Project Υπολογιστική Νοημοσύνη Β Μέρος

Β1. α) Κωδικοποίηση για τα άτομα του πληθυσμού:

Για να κωδικοποιήσουμε τα άτομα του πληθυσμού θα χρησιμοποιήσουμε δυαδική κωδικοποίηση.

Πρώτα, θα πραγματοποιήσουμε Standardization στα δεδομένα, προκειμένου να εξαλείψουμε τα outliers και έπειτα θα πραγματοποιήσουμε Normalization στα δεδομένα, ώστε να τα φέρουμε στο εύρος [0,1]. Με αυτό τον τρόπο, έχουμε αφαιρέσει κατά ένα μεγάλο βαθμό τα outliers που μπορεί να υπήρχαν στα δεδομένα μας, τα οποία θα επηρέαζαν μετρήσεις, όπως ο μέσος όρος ανά κλάση (sitting, walking, etc), τον οποίο θα χρειαστούμε για να αξιολογήσουμε τα άτομα του πληθυσμού, αλλά μειώσαμε και τον αριθμό των bit που θα χρειαστούν για να κωδικοποιήσουμε κάθε τις τιμές των x,y,z των 4 αισθητήρων.

Υποθέτουμε επιθυμητή ακρίβεια λύσεων: 4 δεκαδικά ψηφία. Άρα, σε συνδυασμό με τα παραπάνω, έχουμε:

Διάστημα τιμών των x, y, z είναι το [0,1], με μήκος 1. Το οποίο θα πρέπει να διαχωριστεί σε τουλάχιστον 1\*104=10000 ίσα υπο-διαστήματα για να επιτευχθεί η επιθυμητή ακρίβεια των τεσσάρων δεκαδικών ψηφίων. Άρα θα χρειαστούμε 14 δυαδικά ψηφία για καθένα από τα x, y, z των τεσσάρων αισθητήρων.

β) Με 14 bit μπορούμε να αναπαραστήσουμε 214=16384 διαφορετικούς αριθμούς, ενώ εμείς θέλουμε 10000, αφού τόσα είναι και τα υπο-διαστήματα στα οποία θα χωρίσουμε το διάστημα [0, 1]. Εάν κατά τη διασταύρωση ή τη μετάλλαξη προκύψει για κάποια μεταβλητή (δηλαδή 14αδα bit) ενός χρωμοσώματος τιμή η οποία είναι στο εύρος μεταξύ [10001, 16384], μπορούμε να αντιστοιχίσουμε, τυχαία και με ίση πιθανότητα, την πλεονάζουσα τιμή σε κάποια από τις νόμιμες τιμές. Με αυτό τον τρόπο διατηρούμε την τυχαία φύση του γενετικού αλγορίθμου αλλά μπορεί να “χάσει” πληροφορία με το να επιλέξει τιμή λιγότερο κατάλληλη από αυτή που είχε πριν.

Καθώς έχουμε χωρίσει το διάστημα [0, 1] σε 214 ίσα υπο-διαστήματα θα μπορούσαμε να αντιστοιχίσουμε ένα πάνω και ένα κάτω όριο, που θα δημιουργούνταν από αυτά τα διαστήματα, σε κάθε μία από τις 10000 τιμές που χρειαζόμαστε.

Δεν μπορούμε να αποφύγουμε τις πλεονάζουσες τιμές με χρήση της δυαδικής κωδικοποίησης, καθώς δεν μπορούμε να χωρίσουμε με κάποιο τρόπο το διάστημα [0,1] σε ίσα διαστήματα τα οποία να μπορούμε να τα αντιστοιχίσουμε με ακρίβεια σε κάθε μία από τις 10.000 διαφορετικές τιμές που μπορούν να πάρουν τα x,y,z κάθε αισθητήρα.

γ) Για την δημιουργία αρχικού πληθυσμού διαλέγουμε τυχαία και ισοπίθανα τιμές από το διάστημα [0, 1], τις κωδικοποιούμε με τον τρόπο που αναφέραμε στο α) ερώτημα και ανά 12 τις ενώνουμε ώστε να δημιουργήσουμε ένα χρωμόσωμα.

