ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ Εργαστηριακή Άσκηση Μέρος Β΄ Εαρινό Εξάμηνο 2019-2020

Παναγιώτης Χριστόπουλος 1054409



Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής

Β1. Σχεδιασμός ΓΑ

Κωδικοποίηση

Η προτεινόμενη κωδικοποίηση για τα άτομα του πληθυσμού είναι η Value Encoding. Με τη βοήθειά της και τη χρήση της συνάρτησης fillna της βιβλιοθήκης pandas, δημιουργήθηκε ένας πίνακας με γραμμές τα id των χρηστών, στήλες τα id των ταινιών και κελιά τα ratings, όπως φαίνεται στο παρακάτω screenshot.

movie_id	1	2	3	4	5	1678	1679	1680	1681	1682
user_id										
1	5.0	3.0	4.0	3.0	3.0	NaN	NaN	NaN	NaN	NaN
2	4.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	4.0	3.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
939	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
940	NaN	NaN	NaN	2.0	NaN	NaN	NaN	NaN	NaN	NaN
941	5.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
942	NaN	NaN	NaN	NaN	NaN	 NaN	NaN	NaN	NaN	NaN
943	NaN	5.0	NaN	NaN	NaN	 NaN	NaN	NaN	NaN	NaN
943	NaN	5.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N

Πλεονάζουσες Τιμές

Χρησιμοποιώντας δυαδική κωδικοποίηση θα είχαμε τις ακόλουθες τιμές: Οι αριθμοί από το 1 έως το 5 χρειάζονται 3 δυαδικά ψηφία για την απεικόνισή τους. Ωστόσο, θα είχαμε στον αλγόριθμό μας τις πλεονάζουσες τιμές για 6(110) και 7(111).

Τιμή	Δυαδική Κωδικοποίηση
1	000
2	001
3	010
4	011
5	100
6	110
7	111

Ωστόσο, με την κωδικοποίηση που προτάθηκε παραπάνω (Value Encoding), δεν υπάρχουν αυτές οι πλεονάζουσες τιμές.

Αρχικός πληθυσμός

Τιμές για τις ταινίες που δεν έχει δει

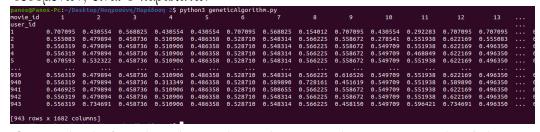
Για τις τιμές που δεν έχει δει, προστέθηκαν στον πίνακα μέσω της συνάρτησης fillna της βιβλιοθήκης pandas, οι τιμές του μέσου όρου των αντιστοίχων ταινιών. Με αυτό τον τρόπο, αγνοείται από το σύνολο των γονιδίων η τιμή μιας ταινίας για την οποία δεν υπάρχει αρχική βαθμολόγηση και δεν επηρεάζεται η μετρική Pearson η οποία χρησιμοποιείται για την εύρεση της γειτονιάς του χρήστη

Τιμές για τις ταινίες που έχει ήδη δει

Για τη μέγιστη απόδοση του αλγορίθμου χρησιμοποιούνται ως γονίδια όλες οι τιμές των ταινιών που έχει αξιολογήσει ο χρήστης ως εξής:

Αρχικά, κεντράρονται οι αξιολογήσεις με βάση τα σύνολα βαθμολογίας του χρήστη. Για λόγους αξιοπιστίας, θεωρείται αδύναμη η σύγκριση χρηστών οι οποίοι έχουν διαφορετικά σύνολα τιμών στις αξιολογήσεις τους. Στη συνέχεια, με τη βοήθεια του MinMax Scaler της βιβλιοθήκης SkLearn οι συνολικές τιμές κανονικοποιούνται στο διάστημα [0,1]

Ο συνολικός πίνακας που δημιουργείται, για την εύχρηστη προσπέλαση των δεδομένων, είναι ο παρακάτω:



