 0 και αντίστροφα, σύμφωνα με τη δοσμένη πιθανότητα μετάλλαξης μέσω του ορίσματος (*p.m*). Η τιμή p.m θα πρέπει ως πιθανότητα να ανήκει στο κλειστό διάστημα [0,1], όπου φυσικά για τιμή ίση με 0 δεν γίνεται καμία μετάλλαξη στο χρωμόσωμα.

Συγκεκριμένα, για κάθε μία συντεταγμένη του χρωμοσώματος (i-οστή συντεταγμένη), παράγεται μία τιμή μεταξύ του μηδενός και του ένα, από την ομοιόμορφη κατανομή (*runif(1)*) και ελέγχεται αν αυτή η τιμή είναι μικρότερη της πιθανότητας μετάλλαξης *p.m*. Αυτό έχει ως ισοδύναμο αποτέλεσμα να επιλεχθεί ένα σημείο στο ευθύγραμμο τμήμα του άξονα των αριθμών μεταξύ 0 και 1. Η πιθανότητα να είναι το σημείο αυτό μικρότερο της πιθανότητας μετάλλαξης του ορίσματος, είναι ίση με το μήκος του τμήματος [0,p.m] προς το μήκος του τμήματος [0,1], δηλαδή την επιθυμητή πιθανότητα. Όταν συμβαίνει αυτό γίνεται η αλλαγή της τιμής της συγκεκριμένης συντεταγμένης του χρωμοσώματος, διαφορετικά δεν αλλάζει η συντεταγμένη αυτή. Αυτό γίνεται με την εντολή καταχώρησης *chro[i]<-1-chro[i]* όπου αν η τιμή της συντεταγμένης ήταν 0, η πράξη θα δώσει 1-0=1, ενώ αν η τιμή της συντεταγμένης ήταν 1 η πράξη θα δώσει 1-1=0. Στην περίπτωση όπου δεν έχουμε την επιθυμητή πιθανότητα p.m τότε δεν εφαρμόζεται καμία εντολή και η αντίστοιχη συντεταγμένη μένει όπως είχε.

Ακόμα, η συνάρτηση στο γενικότερο πλαίσιο της ελευθερίας που αναφέραμε και προηγουμένως, δεν απαιτεί συγκεκριμένο μήκος του χρωμοσώματος αλλά

αναγνωρίζει το μήκος αυτό με την εντολή *length(...)*. Έτσι, μπορεί να εφαρμοστεί σε χρωμοσώματα οποιουδήποτε μήκους.
Στο τέλος η συνάρτηση επιστρέφει το τελικό μεταλλαγμένο χρωμόσωμα με την εντολή (*return(chro)*), το οποίο ενδέχεται να είναι ίδιο και με το αρχικό, αναλόγως την πιθανότητα μετάλλαξης που έχει δεχθεί στο όρισμά της η συνάρτηση.

Το **τρίτο** μέρος του αλγορίθμου αποτελεί και το κυρίως σώμα του. Ονομάζεται **genetikos.alg.lm** συμβολίζοντας έτσι τον *γενετικό αλγόριθμο για γραμμικά μοντέλα.* Μέσα σε αυτό το μέρος του αλγορίθμου γίνεται η παραγωγή του αρχικού τυχαίου πληθυσμού, η "ταξινόμηση" των χρωμοσωμάτων βάσει της τιμής του BIC (συνάρτηση αξιολόγησης), η "αναπαραγωγή - διασταύρωση" των ατόμων του πληθυσμού, και ότι άλλο απαιτεί ένας γενετικός αλγόριθμος για την παραγωγή συγκεκριμένου αριθμού νέων γενεών έως ότου καταλήξει στο βέλτιστο χρωμόσωμα και αντίστοιχα βέλτιστο γραμμικό μοντέλο.
Επειδή, το συγκεκριμένο μέρος του αλγορίθμου είναι και πολύπλοκο και μεγάλο σε όγκο, θα αναπτυχθεί εκτενώς στην παρακάτω ενότητα. Θα δοθεί το σύνολο των μεταβλητών που χρησιμοποιούνται στο αλγόριθμο, και επίσης θα αναπτυχθεί και η λογική πάνω στην οποία βασίστηκε το στήσιμό του.

4.3 ΑΝΑΛΥΣΗ ΓΕΝΕΤΙΚΟΥ ΑΛΓΟΡΙΘΜΟΥ

Όπως έχουμε αναφέρει, το κυρίως σώμα του αλγορίθμου είναι η τρίτη συνάρτηση η οποία ονομάζεται genetikos.alg.lm συμβολίζοντας έτσι τον γενετικό αλγόριθμο για προβλήματα επιλογής μεταβλητών, και είναι το μέρος του αλγορίθμου όπου γίνεται οτιδήποτε απαιτεί ένας γενετικός αλγόριθμος. Δέχεται όλες τις απαραίτητες παραμέτρους για την εφαρμογή εναλλακτικών μεθόδων/τρόπων εύρεσης της βέλτιστης λύσης μέσω του ορίσματος της συνάρτησης αυτής. Κατά την εκτέλεση των εντολών της συνάρτησης αυτής, τυπώνονται στην οθόνη του χρήστη και κάποια από τα ενδιάμεσα αποτελέσματα της αναζήτησης για την βέλτιστη λύση. Όπως έχουμε προαναφέρει, ο αλγόριθμος αναζητεί το άτομο του πληθυσμού με την καλύτερη τιμή ως προς την αντικειμενική συνάρτηση, με άτομα του πληθυσμού να είναι τα δυαδικά διανύσματα που αντιστοιχούν σε γραμμικά μοντέλα με διάφορες επεξηγηματικές μεταβλητές, ενώ η αντικειμενική συνάρτηση είναι η συνάρτηση BIC και είναι μία από τις δύο βοηθητικές συναρτήσεις όπως την αναπτύξαμε στην προηγουμένη ενότητα. Θα αναφερθούν παρακάτω οι μεταβλητές – παράμετροι τις οποίες η συνάρτηση δέχεται στο όρισμά της, όπως και όλες οι υπόλοιπες μεταβλητές οι οποίες χρησιμοποιούνται στην συνάρτηση.

Όρισμα της συνάρτησης

Το όρισμα της συνάρτησης *genetikos.alg.lm* φαίνεται παρακάτω

```
genetikos.alg.lm<-function( X, Y, population=50, epanalipseis=50,
evaluation=eval.fun, p.crossover=1, p.type.crossover=1,
pososto.palaiwn=0.20 , elitist=FALSE, p.mutation=0.01, analogia.0.1=c(1,1) )
{
```

όπου με την σειρά δηλώνονται οι παρακάτω μεταβλητές και πρέπει ο χρήστης να είναι αυστηρός στην δομή και τον τύπο αυτών των μεταβλητών:

- ***X***

  είναι ένας πίνακας διάστασης nxm όπου περιέχει τα δεδομένα/παρατηρήσεις των ανεξάρτητων μεταβλητών. Στις στήλες αντιστοιχούν οι m διαφορετικές ανεξάρτητες μεταβλητές *Xj, j=1,…,m* για τις οποίες αναζητείται ο βέλτιστος συνδυασμός για το γραμμικό μοντέλο, ενώ στις γραμμές αντιστοιχούν οι n παρατηρήσεις κάθε μεταβλητής. Αναλυτικότερα, το στοιχείο στην (i,j) θέση του πίνακα ***X*** αντιστοιχεί στην iοστή παρατήρηση της jοστής μεταβλητής, όπου το i μπορεί να παίρνει τιμές από 1 έως n και το j από 1 έως m.

- ***Y***

  είναι ένα διάνυσμα μήκους n, με τιμές που αντιστοιχούν στην εξαρτημένη μεταβλητή του γραμμικού μοντέλου.

- ***population=50***

  είναι η παράμετρος του γενετικού αλγορίθμου που δηλώνει το μέγεθος του πληθυσμού των ατόμων, όπως έχει αυτός αναφερθεί σε προηγούμενη ενότητα. Η παράμετρος αυτή για τον αλγόριθμο έχει ορισθεί να λαμβάνει την τιμή 50 ως προκαθορισμένη (default) τιμή, δηλαδή εάν δεν δηλωθεί μέσα στο όρισμα από τον χρήστη, η συνάρτηση θα τρέξει με την συγκεκριμένη παράμετρο να έχει την τιμή 50.

- ***epanalipseis=50***

  είναι η παράμετρος του γενετικού αλγορίθμου που δηλώνει το πλήθος των επαναλήψεων που θα γίνουν μέσα στον αλγόριθμο, και αντιστοιχούν στο πλήθος των γενεών που θα παραχθούν έως ότου σταματήσει ο αλγόριθμος. Και για αυτήν την παράμετρο έχει δηλωθεί στο όρισμα προκαθορισμένη (default) τιμή ίση με 50.

- ***evaluation=eval.fun***

  Είναι η παράμετρος που αντιστοιχεί στο περιβάλλον του πληθυσμού για τον γενετικό αλγόριθμο. Συγκεκριμένα, είναι η συνάρτηση που χρησιμοποιείται στον αλγόριθμο ως συνάρτηση αξιολόγησης "του βέλτιστου" των ατόμων για τον πληθυσμό (αντικειμενική συνάρτηση), και έχει δηλωθεί στο όρισμα προκαθορισμένη (default) τιμή για αυτήν η βοηθητική συνάρτηση *eval.fun* δηλαδή η συνάρτηση BIC.

- ***p.crossover=1***

  η μεταβλητή αυτή δηλώνει την πιθανότητα σύμφωνα με την οποία μία διασταύρωση θα πραγματοποιηθεί ή όχι. Στον γενετικό αλγόριθμο υπάρχει η περίπτωση να πραγματοποιείται μία διασταύρωση ή να μην πραγματοποιείται οπότε και προκύπτουν οι δύο απόγονοι ή όχι από τους συγκεκριμένους γονείς.

- ***p.type.crossover=1***

  η μεταβλητή αυτή δηλώνει τον τρόπο με τον οποίο θέλει ο χρήστης να γίνει η διασταύρωση των γονέων. Η μεταβλητή *p.type.crossover* μπορεί να πάρει τρεις διαφορετικές τιμές, συγκεκριμένα μπορεί να είναι ίση με ***1*** οπότε με αυτό ο χρήστης δηλώνει πως η διασταύρωση των γονέων θέλει να γίνει με ένα σημείο (1-point crossover), ή μπορεί να είναι ίση με ***2*** οπότε με αυτό ο χρήστης δηλώνει πως η διασταύρωση θέλει να γίνει με δύο σημεία (2-points crossover), ή τελικά μπορεί να είναι ίση με ***3*** οπότε με αυτό ο χρήστης δηλώνει πως θέλει να γίνει ομοιόμορφη διασταύρωση, δηλαδή με την βοήθεια μάσκας όπως έχουμε αναφέρει στο θεωρητικό μέρος. Και εδώ έχει δηλωθεί προκαθορισμένη (default) τιμή για την μεταβλητή αυτή και είναι ίση με 1, δηλαδή για διασταύρωση ενός σημείου.

- ***pososto.palaiwn=0.20***

  η μεταβλητή αυτή δηλώνει το ποσοστό των καλύτερων ατόμων του πληθυσμού που θέλουμε να κρατάει ο αλγόριθμος από το σύνολο του πληθυσμού σε κάθε επανάληψη/παραγωγή γενιάς. Δηλαδή, στο τέλος των διασταυρώσεων κάθε επανάληψης, το ποσοστό του πληθυσμού που ορίζεται μέσω της μεταβλητής αυτής, είναι το τμήμα του πληθυσμού που κρατά ο αλγόριθμος για την επόμενη γενιά, από τα άτομα του έως τότε πληθυσμού, και θα αποτελέσουν τους πιθανούς γονείς της επόμενης γενιάς. Αυτό βέβαια σημαίνει πως το αντίστοιχο συμπληρωματικό ποσοστό του πληθυσμού θα αναπληρωθεί/παραχθεί τυχαία από τον αλγόριθμο, όπως βέβαια και ο αρχικός πληθυσμός πριν την πρώτη διασταύρωση. Και εδώ έχει δηλωθεί default τιμή για την μεταβλητή αυτή και είναι ίση 0.20, δηλαδή το 80% των όχι καλύτερων ατόμων του πληθυσμού χάνεται και ο αλγόριθμος ανανεώνει τυχαία αυτό το τμήμα του πληθυσμού. Προφανώς ο χρήστης θα πρέπει να δηλώσει τιμές μεταξύ του μηδενός και του ένα για την παράμετρο αυτή.

- **elitist=FALSE**

  η μεταβλητή αυτή δηλώνει για τον χρήστη αν θέλει να εφαρμόζεται ελιτισμός ή όχι μετά από κάθε διασταύρωση. Εάν θέλει ο χρήστης να επιλέγονται τα δύο καλύτερα άτομα από τους δύο γονείς και δύο παιδιά που προκύπτουν από μία διασταύρωση θα πρέπει να δηλώσει ως TRUE την μεταβλητή αυτή, ενώ αν δηλώσει FALSE ο αλγόριθμος μετά από την διασταύρωση θα κρατήσει τα δύο παιδιά στην νέα γενιά είτε είναι καλύτερα από τους δύο γονείς είτε όχι. Μπορεί βέβαια ο χρήστης να επιτύχει το ίδιο χωρίς να δηλώσει την μεταβλητή

αυτή καθώς έχει δηλωθεί ως προκαθορισμένη (default) τιμή για την μεταβλητή αυτήν η FALSE.

- ***p.mutation=0.01***

  η μεταβλητή αυτή χρησιμοποιείται ως πιθανότητα να συμβεί μετάλλαξη σε ένα γονίδιο. Συγκεκριμένα, η τιμή της μεταβλητής αυτής δηλώνει την πιθανότητα να γίνει σε κάθε ένα από τα γονίδια μετάλλαξη, και η πιθανότητα αυτή χρησιμοποιείται από την αντίστοιχη βοηθητική συνάρτηση *Mutation* που έχουμε αναλύσει στην προηγουμένη ενότητα. Και εδώ έχει δηλωθεί προκαθορισμένη (default) τιμή για την μεταβλητή αυτή και είναι ίση με 0.01 δηλαδή κατά 1% μπορεί να γίνει μετάλλαξη σε κάθε γονίδιο.

- ***analogia.0.1=c(1,1)***

  η μεταβλητή αυτή χρησιμοποιείται για την τυχαία παραγωγή δυαδικών χρωμοσωμάτων από τον αλγόριθμο. Συγκεκριμένα, δηλώνει την αναλογία των μηδενικών και των άσσων που θα υπάρχουν σε ένα χρωμόσωμα. Με αυτόν τον τρόπο ο χρήστης μπορεί να ζητήσει να υπάρχουν περισσότερα μηδενικά από ότι άσσοι (ή και αντιστρόφως) σε ένα χρωμόσωμα κάθε φορά που παράγεται τυχαία από τον αλγόριθμο ένα χρωμόσωμα. Και εδώ έχει δηλωθεί default τιμή για την μεταβλητή αυτή και είναι ίση με το διάνυσμα (1,1) που σημαίνει 1:1 αναλογία μηδενικών και άσσων.

<u>Τρόπος λειτουργίας του αλγορίθμου</u>

Όπως κάθε γενετικός αλγόριθμος, έτσι και αυτός σχεδιάστηκε να ξεκινά με μια αρχική γενιά ατόμων, τον αρχικό πληθυσμό δηλαδή, και μέσω διασταυρώσεων των ατόμων του αρχικού πληθυσμού να παράγεται ο επόμενος πληθυσμός, τα λεγόμενα παιδιά του προηγούμενου πληθυσμού. Αυτή η διαδικασία επαναλαμβάνεται αρκετές φορές. Συγκεκριμένα όσες φορές έχει δηλωθεί η μεταβλητή των επαναλήψεων στο όρισμα της συνάρτησης του αλγορίθμου, τόσες θα είναι και οι γενιές οι οποίες θα διαδεχτούν τον αρχικό πληθυσμό. Οι διασταυρώσεις των ατόμων κάθε προηγούμενης γενεάς για να προκύψουν τα παιδιά τα οποία και θα αποτελούν την επόμενη γενιά ακολουθούν συγκεκριμένο τρόπο, ο οποίος δηλώνεται επίσης στο όρισμα της συνάρτησης. Ακόμα, η επιλογή των ατόμων του πληθυσμού τα οποία θα αποτελούν τους γονείς για κάθε γενιά, γίνεται βάσει της συνάρτησης αξιολόγησης (που για εμάς είναι η συνάρτηση BIC) και μέσω πιθανοτήτων οι οποίες βασίζονται σε αυτές τις τιμές BIC, κάθε φορά για τα μεμονωμένα άτομα του πληθυσμού. Για κάθε γενιά και για όλα τα άτομα του πληθυσμού, ορίζονται οι πιθανότητες επιλογής των ατόμων αυτών ως γονείς, με μεγαλύτερη να είναι η πιθανότητα για άτομα με μικρότερη τιμή BIC. Τελικά, καθώς επιλέγονται τα άτομα από τον πληθυσμό με προτεραιότητα αυτά που έχουν καλύτερη τιμή BIC, και τα οποία διασταυρώνονται, προκύπτουν νέα άτομα με πιθανόν καλύτερη (μικρότερη) τιμή BIC, και έτσι ο αλγόριθμος βρίσκει καλύτερα διανύσματα και αντιστοίχως καλύτερα γραμμικά μοντέλα.

Ο αλγόριθμος εκτελεί τις παρακάτω διαδικασίες :

1 Παράγεται τυχαία ένας αρχικός πληθυσμός χρωμοσωμάτων που αποτελούν τα άτομα από τα οποία θα γίνει και η επιλογή των γονέων για την πρώτη διασταύρωση.

2 Υπολογίζεται για κάθε άτομο του πληθυσμού (το αντίστοιχο χρωμόσωμα) η τιμή της αντικειμενικής συνάρτησης, δηλαδή η τιμή BIC που αξιολογεί το χρωμόσωμα ως καλό γραμμικό μοντέλο.

3 Βάσει των τιμών της συνάρτησης BIC, ορίζονται οι πιθανότητες με τις οποίες γίνεται η τυχαία επιλογή με επανάθεση από τα άτομα/χρωμοσώματα του πληθυσμού. Αυτά τα άτομα αποτελούν τους γονείς της νέας γενιάς (του νέου πληθυσμού).

4 Για κάθε δύο γονείς γίνεται η διασταύρωση, βάσει των επιλογών που έχουν δηλωθεί στο όρισμα της συνάρτησης του αλγορίθμου. Συγκεκριμένα για τον τύπο της διασταύρωσης μεταξύ των επιλογών για διασταύρωση ενός ή δύο σημείων ή και ομοιόμορφα (με την χρήση μάσκας όπως έχουμε αναφέρει).

