ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΑΝΑΠΤΥΞΗ ΛΟΓΙΣΜΙΚΟΥ ΓΙΑ ΑΛΓΟΡΙΘΜΙΚΑ ΠΡΟΒΛΗΜΑΤΑ
ΕΡΓΑΣΙΑ 2η

ΑΝΔΡΙΝΟΠΟΥΛΟΥ ΧΡΙΣΤΙΝΑ ΑΜ:1115201500006

Νοέμβριος 2018

Εισαγωγή

Η παρούσα εργασία υλοποιεί όλα τα ζητούμενα της άσκησης. Ο κώδικας έχει αναπτυχθεί σε γλώσσα προγραμματισμού C++ και έχει γίνει χρήση της STL.

Η μεταγλώττιση γίνεται με make και οι εντολές για run είναι:

- ./cluster -i αρχείο_input -c αρχείο.conf -o αρχείο_output -complete 0/1

  το αρχείο_input είναι το αρχείο που περιέχει τα δεδομένα του dataset,
  το αρχείο.conf είναι ένα configuration file με τις αρχικοποιήσεις των τιμών που αναφέρονται στην εκφώνηση της άσκησης,
  το αρχείο_output είναι το αρχείο στο οποίο γράφονται τα αποτελέσματα του προγράμματος
  και το complete σχετίζεται με τον τρόπο παρουσίασης των αποτελεσμάτων (0 = απλός τρόπος, 1 = αναλυτικός τρόπος)

-./cluster

  τα ζητούμενα αρχεία για την εκτέλεση του προγράμματος ζητούνται από την κονσόλα.

- Στα αρχεία υπάρχουν σχόλια, για την καλύτερη κατανόηση του κώδικα.

- Έγινε χρήση git.

- Υπάρχουν ενδεικτικά αρχεία με τα αποτελέσματα του προγράμματος στο παραδοταίο (RESULTS). Όλα τα αρχεία θα είναι διαθέσιμα στο git μου, αλλά και την ημέρα της εξέτασης (για γρηγορότερο upload της εργασίας).

- Το πρόγραμμα ανταποκρίνεται και σε μεγάλα datasets.

- Έγινε χρήση test units (google) για το αρχείο mymath.cpp που περιέχει όλες τις θεμελιώδεις – δομικές μαθηματικές συναρτήσεις του προγράμματος. Τα test βρίσκονται σε ξεχωριστό φάκελο στο παραδοταίο (TESTS).

Περιγραφή

Κατά την εκτέλεση του προγράμματος ο χρήστης καλείται να επιλέξει τον τρόπο που επιθυμεί να γίνει το initialization, το assignment και το update για το πρόγραμμα. Το πρόγραμμα χρησιμοποιεί euclidean ή cosine απόσταση όπως και στην εργασία 1.

Αφού αρχικοποιούνται k σημεία ως κέντρα για τα k clusters (το k προκύπτει από το configuration file), τα υπόλοιπα σημεία ανατίθενται στο πλέον κοντινότερο για αυτά κεντροειδές, σύμφωνα με τον επιλεγμένο αλγόριθμο. Έπειτα από κάθε assignment λαμβάνει χώρα το update, όπου επανακαθορίζει, με βάση τον αλγόριθμο που έχει επιλεχθεί από τον χρήστη, τα κέντρα για κάθε cluster. Έτσι τα νέα κέντρα είναι πιο αντιπροσωπευτικά.

Το πρόγραμμα τερματίζει όταν όλα τα σημεία έχουν ανατεθεί σε κάποιο cluster και κανένα δεν αλλάζει θέση, συνεπώς όταν η τοπολογία έχει πάρει την τελική της μορφή. Ακόμα δίνεται η δυνατότητα να τερματίζει μετά από κάποιον αριθμό επαναλήψεων του αλγορίθμου (σχολιασμένος κώδικας στη main).

Έπειτα, υπολογίζεται η silhouette για κάθε σημείο, ως μέτρο για το evaluate για να αναγνωρίσουμε αν τα παραχθέντα αποτελέσματα είναι σωστά.

Από την εργασία 1 έγινε χρήση του κώδικα που σχετιζόταν με το LSH και το HYPERCUBE για την υλοποιίηση των αντίστοιχων assignments. Δεν υπήρξαν θεμελιώδεις αλλαγές σε αυτόν.

Σύγκριση αλγορίθμων

Το πρόγραμμα φαίνεται να παρουσιάζει καλύτερα αποτελέσματα επιλέγοντας για initialization τη μέθοδο spread out, σε σχέση με τη random επιλογή κεντροειδών. Παρ’ όλα αυτά είναι πιο χρονοβόρα μεθόδος.

Σταθεροποιώντας την επιλογή για initialization και update παρατηρούμε ότι το assignment με LSH λειτουργεί αποδοτικότερα και ως προς την ποιότητα των αποτελεσμάτων και ως προς τον χρόνο εκτέλεσης.

Σε ότι αφοα το update το K-means update, που δημιουργεί ένα νέο κεντροειδές που δεν υπάρχει στο dataset, με βάση τα ήδη υπάρχοντα σημεία του cluster, λειτουργεί ταχύτερα συγκριτικά με το PAM, που ελέγχει κάθε στοιχείο του cluster με κάθε άλλο. Ωστόσο, το PAM φαίνεται να είναι πιο αποδοτικό, καθώς δεν επηρεάζεται από πιθανά σημεία – θόρυβο (σημεία που ανήκουν στο cluster και είναι σχετικά απομακρυσμένα συγκρινόμενα με την πλειοψηφία των σημείων του cluster) που μπορεί να περιέχει το cluster.