δ) Καθώς έχουμε κωδικοποιήσει τις τιμές των αισθητήρων με δυαδική κωδικοποίηση, Οι μετρικές Ευκλείδειας απόστασης, απόστασης Manhattan και συσχέτισης Pearson δεν είναι κατάλληλες για την αξιολόγηση της απόστασης μεταξύ ενός χρωμοσώματος και του μέσου όρου των 5 κλάσεων. Καθώς έχουμε πολύ μεγάλα μητρώα, η Ευκλείδεια απόσταση θα αξιολογεί ως όμοια άτομα τα οποία μπορεί να μοιράζονται πολλές 0 τιμές, χωρίς όμως αυτό να τα κάνει απαραίτητα όμοια. Η απόσταση Manhattan, θα επιστρέφει τον αριθμό των διαφορετικών bit μεταξύ δύο μητρώων, κάτι που δεν είναι επιθυμητό, καθώς μπορεί να τύχει μεταξύ δύο διαφορετικών στάσεων του σώματος, ο αριθμός των bit που διαφέρουν να είναι πολύ κοντά μεταξύ τους, και να μην μπορούμε έτσι να ξεχωρίσουμε σε ποια κλάση είναι πιο κοντά. Η συσχέτιση Pearson μετρά την γραμμική συσχέτιση μεταξύ μεταβλητών συνεχόμενων τιμών, υποθέτοντας ότι αυτές ακολουθούν κανονική κατανομή και καθώς τα δυαδικά δεδομένα δεν ακολουθούν κάποιο γραμμικό μοτίβο, τα αποτελέσματα θα ήταν ανακριβή.

Με τη χρήση ομοιότητας συνημιτόνου μπορούμε να βρούμε την ομοιότητα μεταξύ δύο μητρώων στο χώρο, που είναι η μεταξύ γωνία συνημιτόνου των δύο μητρώων. Η τιμή της είναι μεταξύ [-1,1]. Για τιμή 1, τα δύο μητρώα είναι ευθέως ανάλογα, δηλαδή όμοια. Για τιμή 0, τα δύο μητρώα είναι κάθετα, δηλαδή δεν έχουν ομοιότητα και για τιμή -1 είναι ακριβώς αντίθετα. Στη συγκεκριμένη περίπτωση, καθώς τα μητρώα μας είναι δυαδικά, το εσωτερικό γινόμενο μεταξύ δύο μητρώων θα είναι θετικός ακέραιος αριθμός, (και αφού το μέτρο τους είναι πάντα θετικό), η ομοιότητα συνημιτόνου θα κινείται το [0,1], όπου 0 τα μητρώα δεν σχετίζονται και 1 ταυτίζονται. Με αυτό τον τρόπο μπορούμε να δούμε εύκολα, κατά πόσο ένα χρωμόσωμα είναι κοντά ή όχι στην επιθυμητή κλάση (sitting) και να ενεργήσουμε ανάλογα.

ε)

i) Καθώς το cos(v, ts) κινείται στο [0,1], η ελάχιστη τιμή της F(v) προκύπτει όταν ένα διάνυσμα ατόμου δεν έχει κάποια ομοιότητα με το διάνυσμα sitting, αλλά είναι απολύτως όμοιο με τα άλλα 4 διανύσματα των υπόλοιπων κλάσεων. Άρα, cos(v, ts) = 0 και ∑ cos (𝑣,𝑡 𝑖≠𝑠 𝑖 ) = 4, δηλαδή F(v) = 0 + c (1 – 1/4) / (1 + c) => F(v) = 0.

Η μέγιστη τιμή της F(v), προκύπτει όταν το διάνυσμα ατόμου ταυτίζεται με το διάνυσμα sitting και δεν έχει καμία ομοιότητα με τα άλλα 4 διανύσματα των υπόλοιπων κλάσεων. Άρα, cos(v, ts) = 1 και ∑ cos (𝑣,𝑡 𝑖≠𝑠 𝑖 ) = 0, δηλαδή F(v) = 1 + c (1 – 0) / (1 + c) => F(v) = (1+c)/(1+c) => F(v) = 1.

