Κ23γ: Ανάπτυξη Λογισμικού για Αλγοριθμικά Προβλήματα
Χειμερινό εξάμηνο 2016-17
3η Προγραμματιστική Εργασία
Υλοποίηση αλγορίθμων υπόδειξης (Recommendation) /
Συσταδοποίηση μοριακών διαμορφώσεων στη γλώσσα C/C++

ΣΤΟΙΧΕΙΑ ΦΟΙΤΗΤΩΝ

ΟΝΟΜΑ : ΙΩΑΝΝΗΣ
ΕΠΙΘΕΤΟ : ΤΣΙΩΡΗΣ
ΑΜ : 1115201300188

ΟΝΟΜΑ : ΓΕΩΡΓΙΟΣ
ΕΠΙΘΕΤΟ : ΚΟΚΚΑΛΗΣ
ΑΜ : 1115201300069

Proteins

Λειτουργία :

Το πρόγραμμα διαβάζει το dataset . Αφού το επεξεργαστεί , δημιουργεί τις 2 διεπαφές (c-RMSD , d-RMSD) . Για κάθε μία διεπαφή εκτελεί αλγόριθμο για clustering με διαφορετικό k , κάθε φορά . Στο τέλος μέσω της μετρικής silhouette , βρίσκει το k για το οποίο επιτυγχάνεται το βέλτιστο clustering . Η διεπαφή d-RMSD παράγει 7 τελικά clustering όσα και οι διαφορετικοί συνδυασμοί r και T . Για το c-RMSD , τα αποτελέσματα γράφονται στον αρχείο εξόδου <conform file> και για το d-RMSD στο <experim file> .

Recommendation

Λειτουργία :

Το πρόγραμμα διαβάζει το dataset . Αφού το επεξεργαστεί , δημιουργεί τις 2 διεπαφές (NN-LSH, Clustering) . Για κάθε μία διεπαφή και για κάθε μετρική ισχύουν τα ακόλουθα . Εκτελούμε τον αντίστοιχο αλγόριθμο για την εύρεση των πλησιέστερων γειτόνων . Συγκεκριμένα για το Clustering , βρίσκουμε το k για το οποίο έχουμε το βέλτιστο clustering μέσω της μετρικής silhouette . Στη συνέχεια για κάθε χρήστη βρίσκουμε τις εκτιμώμενες αξιολογήσεις των μη αξιολογημένων αντικειμένων , με χρήση του σταθμισμένου αθροίσματος . Τέλος βρίσκουμε για κάθε χρήστη , τα 5 καλύτερα αντικείμενα από αυτά που αρχικά δεν είχαν αξιολογηθεί . Τα αποτελέσματα γράφονται στο αρχείο εξόδου <output file>.

ΠΑΡΑΔΟΤΕΑ

φάκελος Proteins-Recommendations :

αρχείο README

φάκελος Datasets - περιέχει ενδεικτικά αρχεία εισόδου .

φάκελος Source - περιέχει όλα τα αρχεία πηγαίου κώδικα και επικεφαλίδας .

φάκελος Source :

φάκελος DataStructures - περιέχει αρχεία επικεφαλίδας (library) με τις υλοποιήσεις δομών δεδομένων με templates .

φάκελος General - περιέχει πηγαία αρχεία και αρχεία επικεφαλίδας για διάφορες βοηθητικές δομές .

φάκελος Metrics - περιέχει πηγαία αρχεία και αρχεία επικεφαλίδας για όλες τις μετρικές .

φάκελος Clusters - περιέχει πηγαία αρχεία και αρχεία επικεφαλίδας που αφορούν το clustering .

φάκελος Proteins :

αρχείο Main.cpp - Main .

αρχείο Makefile - Makefile .

αρχεία ProteinsManager.cpp , ProteinsManager.h - κλάση η οποία αποτελεί την διεπαφή του clustering με τον έξω κόσμο . υλοποιεί το update .

φάκελος Recommendation :

αρχείο Main.cpp - Main .

αρχείο Makefile - Makefile .

αρχεία FoldValidation.cpp, FoldValidation.h - Κλάση που υλοποιεί την τεχνική K-Fold Cross Validation .