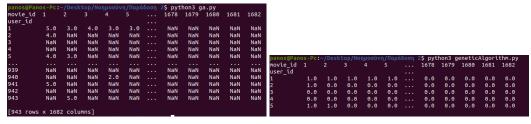
Ο αρχικός πληθυσμός ενός τυχαίου χρήστη μπορεί να απεικονιστεί σε διάνυσμα ως εξής:



Διαδικασία Επιδιόρθωσης

Για τη διαδικασία της επιδιόρθωσης χρησιμοποιήθηκε ένας oneHot Encoded πίνακας, ο οποίος είχε τον αριθμό 1 στα κελιά των ταινιών στα οποία ο χρήστης έχει αξιολογήσει και 0 σε αυτές που δεν υπάρχει αξιολόγηση μέχρι στιγμής. Για την ύπαρξη νομίμων λύσεων στην εκτέλεση του γενετικού αλγορίθμου,

ορίστηκε μία συνάρτηση καταλληλότας, η οποία λαμβάνοντας υπ όψιν τα στοιχεία του παραπάνω πίνακα, αποδέχεται τις μη νόμιμες λύσεις, αλλά εφαρμόζει ποινή σε αυτές της τάξης του 40% κατα τη διάρκεια της μετάλλαξης, αλλά και αυξημένη πιθανότητα επιλογής τους ως γενετικούς τελεστές.



(a) Αρχικός Πίνακας

(b) One-Hot Encoded

Εύρεση γειτονιάς χρήστη

Με βάση την κωδικοποίηση και την κανονικοποίηση που χρησιμοποιήθηκε στα προηγούμενα ερωτήματα, χρησιμοποιείται η μετρική Pearson για την εύρεση της γειτονιάς του χρήστη.

$$r_{xy} = rac{n\sum x_iy_i - \sum x_i\sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Τα στοιχεία του αρχικού πληθυσμού που προέκυψαν από την κανονικοποίηση και το κεντράρισμα είναι κανονικά κατανεμημένα. Επιπλέον, λόγω της συμπλήρωσης των NaN στοιχείων αλλά και της ανεξαρτησίας των παρατηρήσεων, ο συντελεστής συσχέτισης του Pearson θεωρείται η καλύτερη μετρική για τον αλγόριθμο αυτόν.

Συνάρτηση Καταλληλότητας

Για να ελέξγουμε την αποδοτικότητα του συστήματός μας, θα χρησιμοποιήσουμε μια Συνάρτηση Καταλληλότητας. Θα επιχειρήσουμε να προβλέψουμε την αξιολόγηση μιας συγκεκριμένης ταινίας από το χρήστη ως εξής: Στο μέσο όρο αξιολογήσεων ταινιών του χρήστη (πλήρες διάνυσμα αρχικού πληθυσμού), προστίθεται η συσχέτιση Pearson της αξιολόγησης που προέκυψε από την αξιολόγηση των χρηστών της γειτονιάς του. Στη συνέχεια, η νέα αυτή μετρική διαιρείται με το μέγιστο δυνατό αποτέλεσμα ώστε να απεικονίζεται σαν ποσοστό

επιτυχίας. Αφού προσθέσουμε όλα τα νέα νούμερα που προχύπτουν, τα διαιρούμε με τον αριθμό αρχιχού πληθυσμού για να βρούμε το μέσο όρο. Ο νέος αριθμός που προχύπτει είναι ο συντελεστής χαταλληλότητας.

Γενετικοί Τελεστές

Επιλογή τελεστών

Για την επιλογή τελεστών, μετασχηματίστηκε ο one-Hot πίνακας έτσι ώστε τα βάρη των υπαρχόντων αξιολογήσεων να είναι 1 και των συμπληρωμένων να είναι 0.6 (40% πιθανότητα επιλογής αξιολογημένης ταινίας για αρχικό τελεστή). Στη συνέχεια, βάση των υπαρχόντων βαρών, επιλέχθηκε η μέθοδος της ρουλέτας με βάση το κόστος με την εξής λειτουργία:

- Πρόσθεση των τιμών του συνόλου των βαρών ενός χρήστη στο νέο σύνολο S
- Εύρεση x τυχαίων τιμών στο σύνολο S(όπου x: ζητούμενοι αρχικοί τελεστές)
- Εύρεση των ταινιών που αντιστοιχούν σε κάθε τιμή του συνόλου χ

```
| Description |
```

Τα νούμερα ρουλέτας, τα id των επιλεγμένων ταινιών και οι επιλεγμένες ταινίες

Η παραπάνω μέθοδος θεωρείται η βέλτιστη για τον γενετικό μας αλγόριθμο. Η μέθοδος με βάση την κατάταξη θα ήταν χρήσιμη σε ένα dataset του οποίου τα βάρη έχουν μικρή απόκλιση.