5 Ακολούθως, ή γίνεται μετάλλαξη στα νέα άτομα/παιδιά του νέου πληθυσμού είτε ελιτίστικη στρατηγική (elitist), όπου elitist όπως έχουμε αναφέρει σημαίνει να επιλέγονται τα καλύτερα άτομα από τα τέσσερα που υπάρχουν σε μία διασταύρωση, δηλαδή των δύο γονέων και των δύο παιδιών, και βέβαια καλύτερα ορίζονται τα χρωμοσώματα με μικρότερο BIC.

6 Έπειτα, με σκοπό τη βελτίωση του συνόλου των ατόμων του πληθυσμού, και αντιστοίχως των γραμμικών μοντέλων, ακολουθείται στρατηγική κατά την οποία από τον πληθυσμό και των γονέων και των παιδιών, δηλαδή παλαιάς και νέας γενιάς, επιλέγονται τα καλύτερα άτομα του συνόλου, και βέβαια πάντα στον αριθμό του πληθυσμού που έχει δηλωθεί στο όρισμα της συνάρτησης.

7 Τέλος, ο αλγόριθμος κρατά από τον συνολικό πληθυσμό ένα συγκεκριμένο ποσοστό, όπως αυτό έχει δηλωθεί στο όρισμα της συνάρτησης. Αυτό αποτελεί το ποσοστό των καλύτερων ατόμων του έως τώρα πληθυσμού, ενώ παράγει τυχαία διανύσματα χρωμοσωμάτων, νέα άτομα δηλαδή, με τον ίδιο τρόπο που παρήγαγε και τον αρχικό πληθυσμό. Τα νέα άτομα είναι τόσα όσα απαιτούνται ώστε να συμπληρωθεί το 100% του αριθμού των ατόμων που αποτελούν κάθε πληθυσμό.

8 Για κάθε γενιά αποθηκεύεται η καλύτερη τιμή του BIC της γενιάς, με σκοπό να μπορεί να γίνει έλεγχος για το αν καλυτερεύουν τα γραμμικά μοντέλα κατά τη διάρκεια των επαναλήψεων.

9 Αφού τελειώσουν οι επαναλήψεις που έχουν δηλωθεί στο όρισμά της συνάρτησης ο αλγόριθμος δίνει τα αποτελέσματα του τελικού πληθυσμού, αλλά τυπώνει και την καλύτερη τιμή BIC για κάθε επανάληψη ώστε να ελέγχεται η σύγκλιση του αλγορίθμου. Επίσης, ο αλγόριθμος επιστρέφει έξω από την συνάρτηση το σύνολο των ατόμων της τελευταίας γενιάς μαζί με το διάνυσμα των τιμών BIC για κάθε χρωμόσωμα.

10 Στο τέλος κάθε επανάληψης, και κατά την διάρκεια της εκτέλεσης των εντολών του αλγορίθμου, τυπώνονται ο αριθμός της επανάληψης όπως και τα πέντε καλύτερα διανύσματα/χρωμοσώματα με τις αντίστοιχες τιμές BIC, απλά και μόνο για έλεγχο της διαδικασίας σύγκλισης του αλγορίθμου.
Για την όλη διαδικασία ο αλγόριθμος απαιτεί και αρκετές εσωτερικές μεταβλητές ώστε να αποθηκεύει, υπολογίζει, εκτελεί ενέργειες με δοσμένη πιθανότητα κ.α. Όταν μία συνάρτηση χρησιμοποιεί μεταβλητές οι οποίες δεν προέρχονται ούτε έξω από την συνάρτηση ούτε από το όρισμα αυτής, τότε αυτές ονομάζονται εσωτερικές μεταβλητές και δεν επηρεάζουν το περιβάλλον έξω από την συνάρτηση, καθώς υφίστανται μόνο κατά την διάρκεια που εκτελούνται οι εντολές της συνάρτησης και παύουν να ισχύουν όταν τελειώσει η συνάρτηση. Λειτουργούν ουσιαστικά ως βοηθητικές μεταβλητές.
Αυτές οι μεταβλητές παραθέτονται παρακάτω, ενώ μέρη του αλγορίθμου υπάρχουν ανάμεσα στο κείμενο για την ευκολότερη παρακολούθηση αυτού. Ολόκληρος ο αλγόριθμος υπάρχει στο παράρτημα της παρούσας εργασίας.

Εσωτερικές μεταβλητές της συνάρτησης *genetikos.alg.lm*

Οι μεταβλητές της συνάρτησης *genetikos.alg.lm* που χρησιμοποιούνται εσωτερικά είναι οι παρακάτω:

- ***plithos.metabl*** (αριθμός)
  Είναι η μεταβλητή που δηλώνει το (μέγιστο) πλήθος των τυχαίων μεταβλητών του γραμμικού μοντέλου. Ορίζεται με την εντολή plithos.metabl<-dim(X)[2]. Είναι ουσιαστικά και το μήκος του διανύσματος του κάθε χρωμοσώματος, καθώς κάθε συντεταγμένη ενός χρωμοσώματος αντιστοιχεί και σε μία τυχαία μεταβλητή Xi όπως έχουμε αναφέρει.

- ***best.evaluations*** (διάνυσμα)
  Είναι το διάνυσμα όπου αποθηκεύεται η καλύτερη (μικρότερη) τιμή του BIC για κάθε επανάληψη και για το σύνολο των επαναλήψεων. Ορίζεται μέσα στον αλγόριθμο με την εντολή *best.evaluations<-rep(NA,epanalipseis)* όπου απλά δημιουργείται ως διάνυσμα με ελλιπείς τιμές (NA) και θέσεις όσες οι επαναλήψεις, όπως έχει δηλωθεί από τον χρήστη μέσω της μεταβλητής του ορίσματος *epanalipseis*. Παρακάτω στον αλγόριθμο και συγκεκριμένα στο

τέλος κάθε επανάληψης αποθηκεύεται η καλύτερη τρέχουσα τιμή της τιμής BIC στην αντίστοιχη θέση του διανύσματος.

- ***evaluations.values*** (διάνυσμα)
  Είναι το διάνυσμα όπου αποθηκεύονται οι τιμές του BIC για κάθε γενιά. Ορίζεται με την εντολή *evaluations.values<-rep(NA,population)* όπου απλά δημιουργείται ως διάνυσμα με ελλιπείς τιμές (NA) και θέσεις όσες τα άτομα του πληθυσμού όπως έχει δηλωθεί από τον χρήστη μέσω της μεταβλητής του ορίσματος *population*. Μέσα στον αλγόριθμο και σε κάθε επανάληψη, οι τιμές ανανεώνονται καθώς παράγεται κάθε νέα γενιά, καθώς δεν χρειάζεται να κρατούνται στην μνήμη όλος αυτός ο όγκος δεδομένων και τιμών.

- ***new.evaluations.values*** (διάνυσμα)
  Είναι το διάνυσμα όπου αποθηκεύονται οι τιμές του BIC για το σύνολο των ατόμων μίας γενιάς γονέων και μίας γενιάς παιδιών, δηλαδή για δύο διαδοχικές γενιές. Αυτό είναι απαραίτητο για την σύγκριση και την κράτηση από τον αλγόριθμο των καλύτερων ατόμων των δύο διαδοχικών γενεών επομένως και την βελτίωση των ατόμων, και την σύγκλιση του αλγορίθμου. Ορίζεται με την εντολή *new.evaluations.values<-rep(NA,2*population)* όπου απλά δημιουργείται ως διάνυσμα με ελλιπείς τιμές (NA) και θέσεις όσες τα διπλάσια άτομα του πληθυσμού, όπως έχει δηλωθεί από τον χρήστη μέσω της μεταβλητής του ορίσματος *population*. Μέσα στον αλγόριθμο και σε κάθε επανάληψη, οι τιμές ανανεώνονται καθώς παράγεται κάθε νέα γενιά.

- ***Popul.chromos*** (πίνακας)
  Είναι ο πίνακας όπου υπάρχουν τα χρωμοσώματα/άτομα του πληθυσμού. Έχει πλήθος στηλών όσο και το πλήθος των επεξηγηματκών μεταβλητών συν μία και πλήθος γραμμών όσος και ο πληθυσμός που έχει δηλωθεί στο όρισμα της συνάρτησης. Η τελευταία στήλη χρησιμοποιείται για την αποθήκευση της τιμής BIC του αντίστοιχου χρωμοσώματος. Δηλαδή, κάθε γραμμή αντιστοιχεί σε ένα χρωμόσωμα/άτομο μαζί με την τιμή BIC αυτού. Ορίζεται αρχικά στην συνάρτηση με την βοήθεια της εντολής *matrix, Popul.chromos<-matrix(nrow=population,ncol=plithos.metabl)* όπου απλά δημιουργείται ως πίνακας και στην συνέχεια προστίθεται και η στήλη των τιμών BIC. Η ακριβής του λειτουργία είναι αρχικά για την αποθήκευση του αρχικού πληθυσμού, ενώ κατά την διάρκεια των επαναλήψεων χρησιμοποιείται για την αποθήκευση των ατόμων που αποτελούν τους γονείς της νέας γενιάς. Επομένως σε κάθε επανάληψη ανανεώνεται αυτός ο πίνακας.

- ***NEW.Popul.chromos*** (πίνακας)
  Είναι παρόμοιος πίνακας με τον *Popul.chromos* απλά αυτός χρησιμοποιείται για την αποθήκευση της νέας γενιάς σε κάθε επανάληψη. Και αυτός ο πίνακας

ανανεώνεται σε κάθε επανάληψη και αρχικά ορίζεται με την εντολή *NEW.Popul.chromos<- matrix(nrow = population , ncol = plithos.metabl).*

- ***Popul.100*** (πίνακας)
  Είναι ένας πίνακας παρόμοιος με τους δύο παραπάνω, ο οποίος χρησιμοποιείται για την προσωρινή (σε κάθε επανάληψη) αποθήκευση του συνόλου των δύο διαδοχικών πληθυσμών των γονέων και παιδιών. Για τον λόγο αυτό έχει διπλάσιο αριθμό γραμμών από τους αντίστοιχους πίνακες της παλιάς και της νέας γενιάς. Ορίζεται κατά γραμμές με την εντολή *Popul.100<-rbind(Popul.chromos[,1:plithos.metabl] , NEW.Popul.chromos)* παίρνοντας τα χρωμοσώματα από τους δύο πίνακες.

- ***BEST.POPULATION*** (πίνακας)
  Είναι ένας πίνακας που κρατάει για κάθε γενιά τα χρωμοσώματα/άτομα του πληθυσμού σε αύξουσα σειρά ως προς τις τιμές BIC αυτών και χωρίς να έχει διπλότυπα. Και αυτός ο πίνακας ανανεώνεται σε κάθε επανάληψη και είναι και ο πίνακας που η συνάρτηση επιστρέφει ως τελικό πληθυσμό με την πρώτη γραμμή του πίνακα αυτού να αντιστοιχεί και στο καλύτερο χρωμόσωμα, επομένως και στο καλύτερο γραμμικό μοντέλο. Όπως και οι δύο προηγούμενοι πίνακες έχει και αυτός ως τελευταία στήλη τις αντίστοιχες τιμές BIC. Αρχικά ορίζεται με την βοήθεια των εντολών *unique* και *order BEST.POPULATION<-unique(Popul.chromos[order(evaluations.values),]).*

- ***fitness*** (διάνυσμα)
  Είναι η μεταβλητή η οποία χρησιμοποιείται για τον καθορισμό των πιθανοτήτων βάσει των οποίων θα επιλέγονται τα άτομα του πληθυσμού ως γονείς. Όπως έχουμε αναφέρει οι πιθανότητες αυτές εξαρτώνται από τις τιμές BIC των χρωμοσωμάτων και θα αναλυθεί παρακάτω ο ακριβής τρόπος ορισμού τους.

ΓΙΑ ΤΙΣ ΔΙΑΣΤΑΥΡΩΣΕΙΣ ΥΠΑΡΧΟΥΝ ΟΙ ΠΑΡΑΚΑΤΩ ΜΕΤΑΒΛΗΤΕΣ

- ***DIASTAYRWSH.1.POINT*** (λογική)
  Είναι η λογική μεταβλητή που δηλώνει αν πρέπει να γίνει διασταύρωση ενός σημείου. Ορίζεται μέσω του ελέγχου αν η μεταβλητή *p.type.crossover* είναι ίση με 1, με την εντολή *DIASTAYRWSH.1.POINT <- p.type.crossover==1*.

- ***DIASTAYRWSH.2.POINTS*** (λογική)
  Είναι η λογική μεταβλητή που δηλώνει αν πρέπει να γίνει διασταύρωση δύο σημείων. Ορίζεται μέσω του ελέγχου αν η μεταβλητή *p.type.crossover* είναι 2, με την εντολή *DIASTAYRWSH.2.POINTS <- p.type.crossover==2*.

- ***DIASTAYRWSH.UNIFORM*** (λογική)
  Είναι η λογική μεταβλητή που δηλώνει αν πρέπει να γίνει ομοιόμορφη διασταύρωση (με την βοήθεια μάσκας). Ορίζεται μέσω του ελέγχου αν η μεταβλητή *p.type.crossover* είναι ίση με 3, με την εντολή
  DIASTAYRWSH.2.POINTS <- p.type.crossover==3.
- ***temp*** (διάνυσμα)
  Είναι μια βοηθητική μεταβλητή για την επιλογή των γονέων από τον πληθυσμό προς διασταύρωση. Συγκεκριμένα, η ιοστή συντεταγμένη του διανύσματος αυτού είναι ένας δείκτης που δηλώνει τον αριθμό του χρωμοσώματος, από το διατεταγμένο πλήθος των χρωμοσωμάτων, που θα είναι ο ιοστός γονέας για την επόμενη γενιά. Ορίζεται με την βοήθεια της εντολής *sample*, όπου γίνεται επιλογή με επανάθεση από τον πληθυσμό και με πιθανότητες ανάλογες των τιμών που έχουν ορισθεί στην μεταβλητή *fitness temp <- sample(1:length(fitness), population,replace=T, prob=fitness)*. Το σύνολο των γονέων αποθηκεύεται στην μεταβλητή *Popul.chromos* από τον πληθυσμό *BEST.POPULATION*, όπως έχει ορισθεί παραπάνω, με την εντολή *Popul.chromos<-BEST.POPULATION[temp,]*.

- ***goneas.1*** *και* ***goneas.2*** (διανύσματα)
  Είναι οι δύο γονείς για κάθε διασταύρωση και ορίζονται με την σειρά που έχουν επιλεχθεί στον πληθυσμό *Popul.chromos*. Οπότε για την ιοστή διασταύρωση οι δύο γονείς δίνονται με τις εντολές:
  *goneas.1<-Popul.chromos[2*i-1,1:plithos.metabl]*
  *goneas.2<-Popul.chromos[2*i,1:plithos.metabl]*

- ***paidi.1*** *και* ***paidi.2*** (διανύσματα)
  Είναι τα δύο παιδιά από κάθε διασταύρωση και ορίζονται βάσει των επιλογών που έχουν δηλωθεί στην συνάρτηση, όπως είναι ο τύπος της διασταύρωσης και αν εφαρμόζεται ελιτίστικη στρατηγική (elitist) ή μετάλλαξη (mutation). Θα αναλυθεί παρακάτω ο τρόπος ορισμού τους.

- ***elit.4*** (πίνακας)
  Είναι ο πίνακας όπου καταγράφονται κατά γραμμές τα τέσσερα χρωμοσώματα των δύο γονέων και των δύο παιδιών, με σκοπό να επιλεχθούν τα δύο καλύτερα όταν εφαρμόζεται ελιτίστικη στρατηγική. Ορίζεται με την εντολή :
  elit.4<-rbind(paidi.1,paidi.2,goneas.1,goneas.2).

- ***elit.4.eval*** (διάνυσμα)
  Είναι το διάνυσμα όπου αποθηκεύονται οι τιμές του BIC για τα τέσσερα άτομα κάθε διασταύρωσης, δύο γονέων και δύο παιδιών. Χρησιμοποιείται για την περίπτωση όπου στον αλγόριθμο έχει επιλεγεί να εφαρμόζεται η ελιτίστικη στρατηγική (elitist) και έτσι είναι απαραίτητη η σύγκριση των τεσσάρων τιμών BIC των παραπάνω ατόμων. Ορίζεται με την εντολή

elit.4.eval<-rep(NA,4) όπου απλά δημιουργείται ως διάνυσμα με 4 ελλιπείς τιμές (NA). Μέσα στον αλγόριθμο και σε κάθε διασταύρωση, οι τιμές αυτές ανανεώνονται.

- ***choice.where*** (διάνυσμα)
  Είναι το "διάνυσμα" (μίας ή δύο τιμών), που δηλώνει σε ποιό/ά εσωτερικά σημείο/α θα γίνει η τομή των χρωμοσωμάτων των δύο γονέων, για την διασταύρωση του ενός ή δύο σημείων αντιστοίχως. Ορίζεται με τις εντολές:
  choice.where<- sample (1:(plithos.metabl-1) ,1,prob= rep(1,plithos.metabl-1))
  choice.where<-sort(sample(2:(plithos.metabl-1),2,prob=rep(1,plithos.metabl-2)))
  με την ίδια μεταβλητή να ανανεώνεται σε κάθε διασταύρωση είτε αφορά διασταύρωση ενός είτε δύο σημείων.

- ***miso.a1*** *και* ***miso.a2*** *και* ***miso.b1*** *και* ***miso.b2*** (διάνυσμα)
  Είναι τα αντίστοιχα μέρη των χρωμοσωμάτων των δύο γονέων για την διασταύρωση του ενός σημείου. Για τις δύο πρώτες μεταβλητές δίνουμε τις εντολές με τις οποίες ορίζονται (για τις επόμενες οι εντολές είναι ανάλογες):
  miso.a1<-goneas.1[1:choice.where]
  miso.a2<-goneas.1[(choice.where+1):plithos.metabl].

- ***mask*** (διάνυσμα)
  Είναι ένα διάνυσμα με μήκος όσο μήκος ενός χρωμοσώματος και συντεταγμένες που παίρνουν τιμές 0 ή 1 (δυαδικές), και οι οποίες δηλώνουν ως TRUE και FALSE, αν το χρωμόσωμα που θα προκύψει από την διασταύρωση θα λαμβάνει την τιμή του πρώτου γονέα ή όχι. Ορίζεται τυχαία και με επανάθεση με την βοήθεια της εντολής *sample*:
  mask<-sample(c(0,1),plithos.metabl , rep=TRUE).

