

Report

Data mining – Assignment 2

Group members

Pranav Kumar Asthana : 2015A7PS0961H
Uttara Ravi : 2015A7PS0032H
Sahil Sangwan : 2015A7PS0965H
Utkarsh Grover : 2017A7PS1428H

Dataset used

Amino acid sequences:

<http://genome.crg.es/datasets/ggalhsapgenes2005/hg16.311.putative.aa.fa>

Preprocessing done

The data was read in the FASTA format (.fa). Since the data is small and easily fits on memory, we can load the entire data at once.

Once loaded, we computed the distances between each pair of data points. This is a time-intensive step since the number of comparisons that need to be made is quadratic in the number of data points. Moreover, the computation of each distance is of the order of the product of the lengths of the two individual data points.

Total time complexity = $O(N^2 \cdot m^2)$ where N is the number of data points, and m is the average length of each sample.

Since this step is time-intensive and common to all the further steps, it can be performed as part of preprocessing of the data.

Formulae used

1. Distance between clusters:

$$\text{MIN: } \text{dist}(\text{cluster 1}, \text{cluster 2}) = \min(|\text{dist}(p_1, p_2) \forall p_1 \in \text{cluster 1}, p_2 \in \text{cluster 2}|)$$

$$\text{MAX: } \text{dist}(\text{cluster 1}, \text{cluster 2}) = \max(|\text{dist}(p_1, p_2) \forall p_1 \in \text{cluster 1}, p_2 \in \text{cluster 2}|)$$

$$\text{AVG: } \text{dist}(\text{cluster 1}, \text{cluster 2}) = \frac{\sum(|\text{dist}(p_1, p_2) \forall p_1 \in \text{cluster 1}, p_2 \in \text{cluster 2}|)}{|\text{cluster 1}| \cdot |\text{cluster 2}|}$$

2. Average dissimilarity of a point in a cluster:

$$\text{dissimilarity}(\text{point}, \text{cluster}) = \frac{\sum(|\text{dist}(\text{point}, p) \forall p \in \text{cluster}|)}{|\text{cluster}|}$$

3. Diameter of a cluster

$$\text{diameter}(\text{cluster}) = \max(|\text{dist}(p_1, p_2) \forall p_1, p_2 \in \text{cluster}|)$$

Linkage and distance metric used

1. Linkage metric: We have used 3 linkage metrics for agglomerative clustering: Single (MIN), complete (MAX), average (AVG). For computing inter-cluster distance in divisive clustering, we have used the average metric.

2. Distance metric: For computing distance between two points(string), we have used a weighted edit distance with different weights for *match*, *substitution*, and *insert/delete*.

Type of data that can be clustered

We can cluster any kind of short text data where the sequence of letters is of importance, such as gene sequences, amino acid sequences, words, names, etc.

Dendrograms