Η μέθοδος του τουρνουά θα δημιουργούσε πρόβλημα στην απόδοση του αλγορίθμου εξαιτίας της διττής (1,0.6) τιμής των βαρών

Δ ιασταύρωση

• Διασταύρωση μονού σημείου

Η διασταύρωση μονού σημείου είναι η πιο διαδεδομένη επιλογή διασταύρωσης. Η απόδοσή του αλγορίθμου εξαρτάται άμεσα από την ποιότητα των γονέων. Στη διασταύρωση μονού σημείου, η αρχή και το τέλος ενός χρωμοσώματος χωρίζονται και σχηματίζουν απογόνους με την ουρά και το κεφάλι του συζυγούς χρωμοσώματος αντίστοιχα.

• Διασταύρωση πολλαπλού σημείου

Η διασταύρωση πολλαπλού σημείου προσθέτει πάνω από ένα crossover site. Επηρεαζει τις αλληλεπιδράσεις των δομικών στοιχείων. Επιτρέπει όμως διαφορετικά τμήματα ενός γονεά να συνυπάρχουν στον απόγονο.

• Ομοιόμορφη Διασταύρωση

Στην ομοιόμορφη διασταύρωση κάθε γονίδιο του παιδιού επιλέγεται τυχαία από ένα από τα αντίστοιχα γονίδια των χρωμοσωμάτων

• OX(Order Crossover)

Η ΟΧ χρησιμοποιείται για χρωμοσώματα με κωδικοποίηση μετάθεσης. Επιλέγονται αυθαίρετα τα crossover sites και συνδυάζονται. Στη συνέχεια αντιγράφονται τα στοιχεία μετάθεσης μεταξύ των σημείων διασταύρωσης από το cut string απευθείας στον απόγονο

• PMX(Partially Matched Crossover)

Στη διασταύρωση PMX, τα χρωμοσώματα συσχετίζονται και τα crossover sites επιλέγονται αυθαίρετα. Στη συνέχεια, η αρχική ένωση των χρωμοσωμάτων υφίσταται εναλλαγές θέση προς θέση με στόχο τη βέλτιστη απόδοση.

Στην περίπτωση του Γενετικού Αλγορίθμου που εξετάζουμε, η καλύτερη μέθοδος διασταύρωσης είναι η διασταύρωση μονού σημείου, αφού, μετά την κωδικοποίηση κατά τιμή, τα γονίδια είναι ήδη ισχυρά και μπορούν να σχηματίσουν ισχυρούς απόγονους. Η καταλληλότητα του αλγορίθμου, αν και ήδη υψηλή, φαίνεται ότι αυξάνεται με την πάροδο των γενεών

Στο παρακάτω παράδειγμα, ξεκινήσαμε τον αλγόριθμό μας για το χρήστη 10. Βλέπουμε ότι η αρχική καταλληλότητα των παιδιών που παράγονται από τη