Τέλος, ο αλγόριθμος χρησιμοποιεί και βοηθητικές μεταβλητές ως δείκτες για τις επαναληπτικές διαδικασίες (*FOR*) που τρέχει ο αλγόριθμος. Τέτοιες βοηθητικές μεταβλητές είναι οι ***epan*** και ***i.***

<u>Αναλυτικά οι εντολές του αλγορίθμου</u>

Ο αλγόριθμος ξεκινά με την ονομασία του και το όρισμά του, τα οποία έχουν αναλυθεί στα προηγούμενα:

```
genetikos.alg.lm<-function( X, Y, population=50, epanalipseis=50,
evaluation=eval.fun, p. crossover=1, p.type.crossover=1, pososto.palaiwn=0.20 ,
elitist=FALSE, p.mutation=0.01, analogia.0.1=c(1,1) ) {
```

Ακολουθεί ένας έλεγχος για τις σωστές διαστάσεις των δεδομένων, και συγκεκριμένα του πίνακα X και του διανύσματος Y. Εάν οι διαστάσεις ταιριάζουν τότε εκτελούνται

οι λοιπές εντολές του αλγορίθμου, αλλιώς τυπώνεται η αντίστοιχη εντολή λάθους και ο αλγόριθμος τερματίζει. Είναι ένας τρόπος για να αποφύγουμε λάθη που θα οδηγούσαν τον αλγόριθμο σε λανθασμένη εκτέλεση και αντιστοίχως αποτελέσματα ή ακόμα και σε μη εκτέλεσή του.

```
if ( (dim(X)[1]==length(Y)) ) {

                    …

                    …

} else {print( "LATHOS DIASTASH DEDOMENWN" )}

}  # END
```

Ακολουθούν οι εντολές που ορίζουν αρχικώς τις απαραίτητες μεταβλητές του αλγορίθμου, σχηματίζουν τον αρχικό πληθυσμό και υπολογίζουν τις τιμές της αντικειμενικής συνάρτησης (τιμές BIC) που είναι αναγκαίες για την σύγκριση των χρωμοσωμάτων, όπως και ορίζουν τις λογικές μεταβλητές για τον τύπο της διασταύρωσης.

```
plithos.metabl<-dim(X)[2]
best.evaluations<-rep(NA,epanalipseis)
evaluations.values<-rep(NA,population)
new.evaluations.values<-rep(NA,2*population)
elit.4.eval<-rep(NA,4)
Popul.chromos <- matrix(nrow = population , ncol = plithos.metabl)
## ARXIKOS PLITHISMOS  ME ANTISTOIXH ANALOGIA 0-1
for (i in 1:population ) {
    Popul.chromos[i,]<- sample( c(rep(c(0,1),analogia.0.1)) , plithos.metabl , rep=TRUE )
}
## TIMES BIC GIA TON PLITHISMO
for (i in 1:population ) {
    evaluations.values[i]<-eval.fun(Popul.chromos[i,])
}
Popul.chromos<-cbind(Popul.chromos,evaluations.values )
BEST.POPULATION<-unique(Popul.chromos[order(evaluations.values),])
print("ARXIKOS Popul.chromos")
print(Popul.chromos) # ME TA BIC TELEYTAIA STHLH
##EPANALHPSEIS
NEW.Popul.chromos<- matrix(nrow = population , ncol = plithos.metabl)
DIASTAYRWSH.1.POINT   <- p.type.crossover==1
DIASTAYRWSH.2.POINTS <- p.type.crossover==2
DIASTAYRWSH.UNIFORM<- p.type.crossover==3
```

Έως αυτό το σημείο ο αλγόριθμος έχει τον αρχικό πληθυσμό και τις απαραίτητες μεταβλητές για να ξεκινήσει την επαναληπτική διαδικασία των διασταυρώσεων και παραγωγής κάθε νέας γενιάς, ενώ έχει τυπώσει και τον αρχικό πληθυσμό στην κονσόλα του χρήστη. Εδώ υπάρχει και το μεγαλύτερο μέρος των εντολών του αλγορίθμου και εκτελείται όσες φορές έχει δηλωθεί από τον χρήστη.

```
for (epan in 1:epanalipseis) {
        …
        …

} # TELOS EPANALHPSEWN
```

Αρχικά μέσα στις εντολές που αφορούν τις επαναλήψεις, ορίζονται οι πιθανότητες επιλογής των γονέων και το σύνολο των γονέων ως *Popul.chromos*. Οι πιθανότητες ορίζονται μέσω της μεταβλητής *fitness* και όπως έχουμε αναφέρει βασίζονται στις τιμές BIC των χρωμοσωμάτων. Συγκεκριμένα, γίνεται ένα είδος κανονικοποίησης των τιμών αυτών και καταχωρούνται στην μεταβλητή *fitness*. Δηλαδή, για κάθε τιμή BIC υπολογίζεται η διαφορά της με την μικρότερη τιμή BIC του συνόλου των χρωμοσωμάτων, και διαιρείται με το εύρος των τιμών BIC του συνόλου. Οι τιμές της

μεταβλητής *fitness* που προκύπτουν είναι αυτές που θα δώσουν τις πιθανότητες επιλογής κάθε γονέα (να σημειωθεί πως έχουν χρησιμοποιηθεί δύο μεταβλητές *arithmitis* και *paronomastis* για την πιο εύκολη γραφή και καταχώρηση των τιμών της μεταβλητής *fitness*).

```
arithmitis<-BEST.POPULATION[,plithos.metabl+1]-
        min(BEST.POPULATION[,plithos.metabl+1])
paronomastis<-max(BEST.POPULATION[,plithos.metabl+1])-
        min(BEST.POPULATION[,plithos.metabl+1])
fitness<-1-arithmitis/paronomastis
 temp<-sample(1:length(fitness),population,replace=T, prob=fitness)
 Popul.chromos<-BEST.POPULATION[temp,]
```

Η μεταβλητή *temp* δηλώνει όπως έχουμε αναφέρει τον γονέα από το σύνολο του πληθυσμού *BEST.POPULATION* και η επιλογή των αντίστοιχων χρωμοσωμάτων καταχωρείται στην μεταβλητή *Popul.chromos*. Τα άτομα/χρωμοσώματα του πληθυσμού αυτού αποτελούν τους γονείς της νέας γενιάς και οι διασταυρώσεις θα γίνονται ζευγαρώνοντας τα άτομα κατά σειρά, δηλαδή το πρώτο με το δεύτερο, το τρίτο με το τέταρτο, το πέμπτο με το έκτο, κτλ. Να σημειωθεί πως το πλήθος των ατόμων/χρωμοσωμάτων του πληθυσμού *Popul.chromos* είναι ακριβώς όσο και ο πληθυσμός που έχει δηλωθεί στο όρισμα της συνάρτησης, ενώ το πλήθος τον ατόμων/χρωμοσωμάτων του πληθυσμού *BEST.POPULATION* ενδέχεται να είναι μικρότερο, καθώς, εκεί κρατούνται μόνο τα καλύτερα άτομα του πληθυσμού με τα χρωμοσώματα να μην είναι πολλαπλά (δεν υπάρχουν δύο ίδια χρωμοσώματα στον πληθυσμό αυτό).

Έτσι, οι διασταυρώσεις γίνονται με την επαναληπτική εντολή FOR όπως υπάρχει στον αλγόριθμο:

```
for (i in 1:(population/2) ) {  # ARXH "PARAGOGHS" THS NEAS GENIAS
                # EPILOGES ANAPARAGWGHS
goneas.1<-Popul.chromos[2*i-1,1:plithos.metabl]       # ORISMOS TWN 2 GONEWN
goneas.2<-Popul.chromos[2*i,1:plithos.metabl]         # ORISMOS TWN 2 GONEWN
                        ...
                        ...

}  # TELOS "PARAGOGHS" THS NEAS GENIAS
```

Οι διασταυρώσεις γίνονται με έναν από τους τρεις τύπους διασταυρώσεων όπως έχει ορισθεί στο όρισμα της συνάρτησης και για αυτό υπάρχουν οι εντολές *if* και *else if* που κάνουν τον αντίστοιχο διαχωρισμό βάσει των *DIASTAYRWSH.1.POINT*, *DIASTAYRWSH.2.POINTS* και *DIASTAYRWSH.UNIFORM* μεταβλητών. Και για τους τρεις τύπους όπως είναι φυσικό, οι γονείς είναι προκαθορισμένοι όπως φαίνεται και από τις δύο πρώτες εντολές μέσα στην επαναληπτική διαδικασία της παραγωγής της νέας γενιάς.

Παρακάτω φαίνονται οι εντολές που κάνουν τον διαχωρισμό για τον τύπο της διασταύρωσης:

```
if (DIASTAYRWSH.1.POINT){                    # 1-point
        …
} else if (DIASTAYRWSH.2.POINTS){            # 2-points
        …
} else if (DIASTAYRWSH.UNIFORM ){            # uniform – mask
        …
} # TELOS i-OSTHS DIASTAYRWSHS
```

Ακολουθούν οι εντολές για την διασταύρωση, όπου πρώτα ελέγχεται με την εντολή *if* αν θα πραγματοποιηθεί η συγκεκριμένη διασταύρωση, και με την βοήθεια της εντολής *runif(1)* παράγεται μία τιμή στο διάστημα (0,1) και συγκρίνεται με την πιθανότητα διασταύρωσης *p.crossover*. Αν είναι μικρότερη αυτής τότε πραγματοποιείται η διασταύρωση, ειδάλλως τυπικά ως απόγονοι/παιδιά ορίζονται οι γονείς (δεν πραγματοποιείται διασταύρωση). Αναλυτικά οι εντολές για την διασταύρωση ενός σημείου είναι οι παρακάτω, όπου γίνεται η επιλογή του εσωτερικού σημείου του χρωμοσώματος για την διασταύρωση και καταχωρείται στην μεταβλητή *choice.where* και με την βοήθεια των μεταβλητών *miso.a1 και miso.a2 και miso.b1 και miso.b2*, που αποτελούν και τα τμήματα των δύο γονέων που εναλλάσσονται, παράγονται τα δύο νέα χρωμοσώματα *paidi.1* και *paidi.2*.

```
if (runif(1)<=p.crossover) {
      choice.where<-sample( 1:(plithos.metabl-1), 1 , prob= rep(1,plithos.metabl-1))
            # EPILOGH SHMEIOY DIASTAYRWSHS
       miso.a1<-goneas.1[1:choice.where]
       miso.a2<-goneas.1[(choice.where+1):plithos.metabl]
       miso.b1<-goneas.2[1:choice.where]
       miso.b2<-goneas.2[(choice.where+1):plithos.metabl]
       paidi.1<-c(miso.a1,miso.b2)
       paidi.2<-c(miso.b1,miso.a2)
} else {
  paidi.1<-goneas.1
  paidi.2<-goneas.2
}
```

Αντιστοίχως, οι εντολές για την διασταύρωση δύο σημείων είναι οι παρακάτω, όπου γίνεται η επιλογή των δύο εσωτερικών σημείων του χρωμοσώματος όπου θα γίνει η τομή για την αλλαγή των τμημάτων των δύο χρωμοσωμάτων όπως ορίζει η διασταύρωση δύο σημείων. Έτσι, παράγονται τα δύο νέα χρωμοσώματα *paidi.1* και *paidi.2*.

```
if (runif(1)<=p.crossover) {
     choice.where<-sort( sample( 2:(plithos.metabl-1), 2 , prob= rep(1,plithos.metabl-2)) )
                    # EPILOGH SHMEIOY DIASTAYRWSHS
     paidi.1<-goneas.1
    paidi.1[choice.where[1]:choice.where[2]]<-goneas.2[choice.where[1]:choice.where[2]]
    paidi.2<-goneas.2
    paidi.2[choice.where[1]:choice.where[2]]<-goneas.1[choice.where[1]:choice.where[2]]
  } else {
   paidi.1<-goneas.1
   paidi.2<-goneas.2
  }
```

Τέλος, για το τμήμα των διασταυρώσεων, οι εντολές που αφορούν την ομοιόμορφη (*UNIFORM)* διασταύρωση δίνονται παρακάτω, όπου δημιουργείται αρχικώς η "μάσκα" διασταύρωσης με την εντολή *sample* ως δείγμα με επανάθεση 1 και 0, τα οποία δηλώνουν την επιλογή της αντίστοιχης συντεταγμένης του χρωμοσώματος του πρώτου γονέα (άσσος) ή του δεύτερου γονέα (μηδενικό). Ακολούθως, με την βοήθεια της μάσκας επιλέγονται οι συντεταγμένες του πρώτου γονέα και του δεύτερου γονέα ως γινόμενο του χρωμοσώματος με τα στοιχεία της μάσκας και τα συμπληρωματικά αυτής. Συγκεκριμένα, η μάσκα λαμβάνεται εμμέσως ως πολλαπλή λογική μεταβλητή, καθώς πολλαπλασιάζεται κάθε άσσος της μάσκας με την αντίστοιχη συντεταγμένη του γονέα και ορίζεται ως συντεταγμένη του νέου χρωμοσώματος-παιδιού, δηλαδή ως ύπαρξη της συντεταγμένης αυτής στο παιδί, ενώ κάθε μηδενικό πολλαπλασιάζεται με την αντίστοιχη συντεταγμένη του γονέα και ορίζει ως μηδενική την αντίστοιχη συντεταγμένη του νέου χρωμοσώματος-παιδιού, δηλαδή ως μη ύπαρξη αυτής στο χρωμόσωμα. Ομοίως, ορίζεται και το δεύτερο παιδί με αντιμετάθεση των γονέων.

```
        mask<-sample(c(0,1),plithos.metabl , rep=TRUE)
        paidi.1<-goneas.1*mask+goneas.2*(!mask)
        paidi.2<-goneas.2*mask+goneas.1*(!mask)
        } # TELOS i-OSTHS DIASTAYRWSHS
```

Το επόμενο μέρος του αλγορίθμου αφορά την επιλογή για ελιτίστικη στρατηγική (*elitist)* ή μετάλλαξη (*mutation)*. Η επιλογή είναι πάντα να εφαρμοστεί είτε ελιτίστικη στρατηγική (*elitist)* είτε μετάλλαξη(*mutation)* και αυτό ορίζεται μέσω της λογικής μεταβλητής του ορίσματος *elitist*. Εάν αυτή είναι TRUE τότε εκτελούνται οι αντίστοιχες εντολές που επιλέγουν ως νέα μέλη της νέας γενιάς τα καλύτερα δύο χρωμοσώματα από την τετράδα των δύο γονέων και των δύο παιδιών.

Εάν η μεταβλητή *elitist* είναι FALSE τότε για τα δύο παιδιά εφαρμόζεται μετάλλαξη μέσω της συνάρτησης ***Mutation*** που έχει ορισθεί έξω από την συνάρτηση ***genetikos.alg.lm***. Οι δύο τελευταίες εντολές απλά καταχωρούν τα δύο παιδιά στις αντίστοιχες θέσεις του νέου προσωρινού πληθυσμού. Να σημειωθεί πως αυτό γίνεται μετά από τους τρεις τύπους διασταυρώσεων και μετά τις επιλογές για ελιτίστικη στρατηγική (*elitist)* ή μετάλλαξη (*mutation)* δηλαδή γίνεται μία φορά είτε έχει επιλεχθεί ένας από τους τρεις τύπους διασταύρωσης είτε βελτίωση/μετάλλαξη στα χρωμοσώματα.

```
if (elitist) {
  elit.4<-rbind(paidi.1,paidi.2,goneas.1,goneas.2)
  elit.4.eval<-c( eval.fun(paidi.1),eval.fun(paidi.2),eval.fun(goneas.1),eval.fun(goneas.2)
)
  paidi.1<-elit.4[order(elit.4.eval)[1], ]
  paidi.2<-elit.4[order(elit.4.eval)[2], ]
        # EPILOGH KALYTEROY ZEYGARIOY APO TOYS 4 (2 GONEIS + 2 PAIDIA)
} else {
   paidi.1<-Mutation(paidi.1,p.mutation)      # MUTATION GIA TOYS 2 APOGONOYS
   paidi.2<-Mutation(paidi.2,p.mutation)          # MONO AN DEN ISXYEI ELITIST
}

NEW.Popul.chromos[2*i-1,]<-paidi.1
NEW.Popul.chromos[2*i,]<-paidi.2
```

Καθώς, έχει τελειώσει η διαδικασία των διασταυρώσεων, απομένει για τον αλγόριθμο να συγκεντρώσει τους πληθυσμούς της παλιάς και νέας γενιάς και να κρατήσει τα καλύτερα άτομα από το σύνολο αυτό. Ακόμα πρέπει να ανανεωθεί μέρος του πληθυσμού με νέα τυχαία χρωμοσώματα. Το μέρος του πληθυσμού που θα κρατήσει έχει δηλωθεί στο όρισμα της συνάρτησης με την μεταβλητή *pososto.palaiwn*. Συγκεκριμένα, δημιουργείται ένας νέος πίνακας στην μεταβλητή *Popul.100*, όπου συγκεντρώνονται όλα τα παλαιά χρωμοσώματα της προηγούμενης γενιάς (*BEST.POPULATION*), μαζί με τα νέα χρωμοσώματα της νέας γενιάς (παιδιά τα οποία και λαμβάνονται από την μεταβλητή πίνακα *NEW.Popul.chromos*). Ακολούθως, υπολογίζονται οι τιμές *BIC* των χρωμοσωμάτων του συνόλου αυτού με σκοπό να διαταχθούν και οι τιμές αυτές αποθηκεύονται στην μεταβλητή *new.evaluations.values*. Οι εντολές για την συγκέντρωση και επιλογή των καλύτερων χρωμοσωμάτων είναι οι παρακάτω, όπου υπολογίζονται και οι τιμές BIC των χρωμοσωμάτων στην μεταβλητή *new.evaluations.values* :

```
Popul.100<-rbind(BEST.POPULATION[,1:plithos.metabl] , NEW.Popul.chromos )
                  # TIMES BIC GIA TON NEO SYNOLIKO PLITHISMO.100
for (i in 1: ( dim(Popul.100)[1] ) ) {
          new.evaluations.values[i]<-eval.fun(Popul.100[i,])
}
                  # ANTIKASTASTASH NEAS GENIAS
Popul.chromos[,1:plithos.metabl]<-
Popul.100[order(new.evaluations.values)[1:population],]
```