αρχεία *Recommendation*Manager.cpp , *Recommendation*Manager.h - κλάση η οποία αποτελεί την διεπαφή του clustering με τον έξω κόσμο . υλοποιεί το update .

ΟΔΗΓΙΕΣ ΜΕΤΑΓΛΩΤΤΙΣΗΣ ΚΑΙ ΕΚΤΕΛΕΣΗΣ

Proteins

Εκτελώντας την εντολή make μέσα από τον φάκελο Proteins , το πρόγραμμα που αφορά την συσταδοποίηση μοριακών διαμορφώσεων , μεταγλωττίζεται επιτυχώς .

Για την εκτέλεση του προγράμματος ακολουθείται το εξής πρωτόκολλο .

$./proteins –d <input file> -od <experim file> -oc <conform file> -validate -cu

-d : αρχείο dataset

-od : αρχείο εξόδου d-RMSD

-oc : αρχείο εξόδου c-RMSD

-validate : αξιολόγηση μεθόδων

-cu : κάνε Unit testing

Τα argument -validate και -cu δεν είναι απαραίτητο να δοθούν στην γραμμή εντολής . Παρόλο αυτά αν δεν δοθούν κάποιο-α από τα αρχεία , τότε θα ζητούνται από τον χρήστη μέσα από το πρόγραμμα .

Recommendation

Εκτελώντας την εντολή make μέσα από τον φάκελο Recommendation , το πρόγραμμα που αφορά την υλοποίηση αλγορίθμων υπόδειξης , μεταγλωττίζεται επιτυχώς .

Για την εκτέλεση του προγράμματος ακολουθείται το εξής πρωτόκολλο .

$./recommendations –d <input file> -c <config file> -o <output file> -validate

-d : αρχείο dataset

-o : αρχείο results

-validate : αξιολόγηση μεθόδων

-cu : κάνε Unit testing

Τα argument -validate και -cu δεν είναι απαραίτητο να δοθούν στην γραμμή εντολής . Παρόλο αυτά αν δεν δοθούν κάποιο-α από τα αρχεία , τότε θα ζητούνται από τον χρήστη μέσα από το πρόγραμμα .

ΣΗΜΕΙΩΣΕΙΣ

Στο recommendation , για το clustering με την μετρική hamming , ισχύει η εξής παραδοχή . Στο βήμα του cutoff , όλες οι αξιολογήσεις παίρνουν τιμή 1 ενώ οι μη 0.

Στο recommendation , για το NN-LSH με την μετρική hamming , ισχύει η εξής παραδοχή . Στη δημιουργία των συναρτήσεων h , επιλέγονται περισσότερα ψηφία από 1 (συγκεκριμένα 60 , έπειτα από κάποια μελέτη/δοκιμές) .

ΕΚΤΙΜΩΜΕΝΟΙ ΧΡΟΝΟΙ ΕΚΤΕΛΕΣΗΣ

Για τα small datasets του e-class:

Recommendation: 200 seconds (χωρίς validate) , 260 seconds (με validate).

Proteins: 100 seconds.

ΠΑΡΑΤΗΡΗΣΕΙΣ ΠΑΡΑΜΕΤΡΩΝ ΠΡΩΤΕΪΝΩΝ

Παρατηρώντας πολλές εκτελέσεις του αλγορίθμου συσταδοποίησης, μπορούμε να ισχυριστούμε ότι για r = N^(3/2) και T = Random έχουμε τα πιο αξιόπιστα αποτελέσματα. Ο αριθμός των clusters και η σιλουέτα σταθεροποιούνται σε έναν συγκεκριμένο αριθμό, ενώ παράλληλα παράγονται ποιοτικά clusters. Παράλληλα, ο χρόνος εκτέλεσης είναι μικρός, και αντίστοιχος των υπολοίπων συνδυασμών r και T, επομένως μπορούμε να συμπεράνουμε ότι αυτός είναι ο καλύτερος συνδυασμός.

Στις παραπάνω μετρήσεις, ο χρόνος αφορά τον χρόνο εκτέλεσης της καλύτερης συσταδοποίησης και όχι την εξαντλητική αναζήτηση του καλύτερου αριθμού clusters.