DETERMINATION OF K IN K-MEANS CLUSTERING

Hany Harb[1], Ahmed Sobhy[2], Sherif Zaky[3] and Helmi Mahran[4]

[1] Department of Computers & Eng., Faculty of Eng., Al Azhar University, Cairo, Egypt.

[2] Department of Computer Science, Faculty of Computers & Informatics, Suez Canal University, Ismailia, Egypt.

[3] Department of Mathematics, Faculty of Science, Suez Canal University, Ismailia, Egypt.

[4] Department of Basic Science, Faculty of Computers & Informatics, Suez Canal University, Ismailia, Egypt.

Abstract

The *K*-means algorithm is a popular data-clustering algorithm. However, one of its drawbacks is the necessity for specifying the number of clusters, *k*, before the algorithm is applied. In this paper a simple approach based on Mahalanobis Distance will be presented. This approach depends on measuring the distance between each observation (row) in the data set and the common mean (μ_0) of this data set. Then, a number of clusters based on Mahalanobis distance are generated. The validity is calculated as a ratio between *within-cluster* and *between-clusters*. The highest three validity values are selected. Finally, validity indexes measures are calculated to determine the best *K*.

Keywords*: Missing Values, Imputation, Microarray, Regression.*

Published In: Journal of Al Azhar University Engineering Sector Vol. 6, No. 19, April 2011, 575- 591

A New Data Imputing Algorithm

[1]Ahmed Sobhy Sherif , [2]Hany Harb and Sherif Zaky[3]

[1] Department of Computer Science, Faculty of Computers & Informatics,
Suez Canal University, Ismailia, 41511, Egypt
Asse_ahmed_sobhy@yahoo.com

[2] Department of Computers & Eng., Faculty of Engineering,
Al Azhar University, Cairo, 11651, Egypt.
harbhany@yahoo.com

[3] Department of Mathematics, Faculty of Science,
Suez Canal University, Ismailia,41511, Egypt.
Sherif.ibrahim@gmail.com

Abstract

DNA microarray analysis has become the most widely used functional genomics approach in the bioinformatics field. Microarray gene expression data often contains missing values due to various reasons. Clustering gene expression data algorithms requires having complete information. This means that there shouldn't be any missing values. In this paper, a clustering method is proposed, called "Clustering Local Least Square Imputation method (ClustLLsimpute)", to estimate the missing values. In ClustLLsimpute, a complete dataset is obtained by removing each row with missing values. K clusters and their centroids are obtained by applying a non-parametric clustering technique on the complete dataset. Similar genes to the target gene (with missing values) are chosen as the smallest Euclidian distance to the centroids of each cluster. The target gene is represented as a linear combination of similar genes. Undertaken experiments proved that this algorithm is more accurate than the other algorithms, which have been introduced in the literature.

Keywords: *Missing Values, Imputation, Microarray, Regression*.

Published In: IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, May 2011,133-139.

My P.H.D thesis

Design a New Data Model Technique for Data Warehouse Systems

ABSTRACT

Data warehouse is a relational/multidimensional database that is designed for query and analysis rather than transaction processing. One of the most famous biological warehouse storages is DNA microarrays. Recent technology improvement makes us possible to quantify about 500,000 genes activity in a single microarray experiment, and this number is almost equivalent with the number of protein coding genes of human beings. DNA microarray experiments are extensively used to monitor the expression of a large amount of genes under various conditions. Associated with mathematical analysis methods, DNA microarray has important applications in biological and clinical studies. During the laboratory process, some spots on the array may be missing due to various factors (for example, machine error). Because it is often very costly or time consuming to repeat the experiment, molecular biologists, statisticians, and computer scientists have made attempts to recover the missing gene expressions by some ad-hoc and systematic methods. Clustering gene expression data algorithms, require complete information, that is, without any missing values. The *K*-means algorithm is a popular data clustering algorithm. Although k-means is a simple and can be used for a wide variety of data types, it is suffer from two major problems. First one, it is quite sensitive to initial positions of cluster centers (centroid), it is normally done randomly. The final cluster centroids may not be optimal ones as the algorithm can converge to local optimal solution. Second problem, it is the requirement for the number of clusters, *k*, to be specified before algorithm is applied. In this thesis, an efficient method to compute initial centroids for k-means clustering algorithm based on Mahalanobis Distance has been introduced as a solution for first problem. A new method to select the number of clusters ,*K*, for the K-means algorithm has been proposed in this thesis as a solution for second problem. Finally, a novel data imputing algorithm has been introduced based on k-means algorithm and least square regression method.