Η συνάρτηση καταλληλότητας δεν μπορεί να πάρει αρνητικές τιμές, αφού το cos(v, tι) κινείται στο [0, 1].

ii) Η συγκεκριμένη συνάρτηση καταλληλότητας μια καλή επιλογή, καθώς μπορεί να κρίνει κατά πόσο ένα άτομο είναι κατάλληλο ως λύση του προβλήματός μας (δηλαδή κοντά στην επιθυμητή κλάση sitting). Δίνει μεγαλύτερες τιμές αξιολόγησης σε άτομα πιο κοντά στη λύση του προβλήματός μας και μικρότερες σε αυτά που είναι πιο μακριά από αυτή. Τέλος, για διαφορετικούς πληθυσμούς ατόμων θα δίνει παρόμοιες τιμές σε παρόμοια άτομα του πληθυσμού.

iii) Μια τιμή της σταθεράς c θα μπορούσε να είναι c = 1/4. Με αντικατάσταση στην F(v), για c = 1/4, προκύπτει ότι: . Άρα, η ομοιότητα ενός ατόμου με το διάνυσμα κλάσης sitting έχει βαρύτητα 4/5 και η ομοιότητα με τις υπόλοιπες κλάσεις έχει βαρύτητα 1/5. Με αυτό τον τρόπο δεν αμελείται η ομοιότητα με την επιθυμητή κλάση, αλλά ούτε και η ομοιότητα με τις υπόλοιπες κλάσεις.

στ)

i) Τελεστές επιλογής: ρουλέτα με βάση το κόστος, με βάση την κατάταξη και τουρνουά.

Στην περίπτωσή μας η χρήση ρουλέτας με βάση τη κόστος δεν είναι η κατάλληλη επιλογή καθώς αυτή χρησιμοποιείται όταν η συνάρτηση αξιολόγησης είναι και συνάρτηση κόστους, δηλαδή μικρότερες τιμές υποδεικνύουν και καλύτερη αξιολόγηση. Η συνάρτηση κόστους που θα χρησιμοποιήσουμε έχει ως καλύτερη τιμή το 1 και ως χειρότερη το 0.

Με χρήση ρουλέτας με βάση την κατάταξη, τα χρωμοσώματα θα ταξινομηθούν από το καλύτερο προς το χειρότερο με αύξουσα αρίθμηση (ξεκινώντας από το 1) και έπειτα θα τους ανατεθεί μία τιμή πιθανότητας, ανάλογα με τη θέση τους στην κατάταξη (όσο πιο ψηλά στην κατάταξη, τόσο μεγαλύτερη πιθανότητα). Με αυτό τον τρόπο όμως, χρωμοσώματα που βρίσκονται πολύ χαμηλά στην κατάταξη, αλλά θα μπορούσαν σε περίπτωση διασταύρωσης να δώσουν καλούς απογόνους, έχουν πάρα πολύ μικρές πιθανότητες να επιλεγούν.

Η επιλογή με χρήση τουρνουά φαίνεται η κατάλληλη για το πρόβλημά μας. Επιλέγουμε από τον πληθυσμό μία ομάδα χρωμοσωμάτων, τα αξιολογούμε και επιλέγουμε το καλύτερο από αυτά. Η διαδικασία επαναλαμβάνεται μέχρι να επιλέξουμε τον αριθμό χρωμοσωμάτων που θέλουμε. Μπορούμε ακόμα, να ορίσουμε κατάλληλα το μέγεθος των ομάδων του τουρνουά ώστε να αποφύγουμε την αποκλειστική επιλογή χρωμοσωμάτων με τη μεγαλύτερη βαθμολογία.

ii) Διασταύρωση: μονού, πολλαπλού σημείου και ομοιόμορφη

Με τη διασταύρωση μονού σημείου οι απόγονοι θα κινούνται πάντα σε ευθείες πάνω στους άξονες x, y που διέρχονται από κάποιο γονέα, κάτι το οποίο σημαίνει ότι δεν εξερευνάται αρκετά ο χώρος δημιουργίας απογόνων. Αυτό γιατί κατά τη διασταύρωση τα υπάρχοντα δομικά στοιχεία σε κάθε χρωμόσωμα είτε παραμένουν αυτούσια είτε συνδυάζονται αλλά όχι με τρόπο ώστε να δώσουν κάτι πολύ καλύτερο.