διασταύρωση είναι αρχετά μεγάλη(71%). Μετά την πρώτη γενιά, και αφού αλλάζει ο πρώτος γονεάς(ελιτισμός), βλέπουμε μία αύξηση απόδοσης κατά 10%. Για τις δύο επόμενες επαναλήψεις ο αλγόριθμος βελτιώνεται με μικρό ρυθμό, στη συνέχεια σταθεροποιείται και επιτυγχάνει τη μέγιστη βελτιστοποίηση στις τελευταίες επαναλήψεις

```
Panos@Panos-Pc:~/Desktop/Nonpoovn/Napáδoon 2$ python3 geneticAlgorithm.py
User 10 's most similar users are: [800, 384, 849, 888, 369, 33, 728, 631, 583, 571]
Size of Initial Population: 20
User no: 10 selected movies are ['Time to Kill, A (1996)', 'Up Close and Personal (1996)',
nghai Triad (Yao a yao yao dao watpo qiao) (1995)', 'Ninotchka (1939)', 'Batman Returns (1
User no: 10 initial population chromosomes are [0.48638367561292467, 0.4798941890118596, 0
44957019267, 0.3481128555239228, 0.5632758418033369, 0.5419964099014917, 0.486383675612924
Number of Generations 10
Crossover point: 10
Generation: 0 Fittest Parent: 800 Generation Fitness 0.7079523965343794
Crossover point: 2
Generation: 1 Fittest Parent: 728 Generation Fitness 0.8165617784615875
Crossover point: 8
Generation: 2 Fittest Parent: 384 Generation Fitness 0.8961749419885596
Crossover point: 19
Generation: 3 Fittest Parent: 888 Generation Fitness 0.9059753007869658
Crossover point: 15
Generation: 4 Fittest Parent: 888 Generation Fitness 0.9059753007869658
Crossover point: 1
Generation: 5 Fittest Parent: 888 Generation Fitness 0.9059753007869658
Crossover point: 1
Generation: 6 Fittest Parent: 888 Generation Fitness 0.9059753007869658
Crossover point: 1
Generation: 7 Fittest Parent: 888 Generation Fitness 0.9059753007869658
Crossover point: 9
Generation: 8 Fittest Parent: 888 Generation Fitness 0.9059753007869658
Crossover point: 9
Generation: 8 Fittest Parent: 888 Generation Fitness 0.9059753007869658
Crossover point: 1
Generation: 8 Fittest Parent: 888 Generation Fitness 0.9059753007869658
Crossover point: 1
Generation: 9 Fittest Parent: 369 Generation Fitness 0.9589926366550128
Crossover point: 1
Generation: 9 Fittest Parent: 369 Generation Fitness 0.9589926366550128
```

Γενετικός Αλγόριθμος για τον χρήστη [10]

Μετάλλαξη

Για τη μετάλλαξη του πληθυσμού, χρησιμοποιείται η ταχτική του ελιτισμού. Σε κάθε γενιά στην οποία εφαρμόζεται ελιτισμός, ο αλγόριθμος επαναπροσδιορίζει τον πιο κοντινό γείτονα με βάση τα υπάρχοντα χρωμοσώματα. Έτσι διασφαλίζεται η αποφυγή τοπικών μεγίστων και η εύρεση της απόλυτα βέλτιστης λύσης. Για τη διαχείριση του αλγορίθμου μεταξύ γενεών, εφαρμόστηκε πιθανότητα ελιτισμού στην κύρια συνάρτηση, για την ομαλότερη απεικόνιση των αποτελεσμάτων.

Β2. Υλοποίηση ΓΑ

Github

Ο χώδιχας του αλγορίθμου που περιγράφεται στο μέρος B1 υπάρχει στο παραχάτω link: https://github.com/PanosChristopoulos/Genetic-MovieLens

Γλώσσα και Βιβλιοθήκες που χρησιμοποιήθηκαν

Το πρόγραμμα αναπτύχθηκε στη γλώσσα Python 3. Για την εκτέλεση του προγράμματος απαιτούνται οι παρακάτω βιβλιοθήκες:

- pandas
- numpy
- sklearn
- random
- matplotlib

Η πλειοψηφία των παραπάνω βιβλιοθηκών εγκαθίσταται με την εντολή pip3 install του τερματικού.

Τόσο το κύριο πρόγραμμα geneticAlgorithm.py όσο και οι εσωτερικές βιβλιοθήκες που χρησιμοποιήθηκαν είναι εμπλουτισμένα από σχόλια.

