Enliven Archive
Intensifying Thoughts

Review Article www.enlivenarchive.org Enliven: Journal of Genetic, Molecular and Cellular Biology

ISSN: 2379-5700

Assesssing the Effectiveness of Enhanced Clustering Algorithm used in Expression of Gene Data

Janet Titus

Department of Genetics, Stellenbosch University, Jc Smuts Building, De Beer Rd, Stellenbosch Central, Stellenbosch, 7600, South Africa

*Corresponding author: Janet Titus, Department of Genetics, Stellenbosch University, Jc Smuts Building, De Beer Rd, Stellenbosch Central, Stellenbosch, 7600, South Africa, E-mail: titusnet@outlook.com

Received Date: 05th October 2018
Accepted Date: 07th December 2018
Published Date: 8th December 2018

Citation: Titus J (2018) Assesssing the Effectiveness of Enhanced Clustering Algorithm used in Expression of Gene Data. Enliven: J Genet Mol Cell Biol 5(2): 007.

Microarrays allow researchers to perform simultaneous monitoring of the profiles of many gene expressions under diverse experimental environments. Furthermore, it is relied upon to establish the co-expressed genes in certain tissues or cells that are actively utilized in making proteins [1]. Besides, the bioinformatics research largely depends on this technique to examine the gene expression. Additionally, it is evident that clustering of gene expression data is an important process that helps to identify co-expressed genes or the clusters of samples and genes which can be regarded as biological relevant [2]. In the article, the researchers used the K-means clustering with AGMFI or the Automatic Generations of Merge Factor for ISODATA – (Iterative Self-Organizing Data Analysis Techniques) [3]. The results of the experiments on gene expression data prove that the suggested enhanced AGMFI can detect the compact clusters [4].

Clustering can be utilized in several applications ranging from data mining, medicine, biology, as well as engineering. Besides, scientists usually detect the co-expressed genes by clustering the gene expression data [5]. The DNA microarrays, given their high throughput, are popularly regarded as advanced technology with for evaluating the levels of gene expression [6]. Data matrix whereby the columns are used to signify the conditions while rows are relied upon to indicate the data and to comprehensively present the outcome from such investigations [7]. The matrix entries are considered as the measure of the level of the expression of a particular gene under a predetermined condition [8-9].

Furthermore, data analysis indicates the existence of genes with unrecognized functions in addition to developing the practical relations that exist between genes [10]. With regards to analysis of microarray gene expression, the SOM, K-Means clustering, and hierarchical clustering are considered as the most common clustering algorithms [11]. Among the three algorithms, the K-Means clustering as developed by Mac Queen has been proven to be the most popular, more effective as well as very easy to apply [12]. The simplicity of the K-Means clustering algorithm has made it applicable to numerous areas. In addition, it has proved to be successful with regards to producing clusters for several practical applications [13]. Nonetheless, with regards to large data sets, the original K-Means algorithm usually has a very high computational complexity [14]. As a partitioning clustering technique, the K-Means clustering algorithm is usually used to divide the data into K clusters [15]. One disadvantage of the K-Means algorithm includes the prior fixation of several clusters [16]. Cluster analysis seeks to group objects that are the same in a single band and isolate objects that are different by putting them in alternative groups [17].

The user selects a predetermined number of clusters (presuming K clusters) which are used to classify the objects. The idea largely entails choosing unsystematic cluster centers, one for each group and positioned far apart [18]. Furthermore, the K-Means algorithm usually uses the Euclidean distance to determine the distance that separates the centroids and data points [19]. ISODATA uses an iterative strategy to establish the superior cluster centers until it meets certain convergence criteria [20]. Besides, as compared to the K-means algorithm using the ISODATA makes it possible to split or merge the first number of the clusters, and therefore, the last number of clusters might be dissimilar to the number of clusters which have been indicated as a component of the output [21]. Additionally, the majority of the inputs include several extra user-supplied parameters as well as many clusters. There is a need to examine the input microarray data to initialize such parameters with suitable values [22].