Οι εντολές για την κράτηση του δεδηλωμένου ποσοστού στην νέα γενιά και την αντικατάσταση των υπολοίπων χρωμοσωμάτων με καινούργια τυχαία φαίνονται παρακάτω με την βοήθεια της εντολής *sample* όπως έγινε και για την παραγωγή του αρχικού πληθυσμού. Επίσης, υπάρχουν και οι εντολές για τον υπολογισμό και την καταχώρηση των τιμών BIC του πληθυσμού της νέας γενιάς στην μεταβλητή *evaluations.values*, καθώς υπάρχουν και καινούρια μέλη (τα νέα χρωμοσώματα που

παρήχθησαν τυχαία), όπως και οι εντολές για την καταχώρηση στην μεταβλητή *BEST.POPULATION* του πληθυσμού ως σύνολο χρωμοσωμάτων χωρίς επαναλαμβανόμενα μέλη.
Τέλος, καθώς εδώ είναι και το τέλος των εντολών κάθε επανάληψης και έχει προκύψει η νέα γενιά, ο αλγόριθμος καταχωρεί την μικρότερη τιμή BIC του νέου πληθυσμού στο διάνυσμα *best.evaluations* και τυπώνει στην κονσόλα του χρήστη (με την εντολή *print(BEST.POPULATION[1:5,])*) τα πέντε καλύτερα χρωμοσώματα που προέκυψαν με τις αντίστοιχες τιμές BIC να βρίσκονται ως τελευταίο στοιχείο κάθε διανύσματος (ή αλλιώς στην τελευταία στήλη του πίνακα 5x(p+1)).

```
        # ANTIKATASTASH TOY ..% TOY PLITHISMOY ME TYXAIA XROMOSOMATA
for (i in (population*pososto.palaiwn+1):population ) {
  Popul.chromos[i,1:plithos.metabl]<-
                        sample(c(rep(c(0,1),analogia.0.1)),plithos.metabl , rep=TRUE )
}
        #TIMES BIC GIA TON NEO 100% PLITHISMO
for (i in 1:population ) {
   evaluations.values[i]<-eval.fun(Popul.chromos[i,1:plithos.metabl])
}
Popul.chromos[,plithos.metabl+1]<-evaluations.values
        # TIMES BIC GIA THN NEA GENIA WS TELEYTAIA STHLH
BEST.POPULATION<-unique(Popul.chromos[order(evaluations.values),])
best.evaluations[epan]<-min(evaluations.values)
print("BEST UNIQUE POPULATION 1:5")
print(BEST.POPULATION[1:5,])   # TA 5 KALYTERA MONTELA

 } # TELOS EPANALHPSEWN
```

Οι τελευταίες εντολές του αλγορίθμου είναι αυτές που απλά τυπώνουν στην κονσόλα του χρήστη τις τιμές BIC *best.evaluations* για τα καλύτερα χρωμοσώματα κατά τις επαναλήψεις, τον τελικό πληθυσμό *BEST.POPULATION* των χρωμοσωμάτων, όπως και τον επιστρέφουν έξω από την συνάρτηση. Πάντα ο πίνακας *BEST.POPULATION* έχει ως τελευταία στήλη τις τιμές των BIC των αντίστοιχων χρωμοσωμάτων.

```
print("best.evaluations")
print(best.evaluations)
print("TELIKOS BEST.POPULATION")
print(BEST.POPULATION)
return(BEST.POPULATION)
} else {print( "LATHOS DIASTASH DEDOMENWN" )}
}  # END
```

ΚΕΦΑΛΑΙΟ 5
ΠΡΟΣΟΜΟΙΩΜΕΝΑ ΔΕΔΟΜΕΝΑ

5.1 ΕΛΕΓΧΟΣ ΑΛΓΟΡΙΘΜΟΥ

Όπως έχουμε αναφέρει και στο προηγούμενο κεφάλαιο, είναι δόκιμο να ελέγχεται ένας αλγόριθμος ως προς την σωστή λειτουργία του σε δύο βασικά θέματα. Πρώτον, για την σωστή συντακτική του λειτουργία, με την έννοια να εκτελούνται οι εντολές του αλγόριθμου χωρίς να δημιουργείται πρόβλημα όπως π.χ. να καλείται επ'αόριστον κάποια εντολή, ή να ορίζεται κάπου η προσπέλαση ενός διανύσματος ή πίνακα σε στοιχείο εκτός των διαστάσεων αυτού, κ.α. Κατά δεύτερον, θα πρέπει να ελεγχθεί η σωστή/λογική λειτουργία του αλγορίθμου, με την έννοια πως αυτός θα πρέπει να δίνει ικανοποιητικά αποτελέσματα σύμφωνα με τον σκοπό για τον οποίο δημιουργήθηκε.

Για τον έλεγχο της λειτουργίας του αλγορίθμου και των ικανοποιητικών αποτελεσμάτων που αυτός εξάγει, ακολουθείται ο έλεγχος μέσω προσομοιωμένων δεδομένων. Συγκεκριμένα, παράγονται τυχαία δεδομένα για τα οποία γνωρίζουμε τον τρόπο δημιουργίας τους και την μεταξύ τους εξάρτηση, και επομένως γνωρίζουμε το αποτέλεσμα που θα πρέπει να εξάγει ο αλγόριθμος. Καλώντας τον αλγόριθμο με τα συγκεκριμένα προσομοιωμένα δεδομένα ελέγχουμε αν τα αποτελέσματα που δίνει ο αλγόριθμος είναι αυτά που πρέπει. Λαμβάνοντας υπόψη μας ότι τα δεδομένα μας περιέχουν m=15 τυχαίες επεξηγηματικές μεταβλητές (από n=50 παρατηρήσεις η κάθε μία) και μία εξαρτημένη μεταβλητή Y για την οποία έχουμε ορίσει την σχέση της με 3 τυχαίες μεταβλητές από τις 15, και συγκεκριμένα τις X_1 , X_7 και X_{12} αναμένουμε τα σωστά αποτελέσματα του αλγορίθμου να είναι τέτοια ώστε το καλύτερο χρωμόσωμα να έχει στις θέσεις 1, 7 και 12 άσσους και στις άλλες θέσεις μηδενικά. Με αυτόν τον τρόπο, είναι δυνατός ο έλεγχος της απόδοσης του αλγορίθμου.

Υπάρχουν βέβαια και κάποια ζητήματα για την απόδοση του αλγορίθμου, και ένα είναι ο παράγοντας της τύχης, καθώς τα παραγόμενα δεδομένα είναι τυχαία και αυτό μπορεί να αλλοιώσει τα αποτελέσματα του αλγόριθμου, ανεξαρτήτως της επίδρασης των παραμέτρων που χρησιμοποιούνται στον αλγόριθμο. Αυτό μπορεί να αποφευχθεί μειώνοντας την τυπική απόκλιση που ορίζεται για την παραγωγή των προσομοιωμένων δεδομένων. Υπενθυμίζουμε τον τρόπο παραγωγής των δεδομένων:

Συγκεκριμένα έχουμε m=15 τυχαίες μεταβλητές από n=50 παρατηρήσεις η καθεμία που βασίζονται στην κανονική κατανομή με μέση τιμή μηδέν και τυπική απόκλιση ένα. Αυτές οι 15 τυχαίες μεταβλητές παίζουν το ρόλο των ανεξάρτητων τυχαίων μεταβλητών για το γραμμικό μοντέλο *Xi, i=1,2,…,m.*

Επίσης προσομοιώσαμε και 50 παρατηρήσεις από μια μεταβλητή Y η οποία θα είναι και η εξαρτημένη μεταβλητή. Η μεταβλητή Y βασίζεται μόνο σε τρεις από τις 15 τυχαίες επεξηγηματικές μεταβλητές σύμφωνα με τον παρακάτω τύπο :

$$\mathbf{Y_j=3+8*X_{1,j}+4*X_{7,j}-4*X_{12,j}+e_j}\ ,\ j=1,2,\ldots,50.\ \text{Με}\ e_j \sim N(0,s^2)$$

Το χρωμόσωμα που αντιστοιχεί στο βέλτιστο γραμμικό μοντέλο για τα παραπάνω δεδομένα είναι το (**1**,0,0,0,0,0,**1**,0,0,0,0,**1**,0,0,0). Για τον έλεγχό μας η τυπική απόκλιση ορίστηκε στο 1.5.

Ένας πρώτος έλεγχος της απόδοσης του αλγορίθμου είναι με την παρακάτω εντολή εκτέλεσης για πληθυσμό μεγέθους 30 και 20 επαναλήψεις (γενιές), με τύπο διασταύρωσης ενός σημείου και ελιτίστικη στρατηγική ως βασικές παραμέτρους:

```
genetikos.alg.lm(X,Y,population=30,epanalipseis=20,p.crossover=1,
pososto.palaiwn=0.2 , p.type.crossover=1, elitist=TRUE,p.mutation=0.01)
```

Τα αποτελέσματα του αλγορίθμου είναι για τον τελικό πληθυσμό χρωμοσωμάτων και τις αντίστοιχες τιμές *BIC* (παρατηρούμε πως εμφανίζονται μόνον 25 διανύσματα καθώς ο αλγόριθμος κρατά σε κάθε γενιά χρωμοσώματα τα οποία δεν επαναλαμβάνονται) :

"**ΤΕΛΙΚΟΣ ΠΛΗΘΥΣΜΟΣ**"	**ΤΙΜΗ BIC**
1 0 0 0 0 0 1 0 0 0 0 1 0 0 0	**191.1076**
1 1 1 0 0 0 1 1 1 0 1 1 0 0 0	204.7152
1 1 0 0 0 1 1 1 1 1 1 1 1 1 1	214.5425
1 0 0 0 0 0 0 0 1 1 0 1 0 0 1	304.6478
1 0 0 1 0 0 0 1 0 1 0 1 0 1 1	308.7009
1 0 1 0 1 0 1 0 1 0 0 0 0 1 0	310.7034
1 0 1 0 0 0 1 0 1 1 1 0 0 1 1	314.7083
1 1 1 1 0 1 0 0 1 1 0 1 1 1 1	318.5099
1 1 0 1 1 0 0 1 1 1 0 1 0 1 0	319.3320
1 0 1 1 1 1 0 1 0 1 0 1 1 1 0	322.5297
1 1 0 0 0 0 0 1 0 0 1 0 1 0 0	335.4933
1 0 1 1 0 0 0 1 0 1 1 0 0 1 0	347.0638
1 1 1 0 1 1 0 1 0 1 1 0 1 1 0	349.8155
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0	374.8068
0 1 0 1 0 0 0 0 1 0 0 1 1 1 0	374.8118
0 1 0 1 0 1 0 1 0 0 0 1 1 1 0	376.6028
0 0 0 1 1 0 0 0 1 1 0 1 1 0 0	376.7510
0 1 0 0 0 0 0 1 1 0 0 0 0 0 1	379.1098
0 1 0 1 1 0 0 1 1 1 0 1 0 1 0	379.2309
0 1 0 0 1 1 1 1 1 0 0 1 1 0 1	380.0591
0 1 0 0 0 0 0 1 0 1 1 0 0 0 1	381.8638
0 0 0 0 1 1 0 1 1 1 1 1 1 1 1	387.5403
0 1 0 0 0 0 1 1 1 1 1 0 0 0 1	388.8810
0 1 0 1 0 1 1 1 0 0 0 0 1 1 1	390.8948
0 0 1 1 0 1 1 1 0 1 0 0 1 0 1	392.9548

Ως καλύτερο χρωμόσωμα ο αλγόριθμος βρήκε το βέλτιστο χρωμόσωμα που αναμέναμε, και το οποίο αντιστοιχεί στο γραμμικό μοντέλο με τις μεταβλητές X_1, X_7 και X_{12}. Η τιμή BIC είναι η μικρότερη από όλες του συνόλου του πληθυσμού και ίση εδώ με 191.1076. Από αυτά τα αποτελέσματα και μόνο φαίνεται πως ο αλγόριθμος δουλεύει ικανοποιητικά καθώς βρίσκει το βέλτιστο γραμμικό μοντέλο. Με τα αποτελέσματα ο αλγόριθμος δίνει και τις τιμές BIC των καλύτερων χρωμοσωμάτων κατά την διάρκεια της εκτέλεσης της διαδικασίας εύρεσης του καλύτερου χρωμοσώματος και αυτές είναι από την $1^η$ έως την $20^η$ γενιά (κατά γραμμές), όπου φαίνεται πως από την $16^η$ γενιά έχει βρεθεί η καλύτερη λύση:

"ΚΑΛΥΤΕΡΕΣ ΤΙΜΕΣ BIC ΑΠΟ ΚΑΘΕ ΓΕΝΙΑ"

195.7873	195.7873	195.7873	195.7873	193.5781
192.5708	192.4313	192.4313	192.4313	192.4313
192.4313	192.4313	192.4313	192.4313	192.4313
191.1076	**191.1076**	**191.1076**	**191.1076**	**191.1076**

Ένας δεύτερος έλεγχος της απόδοσης του αλγορίθμου με ίδιες παραμέτρους αλλά με τύπο διασταύρωσης δύο σημείων:

```
genetikos.alg.lm(X,Y,population=30,epanalipseis=20,p.crossover=1,
pososto.palaiwn=0.2 , p.type.crossover=2, elitist=TRUE,p.mutation=0.01)
```

Τα αποτελέσματα του αλγορίθμου (για τα 3 πρώτα χρωμοσώματα) είναι :

"ΤΕΛΙΚΟΣ ΠΛΗΘΥΣΜΟΣ"	**ΤΙΜΗ BIC**
1 0 0 0 0 0 1 0 0 0 0 1 0 0 0	**191.1076**
1 0 0 0 0 0 1 1 0 0 0 1 1 1 1	198.53411
1 1 0 0 0 0 1 1 1 1 1 1 0 0 1	207.6543
…	
…	

"ΚΑΛΥΤΕΡΕΣ ΤΙΜΕΣ BIC ΑΠΟ ΚΑΘΕ ΓΕΝΙΑ"

202.0279	198.8525	198.4291	196.2955	194.5753
194.5753	194.5753	200.8089	198.3691	195.2784
195.2511	195.2784	195.2784	195.2784	**191.1076**
191.1076	**191.1076**	**191.1076**	**191.1076**	**191.1076**

όπου φαίνεται πως από την $15^η$ γενιά έχει βρεθεί η καλύτερη λύση.

Παρόμοια βγαίνουν και τα αποτελέσματα με διασταύρωση τύπου *Uniform*.

5.2 ΣΥΝΔΥΑΣΜΟΙ ΠΑΡΑΜΕΤΡΩΝ ΑΛΓΟΡΙΘΜΟΥ

Για τον καλύτερο έλεγχο του αλγορίθμου και της απόδοσης αυτού δοκιμάσαμε κάποιους συνδυασμούς των τιμών των παραμέτρων του αλγορίθμου για τα προσομοιωμένα δεδομένα. Ελέγξαμε εν συνεχεία και το ποσοστό των φορών που ο αλγόριθμος βρίσκει την βέλτιστη λύση αλλά και τον χρόνο που χρειάζεται για την εκτέλεσή του.
Για τον χρόνο εκτέλεσης του αλγόριθμου χρησιμοποιήθηκε η εντολή ***system.time(…)*** η οποία δίνει τον χρόνο εκτέλεσης της εντολής που δίδεται στο όρισμά της.

Επίσης, για μια πιο αξιόπιστη σύγκριση των αποτελεσμάτων, κάθε συνδυασμός δοκιμάστηκε 10 φορές και οι αντίστοιχοι χρόνοι χρησιμοποιήθηκαν για να βγει ο μέσος χρόνος εκτέλεσης, και αντιστοίχως μετρήθηκε το πλήθος των περιπτώσεων (από τις 10), πάντα με τα ίδια αρχικά προσομοιωμένα δεδομένα, που ο αλγόριθμος βρήκε την βέλτιστη λύση (βέλτιστο γραμμικό μοντέλο). Για παράδειγμα, εάν για κάποιον συνδυασμό έχουμε στις 10 δοκιμές, 8 επιτυχείς ευρέσεις της βέλτιστης λύσης αποδίδουμε στον αλγόριθμο για τον συγκεκριμένο συνδυασμό 80% επιτυχία. Ενώ για τον χρόνο εκτέλεσης ορίζουμε τον αριθμητικό μέσο των χρόνων εκτέλεσης που μετρήθηκε στον αλγόριθμο για τις 10 εκτελέσεις αυτού.

Μετά από αρκετές δοκιμές και παρατήρηση των σχετικών αποτελεσμάτων, παρατίθενται στον ***Πίνακα 5.1*** διάφοροι συνδυασμοί των τιμών των παραμέτρων και τα αντίστοιχα αποτελέσματα. Παρατηρήθηκε πως για τα προσομοιωμένα δεδομένα των 50 παρατηρήσεων με τις 15 τυχαίες μεταβλητές, είναι απαραίτητο να ορίσουμε πληθυσμό τουλάχιστον 30, δηλαδή με έναν πρόχειρο υπολογισμό διπλάσιο από το πλήθος των τυχαίων μεταβλητών. Βέβαια όσο μεγαλύτερος είναι ο πληθυσμός τόσο πιο επιτυχής θα είναι και η εύρεση της βέλτιστης λύσης, αλλά θα αυξάνεται και ο χρόνος εκτέλεσης του αλγορίθμου. Αντιστοίχως, όπως και ήταν αναμενόμενο, παρατηρήθηκε πως όταν ήταν μεγαλύτερος ο αριθμός των επαναλήψεων η εύρεση της βέλτιστης λύσης ήταν πιο εύκολη, οπότε και ο αλγόριθμος πιο αποτελεσματικός. Αυτό είναι αναμενόμενο γιατί ο γενετικός αλγόριθμος σε κάθε επανάληψη/παραγωγή νέας γενιάς αξιοποιεί τα υπάρχοντα χρωμοσώματα και είτε μέσω αποτελεσματικών διασταυρώσεων, είτε μέσω μεταλλάξεων είτε ακόμα και μέσω της τυχαίας παραγωγής νέων χρωμοσωμάτων βρίσκει καλύτερα χρωμοσώματα, δηλαδή καλύτερα γραμμικά μοντέλα. Αυτό σημαίνει πως όσο μεγαλύτερος είναι ο αριθμός των επαναλήψεων τόσο ο αλγόριθμος έχει την προοπτική να βρει την καλύτερη λύση καθώς συνεχώς βελτιώνει την υπάρχουσα. Ακόμα, παρατηρήθηκε πως όταν έχει επιλεχθεί να γίνεται *elitist* στο όρισμα του αλγορίθμου, ο χρόνος εκτέλεσης αυξάνεται. Αυτό οφείλεται στην σύγκριση των χρωμοσωμάτων γονέων και παιδιών σε κάθε διασταύρωση, το οποίο είναι χρονοβόρο, όπως και στην αντικατάσταση των 2 παιδιών με το ζεύγος των καλύτερων χρωμοσωμάτων που προκύπτουν από την σύγκριση. Τέλος παρατηρήθηκε πως ο τύπος διασταύρωσης *UNIFORM* είναι λίγο πιο αποτελεσματικός σε σύγκριση με τους τύπους διασταύρωσης ενός ή δύο σημείων.