Με τη διασταύρωση πολλαπλού σημείου οι απόγονοι δεν θα κινούνται απαραίτητα σε ευθείες πάνω στους άξονες x, y που διέρχονται από κάποιο γονέα. Η εξερεύνηση του χώρου δημιουργίας απογόνων αυξάνεται αλλά όχι σημαντικά. Ακόμα, ο αριθμός των σημείων κοπής που θα ορίσουμε από την αρχή μπορεί να μην είναι αρκετά καλός ώστε να πετύχουμε μια όσο το δυνατόν πιο γρήγορη και καλή σύγκλιση. Και εδώ, μπορεί να συνδυάζονται περισσότερα δομικά στοιχεία κάθε φορά, αλλά και πάλι όχι με τρόπο ώστε να μπορέσει να παραχθεί, με μεγαλύτερη πιθανότητα, κάτι πιο αποδοτικό.

Με χρήση ομοιόμορφης διασταύρωσης κάθε bit γονιδίου των χρωμοσωμάτων αποτελεί πιθανό σημείο τομής, με αποτέλεσμα να εξερευνάται πολύ μεγάλο μέρος του χώρου δημιουργίας απογόνων, ειδικά στην αρχή όπου τα χρωμοσώματα διαφέρουν κατά πολύ μεταξύ τους. Όσο προχωρά ο αλγόριθμος τα χρωμοσώματα θα γίνονται ομοιόμορφα και οι απόγονοι θα περιορίζονται στο χώρο. Ακόμα, τα δομικά στοιχεία των γονέων διατηρούνται πιο εύκολα αυτούσια στους απογόνους, δίνοντας έτσι πιο αποδοτικά δομικά στοιχεία. Τέλος, μπορούμε να αποφύγουμε πρόωρη σύγκλιση, καθώς μέσω της μεγάλης μίξης μεταξύ των bit των γονιδίων αυξάνεται η ποικιλομορφία μεταξύ των γονιδίων, αυξάνοντας έτσι την πιθανότητα αποφυγής κάποιου τοπικού ελαχίστου.

Άρα, για την εκπαίδευση του γενετικού μας αλγορίθμου θα χρησιμοποιήσουμε ομοιόμορφη διασταύρωση.

iii) Στόχος της μετάλλαξης είναι η δημιουργία ενός καλύτερου χρωμοσώματος με μια τυχαία αλλαγή σε κάποιο από τα γονίδιά του. Αυτό έχει νόημα για χρωμοσώματα τα οποία έχουν μέτρια ή κακή απόδοση. Χρωμόσωμα που έχει την καλύτερη απόδοση μέσα σε ένα πληθυσμό είναι προτιμότερο να το διασταυρώσεις με κάποιο άλλο(-α), ώστε να προκύψουν καλύτεροι ή ισάξιοι απόγονοι, παρά να το μεταλλάξεις και να δώσει πιθανόν κάτι χειρότερο από πριν. Άρα, για τη μετάλλαξη η χρήση ελιτισμού είναι μια καλή απόφαση.

Σημειώσεις:

In the case of binary matrices, where the values are either 0 or 1, the cosine similarity formula can be simplified. The numerator of the cosine similarity formula would be the dot product of the two binary vectors, and the denominator would be the product of their magnitudes (which is the square root of the sum of squared elements).

Since the dot product of two binary vectors will always be an integer between 0 and the number of elements, and the magnitude of a binary vector will also be an integer, the resulting cosine similarity will generally fall within the range of [0, 1]. However, it is possible for the cosine similarity to be exactly 0 or 1 if the vectors are perfectly dissimilar or identical, respectively.

Therefore, in the case of binary matrices, you can expect the outcome of the cosine similarity to be within the range of [0, 1] without any manual adjustment.