Ανάπτυξη Λογισμικού για Αλγοριθμικά Προβλήματα
Άσκηση 1

Δημήτριος Μπούσουλας 1115201500106
Άγγελος Λάλουσης 1115201500081

Κατάλογος αρχείων:

- main.c : Περιέχει όλες τις κλήσεις στις συναρτήσεις για το input, το clustering και το output.
- input.c : Περιέχει όλες τις συναρτήσεις για το διάβασμα των datasets.
- write_output.c : Περιέχει την υλοποίηση των συναρτήσεων που παράγουν τα output.
- initialization.c : Περίεχει τις συναρτήσεις για την αρχικοποίηση των clusters.
- assignment.c : Περιέχει τις συναρτήσεις για τα assignment των διανυσμάτων και των καμπυλών σε clusters.
- update.c : Περίεχει όλες τις συναρτήσεις για το update των κεντροϊδών.
- manhattan_distance.c : Περιέχει την συνάρτηση για τον υπολογισμό της manhattan distance 2 διανυσμάτων.
- dtw.c : Περιέχει τις συναρτήσεις για το dtw και το backtracking.
- euclidean.c : Υπολογισμός της ευκλείδιας απόστασης.
- lsh.c : Περιέχει τις συναρτήσεις για την αρχικοποίηση, την train και την αναζήτηση του lsh.
- lsh_functions.c : Περιέχει όλες τις περιφεριακές συναρτήσεις που σχεδιάσαμε στην 1η Εργασία για την υλοποίηση του lsh.
- silhouette.c : Περίεχει την υλοποίηση του silhouette.
- structs.h : Περίεχει όλες τις δομές που κατασκευάσαμε.
- functions.h : Περιέχει τους ορισμούς όλων των συναρτήσεων που κατασκευάσαμε.
- Makefile

Το εκτελέσιμο ονομάζεται cluster και η εντολή εκτέλεσης είναι:
./cluster –i <input file> –c <configuration file> -o <output file>

Παραδοχές:

- Στο configuration file τα ορίσματα (k, L κτλ) πρέπει να δίνονται με την σειρά που δίνονται και στο configuration της εκφώνησης και μετά τις άνω-κάτω τελείες πρέπει να υπάρχει κενό (πχ number_of_grids: 2)

- Όταν ενημερωθούν τα κέντρα κάθε διάνυσμα (ή καμπύλη) ψάχνει όλα τα κέντρα και όχι μόνο το κοντινότερο και το δεύτερο κοντινότερο. Αυτό επιλέξαμε να το κάνουμε καθώς οι update σταμάταγαν έπειτα από ελάχιστες επαναλήψεις και δεν δημιουγρούνταν αρκετά καλά clusters.

- Στον αλγόριθμο του DBA κάθε φορά που βρίσκουμε ότι ένα point μίας καμπύλης από το τρέχων cluster κάνει pair με ένα point από το C αυτου του cluster, αντί να το προσθέτουμε σε μία λίστα ώστε στο τέλος να υπολογίσουμε το mean-C. Προσθέτουμε κατευθείαν τις συντεταγμένες του στο καινούργιο mean-C ώστε να δημιουργηθεί το mean-C.

Περιγραφή:

Vectors:

random_selection :

Επιλέγονται τυχάια κ αρχικά κεντροϊδοί από το dataset.

k_means_plus_plus :

Επιλέγονται κ κέντρα με βάση των αλγόριθμο των διαφανειών, ώστε να έχουν μεγάλη πιθανότητα να είναι απομακρυσμένα μεταξύ τους.

Lloyds_assignment :

Για κάθε δίανυσμα του dataset βρίσκει το κοντινότερο και το 2ο κοντινότερο κέντρο κι αποθηκεύει τα cluster αυτων των κέντρων στα nearest και second_nearest.

LSH_assignment :

Πρώτα δημιουργούνται όλες οι δομές του LSH (Hashtables κτλ) με dataset τα διανύσματα του dataset και αποθηκεύονται στην main. Έπειτα για κάθε διάνυσμα βρίσκει το κοντινότερο κέντρο και αποθηκεύει το cluster αυτού του κέντρου στο nearest. Αυτό το πετυχαίνει κάνοντας lsh search για κάθε κέντρο και θεωρώντας εντός ακτίνας όσα διανύσματα βρίσκονται στο ίδιο bucket κι έχουν και το ίδιο g. Για όσα δεν υπήρξε κέντρο με που να χει το ίδιο g γίνεται direct assigment με απλή αναζήτηση μεταξύ όλων των κεντρων. Αν 2 ή περισσότερα κέντρα είχαν το ίδιο g τότε θα κρατήσει το κοντινότερο.

PAM :

Για κάθε cluster, παίρνει όλα τα διανύσματα του cluster και ελέγχει την συνολική απόσταση των υπόλοιπων διανυσμάτων του cluster απ’ αυτά. Αυτό που την ελαχιστοποιεί το ορίζει ως νέο κέντρο. Εφόσον το κάνει για όλα τα cluster καλεί μια συνάρτηση assignment ώστε να ενημερώσει τα clusters. Επαναλαμβάνει αυτήν την διαδικασία μέχρι να μην υπάρξει αλλαγη στα κέντρα ή μέχρι να κάνει 100 επαναλήψεις.

PAMean :

Για κάθε cluster, παίρνει όλα τα διανύσματα του cluster και υπολογίζει τον μέσο όρο τους και τον ορίζει ως νέο κέντρο. Εφόσον το κάνει για όλα τα cluster καλεί μια συνάρτηση

assignment ώστε να ενημερώσει τα clusters. Επαναλαμβάνει αυτήν την διαδικασία μέχρι να μην υπάρξει αλλαγη στα κέντρα ή μέχρι να κάνει 100 επαναλήψεις.

Curves:

random_selection_curve :

Επιλέγονται τυχάια κ αρχικά κεντροϊδοί από το dataset.

Lloyds_assignment_curve :

Για κάθε καμπύλη του dataset βρίσκει το κοντινότερο και το 2ο κοντινότερο κέντρο κι αποθηκεύει τα cluster αυτων των κέντρων στα nearest και second_nearest.

DBA

Πρόκειται για τον αλγόριθμο που υλοποιεί το update για τις καμπύλες. Συγκεκριμένα αρχικοποεί τα C κάθε συστάδας επιλέγοντας το λ και την υπακολουθία από λ points καλώντας τις Initialize_c και random_subsequence. Έπειτα λειτουργεί επαναληπτικά μέχρι να υπάρχει συνολική αλλαγή για όλα τα points των κεντροϊδών μικρότερη από 5% ή μέχρι να συμπληρωθούν 100 επαναλήψεις. Για όλες τις καμπύλες καθενός cluster καλεί την dtw με το τελευταίο όρισμα να είναι 1 ώστε να καλεστεί και η backtracking και να γεμίσει το traversal. Έτσι για κάθε C από κάθε cluster ακολουθείτε η λογική των διαφανειών για τα ταιριασμένα σημεία ώστε να ενημερωθεί το C (το οποίο αποτελεί το κεντροϊδές).

Σύγκριση:

Έπειτα από ακρετές δοκιμές βγάλαμε τα εξής συμπεράσματα.

Vectors:

Σε ότι αφορά τις initialization η random_selection ήταν σαφώς πιο γρήγορη αλλά έκανε μέτρια επιλογή για τα αρχικά κεντροϊδοί, ενώ η k-means++ ήταν πιο αργή αλλά έκανε καλύτερη επιλογή για τα αρχικά κεντροϊδοί. Τελικά είδαμε η κακή επιλογή για τα αρχικά κεντροϊδοί της random_selection δεν οδηγούσε σε κακές συστάδες οπότε επιλέξαμε αυτήν.

Σε ότι αφορά τις assignment η lsh είναι προσεγγιστική αλλά πιο γρήγορη από την Lloyds. Παρατηρήσαμε ότι τα ποσοστά επιτυχίας της ήταν τουλάχιστον 90% και σε ορισμένες περιπτώσεις (συγκεκριμένα σε μερίκα από τα dataset που διανύσματα είχαν 100 συντεταγμένες). Οπότε επιλέξαμε την lsh.

Σε ότι αφορά τις update η μόνη διαφορά που παρατηρήσαμε ήταν ως προς την ταχύητα και όχι ως προς το αποέλεσμα, οπότε επιλέξαμε την πιο γρήγορη δηλαδή την PAMean.

Curves:

Στα curves εφόσον δεν προλάβαμε να ολοκληρώσουμε την LSH_assignment ώστε να την στείλουμε δεν είχαμε να κανούμε δοκιμές μεταξύ αλγορίθμων.