Τα αποτελέσματα κάποιων ενδεικτικών συνδυασμών παρατίθενται στον παρακάτω πίνακα:

Πληθυσμός	Επανα-λήψεις	Τύπος Δ/ρωσης	Ελιτισμός	Ποσοστό Παλαιών	Χρόνος (*) Εκτέλεσης	Ποσοστό Εύρεσης
30	30	2	ΟΧΙ	20%	10,737	100%
30	30	3	ΟΧΙ	20%	10,808	100%
30	30	1	ΟΧΙ	20%	10,852	100%
50	20	3	ΟΧΙ	20%	11,890	100%
50	20	2	ΟΧΙ	20%	11,976	100%
30	40	3	ΟΧΙ	20%	14,139	100%
40	30	3	ΟΧΙ	20%	14,146	100%
40	30	2	ΟΧΙ	20%	14,298	100%
30	40	2	ΟΧΙ	20%	14,326	100%
40	20	3	ΝΑΙ	20%	15,505	100%
30	30	3	ΝΑΙ	20%	17,527	100%
50	20	3	ΝΑΙ	20%	19,357	100%
40	30	3	ΝΑΙ	20%	23,112	100%
40	30	2	ΝΑΙ	20%	23,228	100%
30	20	2	ΟΧΙ	20%	6,887	90%
30	20	3	ΟΧΙ	20%	6,987	90%
40	20	1	ΟΧΙ	20%	9,495	90%
40	20	3	ΟΧΙ	20%	9,534	90%
40	20	2	ΟΧΙ	20%	9,583	90%
50	20	1	ΟΧΙ	20%	11,868	90%
30	40	1	ΟΧΙ	20%	14,006	90%
40	30	1	ΟΧΙ	20%	14,158	90%
50	20	2	ΝΑΙ	20%	19,423	90%
50	20	1	ΝΑΙ	20%	19,492	90%
40	30	1	ΝΑΙ	20%	23,117	90%
30	40	3	ΝΑΙ	20%	24,311	90%
30	20	3	ΝΑΙ	20%	11,136	80%
30	20	2	ΝΑΙ	20%	11,224	80%
40	20	2	ΝΑΙ	20%	15,623	80%
30	30	2	ΝΑΙ	20%	17,409	80%
30	30	1	ΝΑΙ	20%	17,545	80%
30	40	1	ΝΑΙ	20%	23,013	80%
30	40	2	ΝΑΙ	20%	23,156	80%
30	20	1	ΝΑΙ	20%	11,160	70%
40	20	1	ΝΑΙ	20%	15,516	70%
30	20	1	ΟΧΙ	20%	6588	60%

Πίνακας 5.1 Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου και οι αντίστοιχοι μέσοι χρόνοι(*) εκτέλεσης των εντολών του, όπως και τα ποσοστά εύρεσης του καλύτερου χρωμοσώματος με τιμή p.crossover=1.

(*) (Intel Core 2Duo 2.26Hz)

Για την καλύτερη σύγκριση των συνδυασμών του γενετικού αλγορίθμου επιλέγουμε παρακάτω, ειδικά για τις παραμέτρους του πληθυσμού και των επαναλήψεων τέτοιες τιμές ώστε να εκτελείται ο ίδιος αριθμός υπολογισμών. Συγκεκριμένα επιλέγουμε το γινόμενο των δύο αυτών παραμέτρων να είναι σταθερό και ίσο με 1000 για δύο συνδυασμούς 50x20 και 20x50.

Για την παράμετρο της πιθανότητας μετάλλαξης επιλέξαμε τρείς τιμές 0.01, 0.10 και 0.40. Η πιθανότητα 0.01 για μετάλλαξη σε κάθε συντεταγμένη του χρωμοσώματος δηλώνει μικρή μετάλλαξη και αρκετά μικρό ποσοστό αλλαγής των χρωμοσωμάτων. Από την μία αυτό είναι καλό γιατί δεν αλλοιώνονται τα υπάρχοντα χρωμοσώματα, αλλά από την άλλη είναι πολύ μικρή η διερεύνηση του χώρου των χρωμοσωμάτων. Έτσι έχουν επιλεχθεί και οι δύο μεγαλύτερες τιμές των 0.10 και 0.40. Η τελευταία είναι πιθανότατα αρκετά μεγάλη τιμή για πιθανότητα μετάλλαξης αλλά ενδείκνυται για την διερεύνηση της απόδοσης του αλγορίθμου.

Για την παράμετρο του ποσοστού διατήρησης των παλαιών χρωμοσωμάτων επιλέχθηκαν δύο τιμές, 20% και 50%. Η πρώτη είναι σχετικά μικρή και αφήνει περιθώριο για αναζήτηση νέων χρωμοσωμάτων, ενώ η δεύτερη με ποσοστό 50% αξιοποιεί τα υπάρχοντα χρωμοσώματα του πληθυσμού.

Τέλος, για την παράμετρο του ελιτισμού, έχουμε προφανώς δύο περιπτώσεις, της ύπαρξης (όπου παράλληλα δεν γίνεται μετάλλαξη – και αυτό δηλώνεται στους πίνακες με ***), ή μη ύπαρξης ελιτισμού.

Αυτές οι περιπτώσεις συνδυάζονται σε 48 διαφορετικές περιπτώσεις, 16 ανά τύπο διασταύρωσης, για τις οποίες τρέξαμε τον γενετικό αλγόριθμο και τα ίδια προσομοιωμένα δεδομένα κάθε φορά. Τα αποτελέσματα με τους μέσους χρόνους και την απόδοση του αλγορίθμου (ποσοστό εύρεσης του καλύτερου γραμμικού μοντέλου) για 10 επαναλήψεις φαίνονται στους Πίνακες ***5.2***, ***5.3*** και ***5.4*** παρακάτω.

Πληθυσμός x Επαναλήψεις	Πιθανότητα Μετάλλαξης	Τύπος Δ/ρωσης	Ελιτισμός	Ποσοστό Παλαιών	Χρόνος (*) Εκτέλεσης	Ποσοστό Εύρεσης
50x20	0,01	1	ΟΧΙ	50%	11,887	100%
50x20	0,10	1	ΟΧΙ	50%	12,025	100%
50x20	0,01	1	ΟΧΙ	20%	12,238	100%
50x20	0.10	1	ΟΧΙ	20%	12,335	100%
50x20	***	1	ΝΑΙ	50%	19,021	100%
50x20	***	1	ΝΑΙ	20%	19,711	80%
50x20	0,40	1	ΟΧΙ	50%	12,565	30%
50x20	0.40	1	ΟΧΙ	20%	12,546	20%
20x50	0,01	1	ΟΧΙ	50%	11,250	100%
20x50	0,10	1	ΟΧΙ	50%	11,832	100%
20x50	0,01	1	ΟΧΙ	20%	12,049	100%
20x50	0,10	1	ΟΧΙ	20%	12,414	100%
20x50	***	1	ΝΑΙ	50%	18,750	80%
20x50	***	1	ΝΑΙ	20%	19,764	60%
20x50	0,40	1	ΟΧΙ	50%	12,448	20%
20x50	0,40	1	ΟΧΙ	20%	12,529	10%

Πίνακας 5.2 Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου και οι αντίστοιχοι μέσοι χρόνοι(*) εκτέλεσης των εντολών του, όπως και τα ποσοστά εύρεσης του καλύτερου χρωμοσώματος για τύπο διασταύρωσης ενός σημείου και p.crossover=1.

(*) (Intel Core 2Duo 2.26Hz)

Από τον Πίνακα ***5.2*** φαίνεται πως για τύπο διασταύρωσης ενός σημείου (κωδικοποίηση «1» στον πίνακα και στον αλγόριθμο), και αριθμό επαναλήψεων και πληθυσμό τέτοιο ώστε το γινόμενο αυτών να είναι σταθερό (ίσο με 1000), ο γενετικός αλγόριθμος βρίσκει σε ποσοστό 100% το καλύτερο γραμμικό μοντέλο στις περισσότερες των περιπτώσεων. Συγκεκριμένα, όταν η πιθανότητα μετάλλαξης αυξάνεται (περίπτωση 0,40 ή 40%) ο αλγόριθμος δεν αποδίδει καθόλου καλά με τα ποσοστά εύρεσης του καλύτερου γραμμικού μοντέλου να πέφτουν στο 10-30%.
Ακόμα, στην περίπτωση που εφαρμόζεται ελιτισμός, πάλι ο αλγόριθμος δεν αποδίδει καλά, με τα ποσοστά να κυμαίνονται μεταξύ 60-100%. Συγκεκριμένα, όταν ο πληθυσμός είναι μεγάλος και οι επαναλήψεις λιγότερες (50 και 20 αντιστοίχως) ο αλγόριθμος αποδίδει στο 80-100% ενώ όταν ο πληθυσμός είναι μικρότερος και οι επαναλήψεις περισσότερες η απόδοση κυμαίνεται μεταξύ 60-80%. Αυτό οφείλεται πιθανότατα στο ότι όταν έχουμε μεγαλύτερο πληθυσμό τα χρωμοσώματα είναι περισσότερα και έτσι είναι δυνατή η καλύτερη εξερεύνηση του χώρου των χρωμοσωμάτων. Κάτι ακόμα που πρέπει να σημειωθεί είναι και ο μεγαλύτερος μέσος χρόνος τερματισμού των εντολών του αλγορίθμου όταν εφαρμόζεται ελιτισμός. Η αύξηση του χρόνου είναι περίπου στο 50-60% του χρόνου που απαιτείται για την εκτέλεση των εντολών του αλγορίθμου χωρίς να εφαρμόζεται ελιτισμός.

Πληθυσμός x Επαναλήψεις	Πιθανότητα Μετάλλαξης	Τύπος Δ/ρωσης	Ελιτισμός	Ποσοστό Παλαιών	Χρόνος (*) Εκτέλεσης	Ποσοστό Εύρεσης
50x20	0,01	2	ΟΧΙ	50%	12,063	100%
50x20	0,10	2	ΟΧΙ	50%	12,081	100%
50x20	0.10	2	ΟΧΙ	20%	12,360	100%
50x20	***	2	ΝΑΙ	50%	19,148	100%
50x20	0,01	2	ΟΧΙ	20%	12,761	90%
50x20	***	2	ΝΑΙ	20%	19,874	90%
50x20	0.40	2	ΟΧΙ	20%	12,547	30%
50x20	0,40	2	ΟΧΙ	50%	12,841	20%
20x50	0,01	2	ΟΧΙ	50%	11,506	100%
20x50	0,10	2	ΟΧΙ	50%	11,974	100%
20x50	***	2	ΝΑΙ	50%	18,632	100%
20x50	0,01	2	ΟΧΙ	20%	12,717	90%
20x50	0,10	2	ΟΧΙ	20%	12,940	90%
20x50	***	2	ΝΑΙ	20%	19,851	80%
20x50	0,40	2	ΟΧΙ	20%	12,329	20%
20x50	0,40	2	ΟΧΙ	50%	12,442	20%

Πίνακας 5.3 Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου και οι αντίστοιχοι μέσοι χρόνοι(*) εκτέλεσης των εντολών του, όπως και τα ποσοστά εύρεσης του καλύτερου χρωμοσώματος με p.crossover=1.

(*) (Intel Core 2Duo 2.26Hz)

Από τον Πίνακα ***5.3*** και για τύπο διασταύρωσης δύο σημείων (κωδικοποίηση «2» στον πίνακα και στον αλγόριθμο), και αριθμό επαναλήψεων και πληθυσμό τέτοιο ώστε το γινόμενο αυτών να είναι σταθερό (ίσο με 1000), ο γενετικός αλγόριθμος βρίσκει σε ποσοστό 100% το καλύτερο γραμμικό μοντέλο στις περισσότερες των περιπτώσεων. Συγκεκριμένα, όταν η πιθανότητα μετάλλαξης αυξάνεται (περίπτωση 0,40 ή 40%) ο αλγόριθμος δεν αποδίδει καθόλου καλά με τα ποσοστά εύρεσης του καλύτερου γραμμικού μοντέλου να πέφτουν στο 20-30%.

Επίσης, όταν το ποσοστό των παλαιών χρωμοσωμάτων που διατηρούνται στην επόμενη γενιά είναι μικρό (20% αντί του 50%), ο αλγόριθμος αποδίδει λιγότερο καθώς πέφτει το ποσοστό εύρεσης του καλύτερου γραμμικού μοντέλου στο 90%. Η αύξηση του ποσοστού παλαιών χρωμοσωμάτων που διατηρούνται στην επόμενη γενιά (50%) δίνει στον αλγόριθμο ποσοστά εύρεσης του καλύτερου γραμμικού μοντέλου 100%.

Και εδώ, με τον τύπο διασταύρωσης δύο σημείων, στην περίπτωση που εφαρμόζεται ελιτισμός, πάλι ο αλγόριθμος δεν αποδίδει καλά, με τα ποσοστά να κυμαίνονται μεταξύ 80-100%, δηλαδή πέφτουν τα ποσοστά εύρεσης αλλά είναι καλύτερα σε σύγκριση με τα αντίστοιχα του τύπου διασταύρωσης ενός σημείου. Τέλος, και εδώ πρέπει να σημειωθεί πως είναι μεγαλύτερος ο μέσος χρόνος τερματισμού των εντολών του αλγορίθμου όταν εφαρμόζεται ελιτισμός.

Γενικότερα, η απόδοση του αλγορίθμου φαίνεται να είναι καλύτερη από την περίπτωση διασταύρωσης ενός σημείου.

Πληθυσμός x Επαναλήψεις	Πιθανότητα Μετάλλαξης	Τύπος Δ/ρωσης	Ελιτισμός	Ποσοστό Παλαιών	Χρόνος (*) Εκτέλεσης	Ποσοστό Εύρεσης
50x20	0,01	3	ΟΧΙ	50%	11,277	100%
50x20	0,10	3	ΟΧΙ	50%	11,727	100%
50 x20	0,01	3	ΟΧΙ	20%	11,883	100%
50x20	0.10	3	ΟΧΙ	20%	11,983	100%
50 x20	***	3	ΝΑΙ	50%	18,418	100%
50x20	***	3	ΝΑΙ	20%	19,301	100%
50x20	0.40	3	ΟΧΙ	20%	12,180	20%
50 x20	0,40	3	ΟΧΙ	50%	12,262	10%
20x50	0,01	3	ΟΧΙ	50%	11,156	100%
20x50	0,10	3	ΟΧΙ	50%	11,602	100%
20x50	0,01	3	ΟΧΙ	20%	11,772	100%
20x50	0,10	3	ΟΧΙ	20%	11,796	100%
20x50	***	3	ΝΑΙ	20%	15,864	100%
20x50	***	3	ΝΑΙ	50%	18,360	100%
20x50	0,40	3	ΟΧΙ	20%	12,048	30%
20x50	0,40	3	ΟΧΙ	50%	12,070	20%

Πίνακας 5.4 Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου, και οι αντίστοιχοι μέσοι χρόνοι(*) εκτέλεσης των εντολών του, όπως και τα ποσοστά εύρεσης του καλύτερου χρωμοσώματος.(*) (Intel Core 2Duo 2.26Hz)

Από τον Πίνακα ***5.4*** και για τύπο ομοιόμορφης διασταύρωσης (κωδικοποίηση «3» στον πίνακα και στον αλγόριθμο), χρησιμοποιώντας p.crossover=1 και αριθμό επαναλήψεων και πληθυσμό τέτοιο ώστε το γινόμενο αυτών να είναι σταθερό (ίσο με 1000), ο γενετικός αλγόριθμος βρίσκει σε ποσοστό 100% το καλύτερο γραμμικό μοντέλο στις περισσότερες των περιπτώσεων. Συγκεκριμένα, όταν η πιθανότητα μετάλλαξης αυξάνεται (περίπτωση 0,40 ή 40%) ο αλγόριθμος δεν αποδίδει καθόλου καλά με τα ποσοστά εύρεσης του καλύτερου γραμμικού μοντέλου να πέφτουν στο 10-30%.

Ακόμα, το ποσοστό των παλαιών χρωμοσωμάτων που διατηρούνται στην επόμενη γενιά και για τις δύο περιπτώσεις (20% και 50%), ο αλγόριθμος αποδίδει το ίδιο καλά με το ποσοστό εύρεσης του καλύτερου γραμμικού μοντέλου να είναι 100%.

Για τις δύο μικρότερες τιμές της πιθανότητας μετάλλαξης δεν παρατηρείται καμία πτώση των ποσοστών εύρεσης του καλύτερου χρωμοσώματος από τον αλγόριθμο, αλλά όταν η πιθανότητα μετάλλαξης γίνεται 0,40 (ή 40%) ο αλγόριθμος αποδίδει το ίδιο άσχημα όπως και στις αντίστοιχες περιπτώσεις των διασταυρώσεων ενός και δύο σημείων, με τα ποσοστά να πέφτουν στο 10-30%.

Τέλος, με τον τύπο ομοιόμορφης διασταύρωσης, στην περίπτωση που εφαρμόζεται ελιτισμός, ο αλγόριθμος εξακολουθεί και αποδίδει καλά, με τα ποσοστά να διατηρούνται στο 100%, αντιθέτως με τις περιπτώσεις των διασταυρώσεων ενός και

δύο σημείων. Επίσης, και εδώ πρέπει να σημειωθεί πως είναι μεγαλύτερος ο μέσος χρόνος τερματισμού των εντολών του αλγορίθμου όταν εφαρμόζεται ελιτισμός.
Τέλος, θα χρησιμοποιήσουμε και την παράμετρο *p.crossover* η οποία δηλώνει την πιθανότητα να πραγματοποιηθεί κάθε μία από τις διασταυρώσεις σε κάθε επανάληψη και εφαρμόζεται στις περιπτώσεις για διασταύρωση ενός ή δύο σημείων. Στους προηγούμενους συνδυασμούς έχει ληφθεί η πιθανότητα αυτή ίση με 1 (100% να πραγματοποιηθεί η διασταύρωση). Θα ελέγξουμε και άλλες δύο περιπτώσεις για την συγκεκριμένη παράμετρο με τιμές 0,5 και 0,8.

Στον Πίνακα 5.5 έχουμε τις περιπτώσεις όπου η πιθανότητα ***p.crossover* είναι 0.5**.