Εκτέλεση

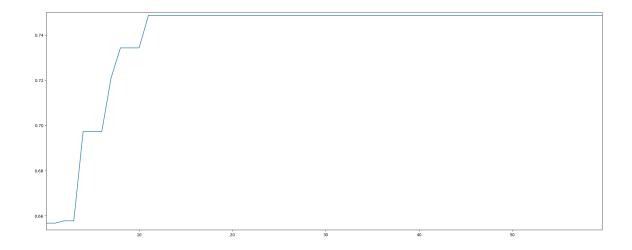
Μετά την εγκατάσταση των βιβλιοθηκών, πατήστε στο τερματικό: python3 geneticAlgorithm.py

Μετά την εκτέλεση, το πρόγραμμα ζητάει το ID του χρήστη/εισόδου στο πρόγραμμα και εμφανίζει τους πιο κοντινούς του γείτονες. Στη συνέχεια, ακολουθούν επιλογές μεγέθους αρχικού πληθυσμού και αριθμού γενεών. Το πρόγραμμα προβάλλει κάθε mutation και κάθε crossover που πραγματοποιείται. Στο τέλος του προγράμματος, εμφανίζονται οι προτάσεις ταινιών που προέκυψαν από τον αλγόριθμο και ένα διάγραμμα με άξονες τον αριθμό απόδοσης και τον αριθμό γενιάς.

```
panos@Panos-Pc:~/Desktop/Noημοσύνη/Παράδοση 2$ python3 geneticAlgorithm.py
Please type the desired input user's ID: 10
User 10 's most similar users are: [800, 384, 849, 888, 369, 33, 728, 631, 583, 571
Size of Initial Population: 30
User no: 10 selected movies as chromosomes names are ['Trigger Effect, The (1996)', he (1998)', 'Shallow Grave (1994)', 'Deceiver (1997)', 'My Family (1995)', 'Mars At
'Maximum Risk (1996)', 'Copycat (1995)', 'Bed of Roses (1996)', 'Belle de jour (19
Number of Generations 400
```

Εισαγωγή αρχικού χρήστη, μεγέθους πληθυσμού και αριθμού γενεών

Προτάσεις ταινιών



Διάγραμμα απόδοση/γενιά

Β3. Αξιολόγηση και Επίδραση Παραμέτρων

Πίνακας μετρήσεων

A/A	ΜΕΓΕΘΟΣ ΠΛΗ- ΘΥΣΜΟΥ	ΠΙΘΑΝΟΤΗΤΑ ΔΙΑΣΤΑΥΡΩΣΗΣ	ΠΙΘΑΝΟΤΗΤΑ ΜΕΤΑΛΛΑΞΗΣ	ΜΕΣΗ ΤΙΜΗ ΒΕΛΤΙΣΤΟΥ	ΜΕΣΟΣ ΑΡΙΘΜΟΣ ΓΕΝΕΩΝ
1	20	0.6	0.00	0.696	60
2	20	0.6	0.01	0.741	72
3	20	0.6	0.10	0.748	105
4	20	0.9	0.01	0.694	60
5	20	0.1	0.01	0.742	98
6	200	0.6	0.00	0.703	60
7	200	0.6	0.01	0.767	218
8	200	0.1	0.01	0.757	264
9	200	0.9	0.01	0.784	302

Καμπύλες

Οι καμπύλες υπάρχουν στο φάκελο plots στο github repository της εργασίας, που αναγράφεται παραπάνω, αριθμημένες σύμφωνα με το A/A

Συμπεράσματα

Η μετρική που χρησιμοποιήθηκε ως μέση τιμή βελτίστου περιγράφεται αναλυτικά στη συνάρτηση καταλληλότητας του ερωτήματος Α Από τον παραπάνω πίνακα συμπεραίνουμε τα εξής:

- Με την αύξηση του πληθυσμού βελτιώνεται η μέση τιμή του βελτίστου
- Με την αύξηση του πληθυσμού βελτιώνεται ο μέσος αριθμός γενεών
- Χωρίς πιθανότητες μετάλλαξης η μέση τιμή βελτίστου παραμένει σχεδόν σταθερή και οι γενιές είναι κατά πολύ μειωμένες