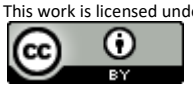
Available online at: https://ijact.in

Date of Submission	26/06/2019
Date of Acceptance	03/09/2019
Date of Publication	05/10/2019
Page numbers	3422-3430(9 Pages)

Cite This Paper: Ameer A. AL-Mshanji, Sura Z.Al-Rashid. Improving Clustering Algorithm for GENE Expression Data Using Hybrid Algorithm, 8(9), COMPUSOFT, An International Journal of Advanced Computer Technology. PP. 3422-3430.



ISSN:2320-0790

An International Journal of Advanced Computer Technology

IMPROVING CLUSTERING ALGORITHM FOR GENE EXPRESSION DATA USING HYBRID ALGORITHM

Ameer A. AL-Mshanji[1], Sura Z.AL-Rashid[2],
Software Department, College of Information Technology, University of Babylon, Hilla, Iraq
ameer.ali@uobabylon.edu.iq[1], sura_os@itnet.uobabylon.edu.iq[2]

Abstract: The technology of DNA Microarray has the ability to measure the levels of gene expression in different experimental conditions. Thousands of genes are generated in microarray experiments. The problem is that not all genes are significant; some of the genes may be noisy and irrelevant. The algorithms of Gene Selection are one of the important steps in the discovery of knowledge to select genes which are more informative. The other central goal of analyzing the data of gene expression is to identify genes that have similar patterns by using clustering processes. Clustering is a crucial process in the processes of data mining. It can divide genes into groups so that genes within the same group have similar features and share common biological functions. In this study, the method of mutual information for gene selection has been applied because it is able to detect nonlinear relationships between genes data. After that, the K-Means algorithm is applied to cluster data. The proposed approach results showed that it is capable of refining the data of gene expression for improved quality of clusters, handling noise effectively, and reducing the computational space.

Keywords: Microarray Technology; Gene Expression Data; Genes Selection; Clustering Algorithms; Clustering validation;

I. INTRODUCTION

With rapid technology development, Microarray Technology has become one of the most powerful tools in bioinformatics. Microarray technology is a good technique to observe the thousands of gene expression levels under different conditions at the same time. The conditions are usually consecutive time points during some environmental changes[1]. It can be beneficial to understand gene networks and functions in addition to its assistance in discovering the effects of medical treatments for diagnosing disease cases. The original gene data faces several problems such as missing values, noise and some variations. Therefore, the pre-processing of data is needed before any analysis [2]. The analysis of expression data can take two forms: it could either be a supervised analysis or an unsupervised one. In the supervised analysis, it is assumed that the structure data of the object is known. This knowledge is useful and can be applied in the analysis process. For the unsupervised analysis, the previously mentioned knowledge is not recognized.[3].Clustering (unsupervised) is a significant stage in the process of analyzing gene expression data and has been put into use in a wide range of fields such as medicine, biology, and engineering. Algorithms of clustering can able to discover the genes groups that exhibit similar expression patterns[4]. Clustering divides data points into sets or groups called clusters so that the homogeneity of elements in the same cluster shares some sort of strong similarity which is otherwise lower for the elements in other clusters. Clustering algorithms can cluster genes that have similar functions into clusters depending on the similarities of gene