UNIVERSITI TEKNOLOGI MARA

THE DOCTORAL RESEARCH ABSTRACTS

Volume: 4, Issue 4 Nov 2013

FOURTH ISSUE

INSTITUTE of GRADUATE STUDIES

Leading You To Greater Heights, Degree by Degree

IPSis Biannual Publication

10

Name :
Ali Bin Seman

Title
Partitional Clustering Algorithms For Highly Similar And Sparseness Y-Short Tandem Repeat Data

Faculty :
Computer & Mathematical Sciences

Supervisor :
Prof. Dr. Zainab Abu Bakar (MS)

Prof. Dr. Mohd Nizam Isa (CS)

Clustering is an overlapping method found in many areas such as data mining, machine learning, pattern recognition, bioinformatics and information retrieval. The goal of clustering is to group any similar objects into a cluster, while the other objects that are not similar in the different clusters. Meanwhile, Y-Short Tandem Repeats (Y-STR) is the tandem repeats on Y-Chromosome. The Y-STR data is now being utilized for distinguishing lineages and their relationships applied in many applications such as genetic genealogy, forensic genetic and anthropological genetic applications. This research tends to partition the Y-STR data into groups of similar genetic distances. The genetic distance is measured by comparing the allele values and their

modal haplotypes. Nevertheless, the distances among the Y-STR data are typically found similar or very similar to each other. They are characterized by the higher degree of similarity of objects in intra-classes and also inter-classes. In some cases, they are quite distant and sparseness. This uniqueness of Y-STR data has become problematic in partitioning the data using the existing partitional clustering algorithms. The main problem was essentially caused by the mode mechanism (problem *P2*) which was unable to handle the characteristics of Y-STR data, thus producing poor clustering results. The problem has become worst when the initial centroid selection which is also known as problem *P0* failed to obtain good centroids. These conditions have led the existing partitional algorithms to local minima and empty clusters problems. As a result, a new idea of problem *P2* using the objects (medoid) themselves was introduced. The idea was incorporated into a new algorithm called, *k*-Approximate Modal Haplotypes (*k*-AMH) algorithm. Six Y-STR data sets were used as a benchmark to evaluate the performances of the algorithm against the other eight partitional clustering algorithms. Out of six data sets, the *k*-AMH algorithm obtained the highest mean accuracy scores for the five data sets and one data set was at equal performance. For the overall performances which were based on the six data sets, the *k*-AMH algorithm recorded the highest mean accuracy scores of 0.93 as compared to the other algorithms: the k-Population (0.91), the *k*-Modes-RVF (0.81), the New Fuzzy k-Modes (0.80), k-Modes (0.76), *k*-Modes-HI (0.76), k-Modes-HII (0.75), Fuzzy *k*-Modes (0.74) and k-Modes-UAVM (0.70). A One-Way ANOVA test also indicated that the clustering accuracy scores of *k*-AMH algorithm was significantly different as compared to the other eight partitional algorithms. In addition, the algorithm was also efficient in terms of time complexity which was recorded as *O (km(n-k)* and considered as linear. Thus, the *k*-AMH algorithm has been bounded with good characteristics of a desired algorithm – scalability.