Πληθυσμός x Επαναλήψεις	Πιθανότητα Μετάλλαξης	Τύπος Δ/ρωσης	Ελιτισμός	Ποσοστό Παλαιών	Χρόνος (*) Εκτέλεσης	Ποσοστό Εύρεσης
50x20	0,01	1	ΟΧΙ	50%	10.971	100%
50x20	0,10	1	ΟΧΙ	20%	11.879	100%
50x20	0.10	1	ΟΧΙ	50%	12.184	100%
50x20	***	1	ΝΑΙ	50%	18.906	80%
50x20	0,01	1	ΟΧΙ	20%	12.351	60%
50x20	***	1	ΝΑΙ	20%	19.092	30%
50x20	0,40	1	ΟΧΙ	50%	12.067	10%
50x20	0,40	1	ΟΧΙ	20%	12.484	10%
20x50	0,01	1	ΟΧΙ	50%	10.59	100%
20x50	0.10	1	ΟΧΙ	50%	11.679	100%
20x50	0,10	1	ΟΧΙ	20%	11.747	100%
20x50	0,01	1	ΟΧΙ	20%	12.301	50%
20x50	***	1	ΝΑΙ	50%	18.583	50%
20x50	***	1	ΝΑΙ	20%	19.884	30%
20x50	0,40	1	ΟΧΙ	20%	12.445	20%
20x50	0,40	1	ΟΧΙ	50%	12.502	10%

Πίνακας 5.5 Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου και οι αντίστοιχοι μέσοι χρόνοι(*) εκτέλεσης των εντολών του, όπως και τα ποσοστά εύρεσης του καλύτερου χρωμοσώματος. (*) (Intel Core 2Duo 2.26Hz)

Παρατηρείται στον Πίνακα 5.5, και σε σύγκριση με τους αντίστοιχους συνδυασμούς των τιμών των παραμέτρων από τον Πίνακα 5.2, πως ο αλγόριθμος γίνεται λιγότερο αποτελεσματικός. Η πιθανότητα διασταύρωσης επομένως όταν είναι σχετικά μικρή (όπως εδώ 0,5 ή 50%), επηρεάζει τον αλγόριθμο αρκετά σε επίπεδο αναζήτησης με αποτέλεσμα, όπως διαφαίνεται από τους πίνακες, να μην βρίσκει σε κάθε επανάληψη/γενιά ικανό αριθμό νέων ατόμων ώστε να γίνεται "καλή" ανανέωση του πληθυσμού και επομένως βελτίωση της λύσης. Ακόμα και στον συνδυασμό 20x50 όπου οι επαναλήψεις είναι περισσότερες, ο αλγόριθμος φαίνεται να χωλαίνει καθώς από τις 10 διασταυρώσεις κάθε γενιάς μόνο οι 5 πραγματοποιούνται κατά μέσο όρο. Οπότε, μπορεί με τις περισσότερες επαναλήψεις να είχαμε καλύτερα αποτελέσματα, αλλά αυτό τελικά δεν συμβαίνει λόγω του μικρού πληθυσμού.

Στον Πίνακα 5.6 έχουμε τις περιπτώσεις όπου η πιθανότητα ***p.crossover* είναι 0.8**.

Πληθυσμός x Επαναλήψεις	Πιθανότητα Μετάλλαξης	Τύπος Δ/ρωσης	Ελιτισμός	Ποσοστό Παλαιών	Χρόνος (*) Εκτέλεσης	Ποσοστό Εύρεσης
50x20	0,01	2	ΟΧΙ	50%	11.305	100%
50x20	0.10	2	ΟΧΙ	50%	11.829	100%
50 x20	0,10	2	ΟΧΙ	20%	12.12	100%
50 x20	***	2	ΝΑΙ	50%	19.451	100%
50x20	0,01	2	ΟΧΙ	20%	11.712	90%
50x20	***	2	ΝΑΙ	20%	20.032	80%
50 x20	0,40	2	ΟΧΙ	20%	12.583	20%
50x20	0,40	2	ΟΧΙ	50%	12.495	10%
20x50	0,01	2	ΟΧΙ	50%	11.358	100%
20x50	0,10	2	ΟΧΙ	20%	11.701	100%
20x50	0.1	2	ΟΧΙ	50%	11.864	100%
20x50	0,01	2	ΟΧΙ	20%	12.286	90%
20x50	***	2	ΝΑΙ	50%	19.034	80%
20x50	***	2	ΝΑΙ	20%	20.281	30%
20x50	0,40	2	ΟΧΙ	50%	11.786	20%
20x50	0,40	2	ΟΧΙ	20%	12.419	10%

Πίνακας 5.6 Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου και οι αντίστοιχοι μέσοι χρόνοι(*) εκτέλεσης των εντολών του, όπως και τα ποσοστά εύρεσης του καλύτερου χρωμοσώματος. (*) (Intel Core 2Duo 2.26Hz)

Ομοίως, παρατηρείται στον Πίνακα 5.6, και σε σύγκριση με τους αντίστοιχους συνδυασμούς των τιμών των παραμέτρων από τον Πίνακα 5.3, πως ο αλγόριθμος γίνεται λιγότερο αποτελεσματικός, αλλά καλύτερος από την περίπτωση όπου η πιθανότητα διασταύρωσης είναι 0.5. Η πιθανότητα διασταύρωσης επομένως όσο αυξάνεται (όπως εδώ από 0.5 σε 0.8 και σε 1), επηρεάζει τον αλγόριθμο δίνοντας όλο και καλύτερα αποτελέσματα. Αυτό το αποτέλεσμα είναι εν μέρει αναμενόμενο καθώς ειδικά στην περίπτωση που εφαρμόζεται ελιτίστικη στρατηγική τα παιδιά συγκρίνονται με τους γονείς και διατηρούνται στον νέο πληθυσμό τα καλύτερα δύο άτομα της τετράδας. Όταν η πιθανότητα διασταύρωσης είναι μικρότερη του 1 τότε κάποια διασταύρωση μπορεί να μην πραγματοποιείται και επομένως δεν υπάρχει η περίπτωση να βρεθεί καλύτερη λύση (τουλάχιστον από την συγκεκριμένη τετράδα). Σε αντίθεση, στην περίπτωση όπου η τιμή της πιθανότητας διασταύρωσης είναι 1, η διασταύρωση πραγματοποιείται σίγουρα, και λόγω ελιτίστικης στρατηγικής η λύση μπορεί να βελτιωθεί (εάν προκύψει παιδί καλύτερο των γονέων).

Στον Πίνακα 5.7 έχουμε κάποιες περιπτώσεις μέσω των οποίων μπορούμε να κάνουμε καλύτερη σύγκριση ως προς την παράμετρο της πιθανότητας διασταύρωσης. Συγκεκριμένα έχουμε για αυτές τις περιπτώσεις πιθανότητα μετάλλαξης σταθερή και

ίση με 0,01, τύπο διασταύρωσης 1 ή 2 σημείων, ποσοστό παλαιών χρωμοσωμάτων να διατηρούνται στην νέα γενιά ίσο με 20%. Οι περιπτώσεις που λαμβάνονται είναι οι συνδυασμοί για

- τύπο διασταύρωσης 1 ή 2 σημείων, με
- 3 περιπτώσεις για πιθανότητα διασταύρωσης (0,5 ή 0,8 ή 1) και
- με πιθανότητα μετάλλαξης 0,01 ή ελιτίστικη στρατηγική.

Πληθυσμός x Επαναλήψεις	Πιθανότητα Μετάλλαξης	Τύπος Δ/ρωσης	Πιθανότητα Δ/ρωσης	Ποσοστό Παλαιών	Χρόνος (*) Εκτέλεσης	Ποσοστό Εύρεσης
50x20	0.01	2	1	20%	12,087	100%
50x20	0.01	1	1	20%	12,261	90%
50x20	0.01	2	0,8	20%	12,405	90%
50x20	**Ελιτισμός**	2	1	20%	19,801	90%
50x20	0.01	1	0,8	20%	11,489	80%
50x20	**Ελιτισμός**	1	1	20%	19,957	80%
50x20	**Ελιτισμός**	2	0,8	20%	21,416	80%
50x20	0.01	1	0,5	20%	11,840	70%
50x20	0.01	2	0,5	20%	12,145	70%
50x20	**Ελιτισμός**	1	0,8	20%	19,989	70%
50x20	**Ελιτισμός**	2	0,5	20%	21,404	40%
50x20	**Ελιτισμός**	1	0,5	20%	21,056	30%

Πίνακας 5.7 Συνδυασμοί των παραμέτρων του γενετικού αλγορίθμου και οι αντίστοιχοι μέσοι χρόνοι(*) εκτέλεσης των εντολών του, όπως και τα ποσοστά εύρεσης του καλύτερου χρωμοσώματος. (*) (Intel Core 2Duo 2.26Hz)

Από τον παραπάνω Πίνακα 5.7, μπορούμε να εξάγουμε την επιρροή που έχει η πιθανότητα διασταύρωσης στην αποτελεσματικότητα του αλγορίθμου. Φαίνεται ξεκάθαρα πως καθώς μειώνεται η πιθανότητα διασταύρωσης τα ποσοστά εύρεσης της βέλτιστης λύσης από τον ΓΑ πέφτουν. Υπάρχει βεβαίως μια εναλλαγή των τιμών της πιθανότητας διασταύρωσης στην 4η στήλη του πίνακα, αλλά πρέπει να συνεκτιμηθεί και η επιρροή των άλλων παραμέτρων (ελιτισμού και τύπου διασταύρωσης).
Η διαπίστωση αυτή είναι αναμενόμενη, καθώς η μείωση της τιμής της πιθανότητας διασταύρωσης, ουσιαστικά επηρεάζει το πλήθος των διασταυρώσεων επομένως χάνονται κάποιες διερευνήσεις στο χώρο των λύσεων (οπότε μειώνονται και οι πιθανώς καλύτερες λύσεις για τον αλγόριθμο).

ΣΥΜΠΕΡΑΣΜΑΤΑ

Από την ανάλυση που προηγήθηκε και τις παρατηρήσεις που έγιναν, φαίνεται πως ο γενετικός αλγόριθμος αποδίδει καλύτερα όταν :

- Χρησιμοποιείται τύπος ομοιόμορφης διασταύρωσης.
- Από άποψη χρόνου δεν είναι καλό να εφαρμόζεται ελιτισμός.
- Ο πληθυσμός είναι αρκετά μεγάλος (περίπου στο τριπλάσιο των παραμέτρων)

- Οι επαναλήψεις δεν πρέπει να είναι λίγες, αν και αυτό μπορεί να φανεί από τις καλύτερες τιμές BIC που δίνει ο αλγόριθμος για κάθε γενιά, καθώς τρέχει ο αλγόριθμος αλλά και στο τέλος αυτού.
- Η πιθανότητα μετάλλαξης να είναι μικρή (1-10%).
- Το ποσοστό των χρωμοσωμάτων που διατηρούνται στην επόμενη γενιά να είναι 20-50%.
- Η πιθανότητα διασταύρωσης είναι καλό να είναι αρκετά κοντά στο 1 για γρηγορότερη σύγκλιση (αν όχι ίση με ένα). Σε αντίθετη περίπτωση θα πρέπει να αυξηθεί το πλήθος πληθυσμός x επαναλήψεις.

Τα παραπάνω αποτελέσματα – συμπεράσματα δηλώνουν :

1) Την καλύτερη απόδοση του γενετικού αλγορίθμου με έναν τύπο διασταύρωσης λίγο πιο σύνθετο (ομοιόμορφη διασταύρωση).
2) Ότι δεν είναι απαραίτητο να εφαρμόζεται ελιτισμός καθώς οι υπόλοιπες παράμετροι μπορούν να κάνουν τον αλγόριθμο να αποδίδει εξίσου ικανοποιητικά.
3) Την αναγκαιότητα όχι μικρού πληθυσμού, με την λογική να μπορούν να γίνονται διασταυρώσεις από ικανό αριθμό χρωμοσωμάτων ώστε να μπορεί να δουλεύει ο αλγόριθμος.
4) Οι επαναλήψεις να μην είναι λίγες, και αυτό είναι λογικό καθώς μέσα από τις επαναλήψεις ο αλγόριθμος μπορεί και συγκλίνει βρίσκοντας το καλύτερο χρωμόσωμα.
5) Η πιθανότητα μετάλλαξης να είναι αρκετά μικρή (1-10%) καθώς μεγαλύτερες τιμές έχουν ως αποτέλεσμα να αλλοιώνονται τα υπάρχοντα χρωμοσώματα και έτσι να μην συγκλίνει ο αλγόριθμος στο βέλτιστο χρωμόσωμα, αλλά από την άλλη θα πρέπει να υπάρχει μετάλλαξη ώστε να μπορεί να γίνει αναζήτηση και σε νέα χρωμοσώματα που τυχών δεν μπορούν να προκύψουν από τις διασταυρώσεις.
6) Το ποσοστό των χρωμοσωμάτων που διατηρούνται στην επόμενη γενιά να μην είναι μικρό, με σκοπό την βελτίωση των χρωμοσωμάτων μέσω των διασταυρώσεων και όχι απλά να βρίσκεται ένα καλύτερο χρωμόσωμα από μια τυχαία παραγωγή αυτού.

Βάσει αυτών των συμπερασμάτων θα δοκιμαστεί ο γενετικός αλγόριθμος στην επόμενη ενότητα για πραγματικά δεδομένα.

ΚΕΦΑΛΑΙΟ 6
ΠΡΑΓΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ

Τα πραγματικά δεδομένα πάνω στα οποία εφαρμόσαμε τον γενετικό αλγόριθμο, αφορούν πραγματικές μετρήσεις και δεδομένα καθημερινής συγκέντρωσης (p.p.m.) του αερίου του όζοντος στην περιοχή Upland της Καλιφόρνια, και 56 (εν δυνάμει) επεξηγηματικές μεταβλητές, με βασικές να είναι οι μετρήσεις για τις 9 τυχαίες μεταβλητές όπως είναι η ημέρα του χρόνου (μία από τις 330 παρατηρήσεις οι οποίες και αντιστοιχούν σε 330 συγκεκριμένες διαφορετικές ημέρες), ο άνεμος, η πίεση (στο Vanderberg και στο αεροδρόμιο του Los Angeles), η υγρασία, η θερμοκρασία (σε βαθμούς F, αλλά και ως διαφορά από την θερμοκρασία από το αεροδρόμιο του Los Angeles), υψόμετρο και ορατότητα. Συγκεκριμένα τα δεδομένα προέρχονται από το παρακάτω paper των ***Breinman, L. and Friedman, J.H. (1985)***.
Όπως αναφέραμε, οι ανεξάρτητες τυχαίες μεταβλητές είναι 56 στο πλήθος και οι παρατηρήσεις είναι 330, οι οποίες αντιστοιχούν σε 330 ημέρες μετρήσεων. Τα δεδομένα που χρησιμοποιήσαμε στην παρούσα εργασία είναι ελαφρώς επεξεργασμένα από τα αρχικά δεδομένα των πραγματικών μετρήσεων και αυτό για την καλύτερη χρήση τους σε γραμμικά μοντέλα. Έχουν χρησιμοποιηθεί οι κανονικοποιημένες τιμές των λογαρίθμων των αρχικών μεταβλητών. Συγκεκριμένα υπάρχουν 9 τυχαίες μεταβλητές που μπορεί να επηρεάζουν την εξαρτημένη μεταβλητή, 9 τυχαίες μεταβλητές $2^{ου}$ βαθμού, 2 τυχαίες μεταβλητές $3^{ου}$ βαθμού (για τις δύο μεταβλητές θερμοκρασίας) που μπορεί να επηρεάζουν επίσης την εξαρτημένη μεταβλητή, όπως και άλλες 36 τυχαίες μεταβλητές όπου δηλώνουν την αλληλεπίδραση των 9 τυχαίων μεταβλητών ανά ζεύγη, που μπορούν να χρησιμοποιηθούν στο γραμμικό μοντέλο.
Πιο αναλυτικά, για την εξαρτημένη μεταβλητή, η οποία είναι οι καθημερινές μετρήσεις της συγκέντρωσης του όζοντος, και η οποία είχε λοξότητα. Για την διόρθωση αυτής ο βέλτιστος μετασχηματισμός, βάσει του βέλτιστου εκθέτη λογαριθμικού μετασχηματισμού *Box-Cox*, είναι ο μετασχηματισμός $(OZON)^{0,17}$. Στην πράξη όμως είναι πιο εύκολος ο μετασχηματισμός του κανονικοποιημένου λογαρίθμου, και πολύ κοντά στον βέλτιστο μετασχηματισμό που προαναφέρθηκε. Συγκεκριμένα η μεταβλητή *OZON* είναι ο λογάριθμος της αρχικής καταγραφείσας τιμής αφού έχει αφαιρεθεί η μέση τιμή αυτής και έχει διαιρεθεί με την τυπική απόκλισή της, και αυτή αποτελεί τελικώς την εξαρτημένη τυχαία μεταβλητή των δεδομένων μας για το πολλαπλό γραμμικό μοντέλο, και την μεταβλητή *Y* για τον γενετικό αλγόριθμο αντιστοίχως.

Έχουμε δηλαδή, από τις αρχικές μετρήσεις 9 βασικές τυχαίες μεταβλητές από τις οποίες παράγουμε άλλες 11 ως τετράγωνα αυτών ή και κύβους αυτών, αλλά και άλλες 36 μεταβλητές ως συνδυασμούς ανά δύο των 9 βασικών τυχαίων μεταβλητών και οι οποίες δηλώνουν την αλληλεπίδρασή τους ανά ζεύγη. Στο σύνολό τους αυτές οι 9+9+2+36=56 είναι οι εν δυνάμει επεξηγηματικές μεταβλητές του πολλαπλού γραμμικού μοντέλου. Δηλαδή, ψάχνουμε ένα γραμμικό μοντέλο που θα εξηγεί επαρκώς την τιμή της συγκέντρωσης του όζοντος, στην συγκεκριμένη περιοχή, και η οποία συγκέντρωση θα επηρεάζεται σε μικρότερο ή μεγαλύτερο βαθμό από κάποιες εκ των 56 τυχαίων μεταβλητών που προαναφέραμε.