In this article, the researchers seek to get the superior quality of clusters by initializing the centroids. Therefore, to get the finest clustering algorithm, the researchers have opted to perform a comparative analysis for the UCI data sets [23]. Furthermore, they sought to enhance the quality of clusters by investigating the ADMFI (Automatic Generations of Merge Factor of ISODATA). The arbitrary choice of the first seed point of the anticipated clusters is regarded as a significant limitation of the Automatic Generations of Merge Factor (AGMFI) [24]. The problem was dealt with by using the Enhanced

AGMFI to find the first centroid algorithms (see Table 1). Nonetheless, the Enhanced Automatic Generations of Merge Factor of ISODATA yields diverse excellent outcomes with the Silhouette-Coefficients measurement since it does not rely on any selection of the number of clusters [25]. The gene expression data was used to test the effectiveness of the two algorithms.

Table 1: Comparative Analysis of Clustering Quality of Measurement [1].

Data Set	Initial Number of Cluster	Finalized Number of Cluster	Cluster Quality by K-Means	Quality by Enhanced with K-means	Cluster Quality by AGMFI	Cluster Quality by EAGMFI
Serum	10	7	-0.566	12.38	22.78	29.594
Yeast	34	18	-0.43	17.27	43.26	51.38
Simulated	10	6	46.74	55.79	58.695	61.79

References

1. Chandrasekhar T, Thangavel K, Elayaraja E (2011) Performance Analysis of Enhanced Clustering Algorithm for Gene Expression Data. IJCSI International Journal of Computer Science Issues 8: 253-257.
2. Fahim AM, Salem AM, Torkey FA, Ramadan MA (2006) An Efficient Enhanced K-Means Clustering Algorithm. Journal of Zhejiang University-Science A 7: 1626-1633.
3. Basah Al Shboul, Sung-Hyon Myaeng (2009) Initializing K-Means Using Genetic Algorithms. In International Conference on Computational Intelligence and Cognitive Informatics (ICCICI 09) 2: 114-118.
4. Chen Zhang, Shixiong Xia (2009) K-Means Clustering Algorithm with Improved Initial Center. In WKDD. Second International Workshop on Knowledge Discovery and Data Mining: 790-792.
5. Fang Yuan, Zeng-Hui Meng, Hong-Xia Zhang, Chun-Ru Dong (2004) A New Algorithm to Get the Initial Centroids. In Proceedings of 2004 International Conference on Machine Learning and Cybernetics 2: 1191-1193.
6. Anil Chaturvedi, Paul E. Green, J. Douglas Caroll (2001) K-Modes Clustering. Journal of Classification 18: 35-55.
7. Dembele D, Kastner P (2003) Fuzzy C-Means Method For Clustering Microarray Data. Bioinformatics 19: 973-980.
8. Jiang D, Pei J, Zhang A (2005) An Interactive Approach to Mining Gene Expression Data. IEEE Transactions on Knowledge and Data Engineering 10: 1363-1378.
9. Pavan Karteeka K, Allam Appa Rao, AV Dattatreya Rao, GR Sridhar (2008) Automatic Generation of Merge Factor for Clustering Microarray Data. IJCSNS 8: 127.
10. Kohei Arai, Ali Ridho Barakbah (2007) Hierarchical K-Means: An Algorithm for Centroids Initialization for K-Means. Rep Fac Sci Engrg 36: 25-31.
11. Bhattacharya A, De RK (2008) Divisive Correlation Clustering Algorithm (DCCA) For Grouping of Genes: Detecting Varying Patterns in Expression Profiles. Bioinformatics 24: 1359-1366.
12. Abdul Nazeer KA, Sebastian MP (2009) Improving the Accuracy and Efficiency of the K-Means Clustering Algorithm. In Proceedings of the World Congress on Engineering 1:1-3.
13. Lleti R, Ortiz MC, Sarabia LA, Sanchez MS (2004) Selecting Variables for K-Means Cluster Analysis by Using a Genetic Algorithm That Optimises the Silhouettes. Analytica Chimica Acta 515: 87-100.
14. Moh'D Belal Al-Zoubi, Mohammad al Rawi (2008) An Efficient Approach for Computing Silhouette Coefficients. Journal of Computer Science 4: 252.
15. Dunham Margaret H (2006) Data mining: Introductory and Advanced Topics. Pearson Education India.
16. