Κ23γ: Ανάπτυξη Λογισμικού για Αλγοριθμικά Προβλήματα

Χειμερινό εξάμηνο 2014-15

2ηΠρογραμματιστική Εργασία

Υλοποίηση του αλγορίθμου ομαδοποίησης K-medoids στη γλώσσα C

Α)Στοιχεία φοιτητών που ανέπτυξαν το πρόγραμμα.

Επώνυμο:	**Μαντάς**	**Μεγκραμπιάν**
Όνομα:	**Ιωάννης**	**Αρσάκ**
Α.Μ:	**1115200700298**	**1115201100049**
email:	**sdi0700298@di.uoa.gr**	**sdi1100049@di.uoa.gr**

Β)Τίτλος και περιγραφή του προγραμματός

Ο τίτλος του προγράμματος είναι **SDproj2.**

Λειτουργικό Σύστημα Υλοποίησης:	Ubuntu 14.04
IDE Υλοποίησης:	Geany v.1.23.1
Χρόνος Υλοποίησης:	3 εβδομάδες

Μέθοδοι Αρχικοποιήσης (Initialization)

- **Concentrate:** Ο αλγόριθμος αυτός υλοποιήθηκε όπως αναφέρεται στο *"A simple and fast algorithm for K-medoids clustering" [Park-Jun '09]*, του οποίου το implementation αναφέρεται και στις διαφάνειες του μαθήματος. Το πρόγραμμα μας υλοποιεί πλήρως τον αλγόριθμο αυτό.
- **K-medoids++:** Ο αλγόριθμος αυτός υλοποιήθηκε όπως αναφέρεται στο *"K-means++: The Advantages of Careful Seeding" [Arthur - Vassilvitskii '07]*, του οποίου το implementation αναφέρεται και στις διαφάνειες του μαθήματος. Το πρόγραμμα μας υλοποιεί πλήρως τον αλγόριθμο αυτό.

Μέθοδοι Ανάθεσης (Assignment)

- **PAM assignment(simplest approach):** Ο αλγόριθμος αυτός για το assignment, όπως αναφέρει και το όνομα του, αποτελεί την απλούστερη μορφή για ανάθεση των αντικειμένων του συνόλου στα κέντρα των clusters. Δεν χρησιμοποιεί κανενός είδους βελτίωση απλώς ελέγχει όλα τα σημεία με τα κέντρα εξαντλητικά, όπως μας δίνεται και στις διαφάνειες.
- **Assignment by LSH/DBH(Reverse Approach):** (Edit: Δεν προλάβαμε τελικά να υλοποιήσουμε να κάνει compile και να είναι πλήρως συμβατός με το πρόγραμμα το LSH assignment.)Μπορούμε εφόσον μας δοθεί η δυνατότητα να παραδώσουμε εντός ημερών, καθώς βρισκόμασταν πολύ

κοντά στην ολοκλήρωση).

Μέθοδοι Ενημέρωσης (Update)

- **"A la Lloyd's"(improved PAM):** Ο αλγόριθμος αυτός για το update, όπως αναφέρεται στο *"A simple and fast algorithm for K-medoids clustering" [Park-Jun '09]*, του οποίου το implementation αναφέρεται και στις διαφάνειες του μαθήματος αλλά και συζητήθηκε στο φροντιστήριο Το πρόγραμμα μας υλοποιεί πλήρως τον αλγόριθμο αυτό.
- **CLARANS:** Ο αλγόριθμος αυτός υλοποιήθηκε όπως ακριβώς αναφέρεται στο *"CLARANS: A Method for Clustering Objects for Spatial Data Mining" [Ng - Han '94]*. Δηλαδή χρησιμοποιεί το cost differentials και κάνει τα ανάλογα swaps μέσω του αλγορίθμου PAM. Λόγω αδυναμίας για υπολογισμού του 2nd best centroid μέσω του LSH assignment, δεν είμαστε σε θέση να συνδυάσουμε αυτές τις δύο μεθόδους.

Μέθοδοι Αποτίμησης (Evaluation)

- **Silhouettes(simplest approach):** Ο αλγόριθμος αυτός για το evaluation του clustering, υλοποιήθηκε όπως αναφέρεται στο *"Silhouettes: a graphical aid to the interpretation and validation of cluster analysis" [Rousseeuw '87]*. Ο αλγόριθμος

- Σε κάθε μέθοδο επίσης μετράται ο χρόνος, όπως ζητείται, για να κάνουμε τους αντίστοιχους ελέγχους.
- Για το LSH πήραμε τα "χρήσιμα" κομμάτια κώδικα από την προηγούμενη άσκηση και τα χρησιμοποιήσαμε για το LSH assignment.

Γ)Κατάλογος των αρχείων κώδικα /κεφαλίδων και περιγραφή τους.

Τα αρχεία έχουν χωριστεί σε λογικές ομάδες για λόγους επεκτασιμότητας, απόκρυψης δεδομένων, καλύτερης διαχείρισης των αρχείων αλλά για λόγους καλύτερης αντίληψης των προγραμμάτων.

Πηγαία Αρχεία

./main.c	Τρέχει το πρόγραμμα και καλεί τις ανάλογες συναρτήσεις
./aEucMain.c	Η κύρια συνάρτηση clustering για Euclidean Spaces
./aHamMain.c	Η κύρια συνάρτηση clustering για Hamming Distance.
./aMatMain.c	Η κύρια συνάρτηση clustering για Distance Matrixes.
./cluster.c	Συναρτήσεις για τα cluster.
./generalfuns.c	Περιέχει διάφορες γενικές συναρτήσεις
./algorithms.c	Περιέχει τις υλοποιήσεις των ζητούμενων αλγόριθμων.

./Euc/eucfuns.c	Περιέχονται συναρτήσεις σχετικές με aEucMain.c
./Ham/hamfuns.c	Περιέχονται συναρτήσεις σχετικές με aHamMain.c
./Mat/matfuns.c	Περιέχονται συναρτήσεις σχετικές με aMatMain.c

■

Αρχεία Κεφαλίδων

./headers.h	Περιέχει όλες τα υπόλοιπα αρχεία κεφαλίδων
./mainfuns.h	Περιέχει τα ορίσματα των 3 κύριων συναρτήσεων.
./cluster.h	Τα ορίσματα του αντίστοιχου random.c
./defines.h	Περιέχει τα defines του προγράμματος.
./generalfuns.h	Τα ορίσματα του αντίστοιχου generalfuns.c
./algorithms.h	Τα ορίσματα του αντίστοιχου algorithms.c
./Euc/eucfuns.h	Τα ορίσματα του αντίστοιχου aEucMain.c
./Ham/hamfuns.h	Τα ορίσματα του αντίστοιχου aHamMain.c
./Mat/matfuns.h	Τα ορίσματα του αντίστοιχου aMatMain.c

Λοιπά

./makefile	Το αρχείο makefile.
./cluster.conf	Το configurator file.
./LSH	Directory που περιέχει όλα τα αρχεία για το LSH/DBH.

Δ)Οδηγίες μεταγλώττισης του προγράμματος.

Εντολές του makefile

make	Δημιουργεί όλα τα αντικειμενικά αρχεία και το εκτελέσιμο.
make clean	Διαγράφει όλα τα αντικειμενικά αρχεία και το εκτελέσιμο.
make count	Μετράει πλήθος γραμμών, λέξεων και χαρακτήρων.

Ε)Οδηγίες χρήσης του προγράμματος

Το πρόγραμμα εκτελείται γράφοντας

./SDproj1 –d <input file> –c <configuration file> -o <output file> -complete

Όλα τα ορίσματα είναι προαιρετικά. Σε περίπτωση που δεν δοθούν τα L και k έχουν default τιμές ενώ τα διάφορα files ζητούνται στην συνέχεια.

Ο χρήστης θα δώσει ένα dataset, σαν τα ενδεικτικά που μας δόθηκαν και το πρόγραμμα θα του επιστρέψει σε ένα αρχείο για όλους τους συνδυασμούς των παραπάνω αλγόριθμων που αναφέρθηκαν ό,τι ακριβώς ζητείται στην εκφώνηση της άσκησης.