Συγκεκριμένα, οι 9 βασικές επεξηγηματικές μεταβλητές (από την X_1 έως και την X_9) της γραμμικής παλινδρόμησης, όπως και η εξαρτημένη μεταβλητή (Y), από τις οποίες παράγονται οι υπόλοιπες 47 επεξηγηματικές μεταβλητές (από την X_{10} έως και την X_{56}), είναι αυτές που φαίνονται στον ***Πίνακα 6.1***:

ΒΑΣΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ (Συμβολισμός)	**ΠΕΡΙΓΡΑΦΗ**
Y	Συγκέντρωση (p.p.m.) του αερίου του όζοντος στην περιοχή Upland της Καλιφόρνια.
$\mathbf{X_1}$	Ημέρα του χρόνου.
$\mathbf{X_2}$	Ταχύτητα Ανέμου (μίλια/ώρα) - στο LAX.
$\mathbf{X_3}$	500 mb ατμοσφαιρικής πίεσης (σε μέτρα) - στο VAFB.
$\mathbf{X_4}$	Ποσοστό Υγρασίας - στο LAX.
$\mathbf{X_5}$	Θερμοκρασία (σε βαθμούς F) – στο Σάντμπουργκ.
$\mathbf{X_6}$	Αντίστροφη βάση ύψους (σε πόδια) – στο LAX.
$\mathbf{X_7}$	Πίεση (σε mm Hg) – στο LAX.
$\mathbf{X_8}$	Αντίστροφη βάση θερμοκρασίας (σε βαθμούς F) – στο LAX.
$\mathbf{X_9}$	Ορατότητα (σε μίλια) – στο LAX.

Πίνακας 6.1. Οι βασικές μεταβλητές οι οποίες καταγράφηκαν στα πραγματικά δεδομένα, και από τις οποίες παρήχθησαν οι 56 μεταβλητές για την γραμμική παλινδρόμηση του προβλήματος, μαζί με την εξαρτημένη μεταβλητή.

Εάν επιχειρούσαμε να ελέγξουμε όλα τα δυνατά γραμμικά μοντέλα για αυτά τα δεδομένα, θα είχαμε 2^{56} διαφορετικά γραμμικά μοντέλα τα οποία προκύπτουν από την ύπαρξη ή όχι κάθε μίας εκ των 56 μεταβλητών στην γραμμική παλινδρόμηση. Αυτά θα έπρεπε να τα συγκρίνουμε βάσει της τιμής της συνάρτησης *BIC*, ώστε να επιλέξουμε το βέλτιστο. Ο υπολογισμός όλων αυτών των γραμμικών μοντέλων είναι εξαιρετικά χρονοβόρος και πιθανότατα αδύνατος λόγω απαίτησης μεγάλης υπολογιστικής ισχύος και μνήμης. Αυτός είναι ο λόγος που είναι απαραίτητος ένας γενετικός αλγόριθμος (ή μια γενικότερη μέθοδος εύρεσης βέλτιστης λύσης, χωρίς την αναλυτική σάρωση σε όλο τον χώρο των λύσεων).

Για αυτά τα δεδομένα λοιπόν, και για τις 56 επεξηγηματικές μεταβλητές που διαθέτουμε, αντιστοιχούμε την ύπαρξη ή όχι κάθε μίας μεταβλητής στο γραμμικό μοντέλο, με ένα δυαδικό διάνυσμα 56 θέσεων (χρωμόσωμα). Κάθε συντεταγμένη του χρωμοσώματος δηλώνει με 1 την ύπαρξη της αντίστοιχης μεταβλητής στο γραμμικό μοντέλο και με 0 την απουσία αυτής. Δηλαδή, τα άτομα του πληθυσμού για τον ΓΑ είναι τα εν λόγω χρωμοσώματα.

Ο γενετικός αλγόριθμος βρίσκει το καλύτερο χρωμόσωμα που αντιστοιχεί στα δεδομένα αυτά, δηλαδή το πλήθος των επεξηγηματικών μεταβλητών αλλά και ποιες είναι οι επεξηγηματικές μεταβλητές που συμπεριλαμβάνονται στο βέλτιστο γραμμικό μοντέλο. Βέβαια, ως βέλτιστο μοντέλο χαρακτηρίζεται από τον αλγόριθμο ένα γραμμικό μοντέλο βάσει του κριτηρίου της μικρότερης τιμής *BIC* μεταξύ των γραμμικών μοντέλων.

Για τα δεδομένα, λοιπόν, που διαθέτουμε επιλέξαμε για τον αλγόριθμο τις παρακάτω τιμές των παραμέτρων:

ΠΛΗΘΥΣΜΟΣ	150
ΕΠΑΝΑΛΗΨΕΙΣ	100
ΠΟΣΟΣΤΟ ΠΑΛΑΙΩΝ	0,20 (20%)
ΔΙΑΣΤΑΥΡΩΣΗ ΟΜΟΙΟΜΟΡΦΗ	“p.type.crossover= 3”
ELITIST	FALSE
ΠΙΘΑΝΟΤΗΤΑ ΜΕΤΑΛΛΑΞΗΣ	1%
ΠΙΘΑΝΟΤΗΤΑ ΔΙΑΣΤΑΥΡΩΣΗΣ	100%

και αυτό λαμβάνοντας τον πληθυσμό περίπου τριπλάσιο από το πλήθος των μεταβλητών, 100 επαναλήψεις ως επαρκής αριθμός επαναλήψεων, 20% κράτηση παλαιών ατόμων/χρωμοσωμάτων κάθε γενιάς (και αυτό σημαίνει 80% παραγωγής νέων ατόμων με την προοπτική να βρεθεί κάποιο χρωμόσωμα καλύτερο στην επόμενη γενιά), τύπο ομοιόμορφης διασταύρωσης (uniform)-μάσκας όπου φάνηκε πως έδινε καλύτερα αποτελέσματα στα προσομοιωμένα δεδομένα, και χωρίς ελιτίστικη στρατηγική (elitist) για γρηγορότερα αποτελέσματα (με πιθανότητα μετάλλαξης 1% και πιθανότητα διασταύρωσης 100%).

Ο αλγόριθμος έχει συγκλίνει στο βέλτιστο χρωμόσωμα από την 34 επανάληψη/γενιά και αυτό είναι το χρωμόσωμα :

1100101111-0110110101-0000000000-0000000000-0010000000-000000
(παρατίθεται ανά 10 στοιχεία για ευκολία στην παρατήρηση)

με τις συντεταγμένες ίσες με 1 να αντιστοιχούν στις τυχαίες μεταβλητές :
X1, X2, X5, X7, X8, X9, X10, X12, X13, X15, X16, X18, X20, X43.
Η τιμή *BIC* για το χρωμόσωμα αυτό είναι ***419.8924*** .

Τα 6 καλύτερα χρωμοσώματα του πληθυσμού με τις τιμές *BIC* είναι τα παρακάτω:

1100101111-0110110101-0000000000-0000000000-0010000000-000000	419.8924
1100001111-0110110101-0000000000-0000000000-0010000000-000000	419.9602
1100001111-0110110101-0000000000-0000000000-0110000000-000000	420.1448
1100101111-0110110101-0000000000-0000000000-0110000000-000000	420.2546
1100101111-0110110101-0000010000-0000000000-0110000000-000000	420.8757
1100101111-0110110101-0000010000-0000000000-0010000000-000000	421.2664

Παρατηρώντας βλέπουμε πως κάθε ένα από τα παραπάνω μοντέλα διαφοροποιείται από το προηγούμενό του ως προς μία και μόνο μεταβλητή κάθε φορά. Συγκεκριμένα, το $2^{ο}$ μοντέλο βγάζει την $5^{η}$ μεταβλητή, το $3^{ο}$ εισάγει την $42^{η}$ μεταβλητή, το $4^{ο}$ επανεισάγει την $5^{η}$ μεταβλητή, το $5^{ο}$ εισάγει την $26^{η}$ μεταβλητή και το $6^{ο}$ βγάζει την $42^{η}$ μεταβλητή. Οι διαφορές στην τιμή BIC είναι πολύ μικρές για τα έξι αυτά γραμμικά μοντέλα. Ενώ οι τιμές του προσαρμοσμένου συντελεστή προσδιορισμού διαφοροποιούν την κατάταξη των 6 αυτών μοντέλων, κάτι που είναι σύνηθες.

Η R για το συγκεκριμένο βέλτιστο γραμμικό μοντέλο έδωσε αναλυτικά στοιχεία για την γραμμική παλινδρόμηση τα παρακάτω, όπου με Xi έχουμε ορίσει το παραπάνω βέλτιστο χρωμόσωμα που προέκυψε από τον αλγόριθμο:

```
lm(formula = Y ~ X[, Xi])
Residuals:
    Min      1Q  Median      3Q     Max
-1.29823 -0.23951  0.02538  0.26569  0.88550

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.89696    0.07594  11.812  < 2e-16      ***
X[, Xi]1    -0.18415    0.02469  -7.459 8.52e-13      ***
X[, Xi]2    -0.10147    0.02696  -3.764 0.000199      ***
X[, Xi]5     0.14913    0.06274   2.377 0.018049      *
X[, Xi]7    -0.17334    0.04578  -3.786 0.000183      ***
X[, Xi]8     0.42482    0.07272   5.842 1.29e-08      ***
X[, Xi]9    -0.19582    0.03334  -5.873 1.08e-08      ***
X[, Xi]10   -0.47687    0.04868  -9.797  < 2e-16      ***
X[, Xi]12   -0.04749    0.01464  -3.243 0.001309      **
X[, Xi]13   -0.10100    0.02829  -3.570 0.000413      ***
X[, Xi]15   -0.17747    0.03945  -4.499 9.61e-06      ***
X[, Xi]16   -0.18223    0.02432  -7.492 6.89e-13      ***
X[, Xi]18    0.07308    0.02070   3.530 0.000478      ***
X[, Xi]20   -0.05512    0.01776  -3.103 0.002088      **
X[, Xi]43   -0.15205    0.02414  -6.298 1.01e-09      ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4065 on 315 degrees of freedom
  (36 observations deleted due to missingness)
Multiple R-squared: 0.8399,     Adjusted R-squared: 0.8328
F-statistic:   118 on 14 and 315 DF,  p-value: < 2.2e-16
```

Όπου παρατηρούμε πως όλες οι τυχαίες μεταβλητές (14 στο πλήθος από τις 56) που έδωσε ο αλγόριθμος, είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας τουλάχιστον **1%.**
Συγκεκριμένα **1%** για την **X5 , 1‰** για τις **X12 και X20** και 1/1000000 για τις υπόλοιπες. Επίσης, ο προσαρμοσμένος συντελεστής προσδιορισμού (R-Adjusted) είναι στο 0.8328 το οποίο σημαίνει πως το 83.28% της εξαρτημένης μεταβλητής εξηγείται από το συγκεκριμένο γραμμικό μοντέλο, και αυτό είναι αρκετά μεγάλο ποσοστό για γραμμικά μοντέλα με πραγματικά δεδομένα. Όπως σημαντική είναι και η τιμή p-value για το F test που προκύπτει και είναι μόλις στο 2.2*10^-16 (2.2e-16), δηλαδή εξαιρετικά μικρό! Με αυτό να σημαίνει πως απορρίπτεται η υπόθεση οι τιμές των συντελεστών των τυχαίων μεταβλητών του γραμμικού μοντέλου να είναι όλες μηδέν.

Ο αλγόριθμος ως 2^{ο} καλύτερο χρωμόσωμα (και γραμμικό μοντέλο), έδωσε το:

1100001111-0110110101-0000000000-0000000000-0010000000-000000

έχοντας αφαιρέσει την 5^{η} μεταβλητή και με τιμή BIC ***419.9602***. Δηλαδή, δίνει το γραμμικό μοντέλο με τις παρακάτω επεξηγηματικές μεταβλητές:
X1, X2, X7, X8, X9, X10, X12, X13, X15, X16, X18, X20, X43.

Η R για το συγκεκριμένο βέλτιστο γραμμικό μοντέλο έδωσε αναλυτικά στοιχεία για την γραμμική παλινδρόμηση τα παρακάτω, όπου με Xi έχουμε ορίσει το παραπάνω χρωμόσωμα που προέκυψε από τον αλγόριθμο:

```
lm(formula = Y ~ X[, Xi2])

Residuals:
    Min      1Q  Median      3Q     Max
-1.32343 -0.23881  0.01823  0.28710  0.90043

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.92461    0.07559  12.232  < 2e-16 ***
X[, Xi]1    -0.17270    0.02439  -7.081 9.34e-12 ***
X[, Xi]2    -0.09627    0.02707  -3.557 0.000433 ***
X[, Xi]7    -0.14068    0.04399  -3.198 0.001524 **
X[, Xi]8     0.53782    0.05544   9.702  < 2e-16 ***
X[, Xi]9    -0.19031    0.03350  -5.680 3.05e-08 ***
X[, Xi]10   -0.50790    0.04724 -10.752  < 2e-16 ***
X[, Xi]12   -0.05203    0.01463  -3.557 0.000432 ***
X[, Xi]13   -0.09543    0.02840  -3.360 0.000875 ***
X[, Xi]15   -0.16842    0.03955  -4.258 2.72e-05 ***
X[, Xi]16   -0.19046    0.02425  -7.854 6.35e-14 ***
X[, Xi]18    0.06897    0.02078   3.318 0.001011 **
X[, Xi]20   -0.05108    0.01781  -2.868 0.004406 **
X[, Xi]43   -0.16917    0.02321  -7.288 2.53e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4095 on 316 degrees of freedom
Multiple R-squared: 0.837,      Adjusted R-squared: 0.8303
F-statistic: 124.8 on 13 and 316 DF,  p-value: < 2.2e-16
```

Όπου παρατηρούμε πως όλες οι τυχαίες μεταβλητές (13 στο πλήθος από τις 56) που έδωσε ο αλγόριθμος, είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας τουλάχιστον **1‰**. Ο προσαρμοσμένος συντελεστής προσδιορισμού (R-Adjusted) είναι στο 0.8303 το οποίο σημαίνει πως το 83.03% της εξαρτημένης μεταβλητής εξηγείται από το συγκεκριμένο γραμμικό μοντέλο, και είναι 0.25% μικρότερο από το αντίστοιχο του 1ου χρωμοσώματος.

Ο αλγόριθμος ως 3^{ο} καλύτερο χρωμόσωμα (και γραμμικό μοντέλο), έδωσε το:

1100001111-0110110101-0000000000-0000000000-0110000000-000000

έχοντας προσθέσει την **42^{η}** μεταβλητή και με τιμή BIC ***420.1448***. Δηλαδή, δίνει το γραμμικό μοντέλο με τις παρακάτω επεξηγηματικές μεταβλητές:
X1, X2, X7, X8, X9, X10, X12, X13, X15, X16, X18, X20, X42, X43.

Η R για το συγκεκριμένο βέλτιστο γραμμικό μοντέλο έδωσε αναλυτικά στοιχεία για την γραμμική παλινδρόμηση τα παρακάτω, όπου με Xi έχουμε ορίσει το παραπάνω χρωμόσωμα που προέκυψε από τον αλγόριθμο:

```
lm(formula = Y ~ X[, Xi3])

Residuals:
    Min      1Q  Median      3Q     Max
-1.31446 -0.23584  0.01521  0.27372  0.95772

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.89322    0.07628  11.710  < 2e-16 ***
X[, Xi]1    -0.17443    0.02423  -7.198 4.50e-12 ***
X[, Xi]2    -0.09337    0.02691  -3.470 0.000593 ***
X[, Xi]7    -0.13215    0.04384  -3.014 0.002785 **
X[, Xi]8     0.52892    0.05519   9.584  < 2e-16 ***
X[, Xi]9    -0.18351    0.03340  -5.494 8.11e-08 ***
X[, Xi]10   -0.49223    0.04739 -10.386  < 2e-16 ***
X[, Xi]12   -0.05385    0.01455  -3.702 0.000252 ***
X[, Xi]13   -0.11297    0.02920  -3.869 0.000133 ***
X[, Xi]15   -0.15122    0.03997  -3.784 0.000185 ***
X[, Xi]16   -0.18954    0.02409  -7.869 5.78e-14 ***
X[, Xi]18    0.06549    0.02069   3.165 0.001703 **
X[, Xi]20   -0.04638    0.01780  -2.605 0.009622 **
X[, Xi]42    0.07608    0.03272   2.325 0.020712 *
X[, Xi]43   -0.14806    0.02477  -5.976 6.17e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4067 on 315 degrees of freedom
Multiple R-squared: 0.8398,     Adjusted R-squared: 0.8326
F-statistic: 117.9 on 14 and 315 DF,  p-value: < 2.2e-16
```

Όπου παρατηρούμε πως όλες οι τυχαίες μεταβλητές (14 στο πλήθος από τις 56) που έδωσε ο αλγόριθμος, είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας τουλάχιστον **1%.** Ο προσαρμοσμένος συντελεστής προσδιορισμού (R-Adjusted) είναι στο 0.8326 το οποίο σημαίνει πως το 83.26% της εξαρτημένης μεταβλητής εξηγείται από το συγκεκριμένο γραμμικό μοντέλο, και είναι 0.02% μικρότερο από το αντίστοιχο του 1ου χρωμοσώματος.

Ο αλγόριθμος ως 4^{ο} καλύτερο χρωμόσωμα (και γραμμικό μοντέλο), έδωσε το:

1100101111-0110110101-0000000000-0000000000-0110000000-000000

έχοντας προσθέσει ξανά την **5^{η}** μεταβλητή και με τιμή BIC ***420.2546***. Δηλαδή, δίνει το γραμμικό μοντέλο με τις παρακάτω επεξηγηματικές μεταβλητές:
X1, X2, X5, X7, X8, X9, X10, X12, X13, X15, X16, X18, X20, X42, X43.

Η R για το συγκεκριμένο βέλτιστο γραμμικό μοντέλο έδωσε αναλυτικά στοιχεία για την γραμμική παλινδρόμηση τα παρακάτω, όπου με Xi έχουμε ορίσει το παραπάνω χρωμόσωμα που προέκυψε από τον αλγόριθμο:

```
lm(formula = Y ~ X[, Xi4])

Residuals:
   Min     1Q  Median     3Q    Max
-1.2901 -0.2289  0.0269  0.2688  0.9417

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.86697    0.07657  11.323  < 2e-16 ***
X[, Xi]1    -0.18558    0.02453  -7.565 4.33e-13 ***
X[, Xi]2    -0.09852    0.02681  -3.675 0.000280 ***
X[, Xi]5     0.14568    0.06234   2.337 0.020081 *
X[, Xi]7    -0.16426    0.04565  -3.598 0.000372 ***
X[, Xi]8     0.41876    0.07229   5.793 1.68e-08 ***
X[, Xi]9    -0.18906    0.03325  -5.686 2.98e-08 ***
X[, Xi]10   -0.46231    0.04877  -9.479  < 2e-16 ***
X[, Xi]12   -0.04938    0.01457  -3.389 0.000791 ***
X[, Xi]13   -0.11799    0.02907  -4.059 6.24e-05 ***
X[, Xi]15   -0.16048    0.03988  -4.024 7.19e-05 ***
X[, Xi]16   -0.18152    0.02416  -7.512 6.09e-13 ***
X[, Xi]18    0.06960    0.02062   3.375 0.000832 ***
X[, Xi]20   -0.05043    0.01776  -2.839 0.004817 **
X[, Xi]42    0.07424    0.03250   2.284 0.023045 *
X[, Xi]43   -0.13185    0.02556  -5.158 4.42e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4039 on 314 degrees of freedom
Multiple R-squared: 0.8425,     Adjusted R-squared: 0.835
F-statistic:   112 on 15 and 314 DF,  p-value: < 2.2e-16
```

Όπου παρατηρούμε πως όλες οι τυχαίες μεταβλητές (15 στο πλήθος από τις 56) που έδωσε ο αλγόριθμος, είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας τουλάχιστον **1%**. Ο προσαρμοσμένος συντελεστής προσδιορισμού (R-Adjusted) είναι στο 0.835 το οποίο σημαίνει πως το 83.5% της εξαρτημένης μεταβλητής εξηγείται από το συγκεκριμένο γραμμικό μοντέλο, και είναι 0.22% μεγαλύτερο από το αντίστοιχο του 1ου χρωμοσώματος, αλλά η τιμή BIC είναι μεγαλύτερη καθώς ποινικοποιεί το πλήθος των μεταβλητών.

Ο αλγόριθμος ως 5^{ο} και 6^{ο} καλύτερο χρωμόσωμα (και γραμμικό μοντέλο), έδωσε τα:

1100101111-0110110101-0000010000-0000000000-0110000000-000000 420.8757
1100101111-0110110101-0000010000-0000000000-0010000000-000000 421.2664

έχοντας προσθέσει την **26^{η}** μεταβλητή και αφαιρέσει αντίστοιχα την **42^{η}** και με τιμές BIC ***420.8757*** και ***421.2664***. Δηλαδή, δίνει τα γραμμικά μοντέλα με τις παρακάτω επεξηγηματικές μεταβλητές:

X1, X2, X5, X7, X8, X9, X10, X12, X13, X15, X16, X18, X20, X26, X42, X43.
X1, X2, X5, X7, X8, X9, X10, X12, X13, X15, X16, X18, X20, X43.

Για το 5^{ο} γραμμικό μοντέλο του ΓΑ ο προσαρμοσμένος συντελεστής προσδιορισμού (R-Adjusted) είναι στο 0.837 το οποίο σημαίνει πως το 83.7% της εξαρτημένης μεταβλητής εξηγείται από το συγκεκριμένο γραμμικό μοντέλο, και είναι ακόμα καλύτερο ποσοστό από αυτό του πρώτου μοντέλου κατά 0.42%, αλλά όπως έχουμε αναφέρει το κριτήριο για τον αλγόριθμο είναι η τιμή BIC είναι μεγαλύτερη και αυτή είναι μεγαλύτερη για το 6^{ο} χρωμόσωμα σε σύγκριση με το 1^{ο}.

Αντιστοίχως για το 6^{ο} γραμμικό μοντέλο του ΓΑ ο προσαρμοσμένος συντελεστής προσδιορισμού (R-Adjusted) είναι στο 0.8345 το οποίο σημαίνει πως το 83.45% της εξαρτημένης μεταβλητής εξηγείται από το συγκεκριμένο γραμμικό μοντέλο, αλλά το μοντέλο αυτό είναι χειρότερο από το πρώτο ως προς την τιμή BIC.

Να σημειωθεί πως όλες οι επεξηγηματικές μεταβλητές που εισάγονται ή εξάγονται στα 6 πρώτα γραμμικά μοντέλα σε σύγκριση με το πρώτο γραμμικό μοντέλο έχουν συντελεστές που είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας τουλάχιστον **5%** ενώ όλες οι άλλες σε ποσοστό 1%, 1‰ ή και 1/1000000**.**

Για την καλύτερη σύγκριση των καλύτερων χρωμοσωμάτων ακολουθεί ο παρακάτω πίνακας που περιλαμβάνει τις 56 επεξηγηματικές μας μεταβλητές και τα έξι καλύτερα χρωμοσώματα. Στο χρωμόσωμα που περιλαμβάνεται η αντίστοιχη επεξηγηματική μεταβλητή έχουμε βάλει *, ενώ τα κενά κελιά δηλώνουν πως το χρωμόσωμά μας δεν περιλαμβάνει τη συγκεκριμένη επεξηγηματική μεταβλητή.

ΕΠΕΞΗΓΗΜΑΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ	**1^{ο} χρωμόσωμα**	**2^{ο} χρωμόσωμα**	**3^{ο} χρωμόσωμα**	**4^{ο} χρωμόσωμα**	**5^{ο} χρωμόσωμα**	**6^{ο} χρωμόσωμα**
X_1	*	*	*	*	*	*
X_2	*	*	*	*	*	*
X_3						
X_4						
X_5	*			*	*	*
X_6						
X_7	*	*	*	*	*	*

X_8	*	*	*	*	*	*
X_9	*	*	*	*	*	*
X_{10}	*	*	*	*	*	*
X_{11}						
X_{12}	*	*	*	*	*	*
X_{13}	*	*	*	*	*	*
X_{14}						
X_{15}	*	*	*	*	*	*
X_{16}	*	*	*	*	*	*
X_{17}						
X_{18}	*	*	*	*	*	*
X_{19}						
X_{20}	*	*	*	*	*	*
X_{21}						
X_{22}						
X_{23}						
X_{24}						
X_{25}						
X_{26}					*	*
X_{27}						
X_{28}						
X_{29}						
X_{30}						
X_{31}						
X_{32}						
X_{33}						
X_{34}						
X_{35}						
X_{36}						
X_{37}						
X_{38}						
X_{39}						
X_{40}						
X_{41}						
X_{42}			*	*	*	
X_{43}	*	*	*	*	*	*
X_{44}						
X_{45}						
X_{46}						
X_{47}						
X_{48}						
X_{49}						
X_{50}						
X_{51}						
X_{52}						
X_{53}						

X_{54}						
X_{55}						
X_{56}						

Πίνακας 6.2. Σύγκριση των έξι καλύτερων μοντέλων

ΚΕΦΑΛΑΙΟ 7
ΠΑΡΑΡΤΗΜΑ

Παρατίθενται παρακάτω οι κώδικες σε **R** σχετικά με τον γενετικό αλγόριθμο για την εύρεση καλύτερου γραμμικού μοντέλου πολλών μεταβλητών. Συγκεκριμένα, οι κώδικες α) για την προσομοίωση των δοκιμαστικών δεδομένων, β) για τον υπολογισμό της τιμής *BIC* για κάθε χρωμόσωμα/γραμμικό μοντέλο, γ) για την μετάλλαξη του κάθε χρωμοσώματος όπως αυτή έχει αναλυθεί στην παρούσα εργασία και τελικώς δ) τον κώδικα για τον ίδιο τον γενετικό αλγόριθμο.

7.1 ΠΡΟΣΟΜΟΙΩΣΗ ΔΕΔΟΜΕΝΩΝ (15 ΜΕΤΑΒΛΗΤΩΝ)

```
Ncol=15
Nrow=50
#x<-rnorm(Nrow*Ncol,0,1)    # dianysma
#X<-matrix(x,ncol=Ncol)     # PINAKAS

X<-matrix(rep(0,Nrow*Ncol),ncol=Ncol)
for (i in 1:Ncol){     # GIA KATHE STHLH PARATHRHSEIS TOY KATHE X
        X[,i]<-rnorm(Nrow,0,1)
}
Y<-rnorm(Nrow,3+8*X[,1]+4*X[,7]-4*X[,12] ,1.5)
```

7.2 ΥΠΟΛΟΓΙΣΜΟΣ ΚΑΙ ΕΞΑΓΩΓΗ ΤΙΜΗΣ *BIC*

```
eval.fun<-function(chromos){
        Nrow=dim(X)[1]
        BIC<-AIC(lm(Y~X[,chromos==1 ]), k = log(Nrow) )
        #print(BIC)
        return(BIC)
}
```

7.3 ΜΕΤΑΛΛΑΞΗ ΧΡΩΜΟΣΩΜΑΤΟΣ

```
 Mutation<-function(chro,p.m){       # METALLAXH GIA TO XROMOSOMA
                        # ME PITHANOTHTA p.m, SE KATHE SHMEIO TOY
        for (i in 1:length(chro) ){
          if (runif(1)<=p.m){
                chro[i]<-1-chro[i]
          }
        }
 return(chro) # EPISTREFEI TO IDIO 'H TO METALLAGMENO XROMOSWMA
}
```

7.4 ΓΕΝΕΤΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ

```
genetikos.alg.lm<-function(X,Y,population=50,epanalipseis=50, evaluation=eval.fun,
p.crossover=1, p.type.crossover=1, pososto.palaiwn=0.20 , elitist=FALSE,
p.mutation=0.01, analogia.0.1=c(1,1) ) {
# p.type.crossover = 1 FOR single_point/ = 2 FOR 2_points / =3 FOR Uniform
if ( (dim(X)[1]==length(Y)) ) {
        plithos.metabl<-dim(X)[2]
                # DINEI TO PLITHOS TWN METABLHTWN
                # GIA TA APOTELESMATA
        best.evaluations<-rep(NA,epanalipseis)
                # KALYTERO APOTELESMA GIA KATHE EPAN./GENIA
        evaluations.values<-rep(NA,population)
                #TIMES (BIC) GIA MIA EPAN./GENIA
        new.evaluations.values<-rep(NA,2*population)
                # TIMES (BIC) GIA MIA EPAN./GENIA
        elit.4.eval<-rep(NA,4)
        Popul.chromos <- matrix(nrow = population , ncol = plithos.metabl)
                # ARXIKOS PLITHISMOS ME ANALOGIA 0-1
                for (i in 1:population ) {     # TYXAIOS ARXIKOS PLITHISMOS
                        Popul.chromos[i,]<- sample( c(rep(c(0,1),analogia.0.1)) ,
plithos.metabl , rep=TRUE )
                }
        # TIMES BIC GIA TON PLITHISMO
                for (i in 1:population ) {
                        evaluations.values[i]<-eval.fun(Popul.chromos[i,])
                }
Popul.chromos<-cbind(Popul.chromos,evaluations.values )
BEST.POPULATION<-unique(Popul.chromos[order(evaluations.values),])
print("ARXIKOS Popul.chromos")
print(Popul.chromos) # ME TA BIC TELEYTAIA STHLH
        # EPANALHPSEIS
        NEW.Popul.chromos<- matrix(nrow = population , ncol = plithos.metabl)
                # NEOS PLITHISMOS
        DIASTAYRWSH.1.POINT <- p.type.crossover==1
        DIASTAYRWSH.2.POINTS<- p.type.crossover==2
        DIASTAYRWSH.UNIFORM <- p.type.crossover==3
        for (epan in 1:epanalipseis) {
                arithmitis<-BEST.POPULATION[,plithos.metabl+1]-
min(BEST.POPULATION[,plithos.metabl+1])
                paronomastis<-max(BEST.POPULATION[,plithos.metabl+1])-
min(BEST.POPULATION[,plithos.metabl+1])
                fitness<-1-arithmitis/paronomastis
                temp<-sample(1:length(fitness),population,replace=T, prob=fitness)
```

```
                Popul.chromos<-BEST.POPULATION[temp,]
for (i in 1:(population/2) ) {
                        # ARXH "PARAGOGHS" THS NEAS GENIAS
                        # EPILOGES ANAPARAGWGHS
                goneas.1<-Popul.chromos[2*i-1,1:plithos.metabl]
                        # ORISMOS TWN 2 GONEWN
                goneas.2<-Popul.chromos[2*i,1:plithos.metabl]
                        # ORISMOS TWN 2 GONEWN
if (DIASTAYRWSH.1.POINT){
                        # 1-point
  if (runif(1)<=p.crossover) {
    choice.where<-sample( 1:(plithos.metabl-1), 1 , prob= rep(1,plithos.metabl-1))
                        # EPILOGH SHMEIOY DIASTAYRWSHS
    miso.a1<-goneas.1[1:choice.where]
    miso.a2<-goneas.1[(choice.where+1):plithos.metabl]
    miso.b1<-goneas.2[1:choice.where]
    miso.b2<-goneas.2[(choice.where+1):plithos.metabl]
    paidi.1<-c(miso.a1,miso.b2)
    paidi.2<-c(miso.b1,miso.a2)
  } else {
    paidi.1<-goneas.1
    paidi.2<-goneas.2
  }
} else if (DIASTAYRWSH.2.POINTS){
                        # 2-points
  if (runif(1)<=p.crossover) {
     choice.where<-sort(sample(2:(plithos.metabl-1),2, prob= rep(1,plithos.metabl-2)) )
                        # EPILOGH SHMEIOY DIASTAYRWSHS
     paidi.1<-goneas.1
     paidi.1[choice.where[1]:choice.where[2]]<-
                        goneas.2[choice.where[1]:choice.where[2]]
     paidi.2<-goneas.2
     paidi.2[choice.where[1]:choice.where[2]]<-
                        goneas.1[choice.where[1]:choice.where[2]]

  } else {
    paidi.1<-goneas.1
    paidi.2<-goneas.2
  }
} else if (DIASTAYRWSH.UNIFORM ){
                        # uniform - mask
mask<-sample(c(0,1),plithos.metabl , rep=TRUE)
paidi.1<-goneas.1*mask+goneas.2*(!mask)
paidi.2<-goneas.2*mask+goneas.1*(!mask)
} # TELOS i-OSTHS DIASTAYRWSHS
if (elitist) {
elit.4<-rbind(paidi.1,paidi.2,goneas.1,goneas.2)
elit.4.eval<-c(eval.fun(paidi.1),eval.fun(paidi.2),eval.fun(goneas.1),eval.fun(goneas.2) )
```

```
paidi.1<-elit.4[order(elit.4.eval)[1], ]    # EPILOGH KALYTEROY ZEYGARIOY
paidi.2<-elit.4[order(elit.4.eval)[2], ]    # APO TOYS 4 (2 GONEIS + 2 PAIDIA)
} else {
paidi.1<-Mutation(paidi.1,p.mutation)  # MUTATION GIA TOYS 2 APOGONOYS
paidi.2<-Mutation(paidi.2,p.mutation)  # MONO AN DEN ISXYEI ELITIST
}
NEW.Popul.chromos[2*i-1,]<-paidi.1
NEW.Popul.chromos[2*i,]<-paidi.2
}  # TELOS "PARAGOGHS" THS NEAS GENIAS
print(epan)



Popul.100<-rbind( BEST.POPULATION[,1:plithos.metabl] , NEW.Popul.chromos )
        # TIMES BIC GIA TON NEO SYNOLIKO PLITHISMO.100
new.evaluations.values<-rep(NA,dim(Popul.100)[1])
for (i in 1:( dim(Popul.100)[1] ) ) {
        new.evaluations.values[i]<-eval.fun(Popul.100[i,])
}
        # ANTIKASTASTASH NEAS GENIAS
Popul.chromos[,1:plithos.metabl]<-Popul.100[order(new.evaluations.values)[1:population],]
        # ANTIKATASTASH TOY 80% TOY PLITHISMOY ME TYXAIA XROM.
for (i in (population*pososto.palaiwn+1):population ) {
        # TYXAIOS PLITHISMOS 80%
Popul.chromos[i,1:plithos.metabl]<-
sample(c(rep(c(0,1),analogia.0.1)),plithos.metabl,rep=TRUE )
}
        # TIMES BIC GIA TON NEO 100% PLITHISMO
for (i in 1:population ) {
        evaluations.values[i]<-eval.fun(Popul.chromos[i,1:plithos.metabl])
}
Popul.chromos[,plithos.metabl+1]<-evaluations.values
        # TIMES BIC GIA THN NEA GENIA WS TELEYTAIA STHLH
BEST.POPULATION<-unique(Popul.chromos[order(evaluations.values),])
best.evaluations[epan]<-min(evaluations.values)
print("BEST UNIQUE POPULATION 1:5")
print(BEST.POPULATION[1:5,])   # TA 5 KALYTERA MONTELA
          } # TELOS EPANALHPSEWN
        print("best.evaluations")
        print(best.evaluations)
        print("TELIKOS BEST.POPULATION")
        print(BEST.POPULATION)
return(BEST.POPULATION)
        } else {print( "LATHOS DIASTASH DEDOMENWN" )}
}  # END
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Breinman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association, 80*, 580-598.
- S.N.Sivanandam, S.N.Deepa,(2008). Introduction to Genetic Algorithms. *Springer Berlin Heidelberg New York.*
- M.Mitsell,(1996) An Introduction to Genetic Algorithms, *MIT press*.
- Lance Chambers,(2001).The Practical Handbook of Genetic Algorithms, Second Edition. *CHAPMAN & HALL/CRC*,.
- Colin R. Reeves, Jonathan E.Rowe,(2002).GENETIC ALGORITHMS-PRINCIPLES AND PERSPECTIVES, A Guide to GA Theory. *KLUWER ACADEMIC PUBLISHERS.*
- Gottlieb, Jens, et al.,et al. Prufer Numbers (2001). A Poor Representation of Spanning Trees for Evolutionary Search. *Procceedings of the Genetic and Evolutionary Computation Conference*.
- J. D. Bagley (1967).The behavior of adaptive systems which employ genetic and correlation algorithms, [PhD Thesis]. *University of Michigan*..
- J. H. Holland (1975). Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. *University of Michigan Press*.
- R. Rosenberg (1967). Simulation of genetic populations with biochemical properties, [PhD Thesis]. *University of Michigan*.
- K. A. De Jong (1975). An analysis of the behavior of a class of genetic adaptive systems, [PhD Thesis]. *University of Michigan Ann Arbor, MI, USA.*
- Geof H. Givens, Jennifer A. Hoeting,(2013). Computational Statistics. *Department of Statistics, Colorado State University, John Wiley& Sons Publication.*
- Dimitris Fouskakis, (2000). Stochastic Optimization Methods for Cost-Effective Assessment in Health, [Ph.D. Thesis]. *University of Bath U.K.*
- Lawrence Davis (1991). Hanbook of Genetic Algorithms. *Van Nostrand Reinhold,* New York.

ΙΣΤΟΣΕΛΙΔΕΣ

https://stat.ethz.ch/R-manual/R-devel/library/base/html/system.time.html

http://www4.ncsu.edu/~shu3/Presentation/AIC.pdf

http://www.math.ntua.gr/~fouskakis/Conferences/GS/Presentation_Crete.pdf

http://www.math.ntua.gr/~fouskakis/EPIPSI/05.pdf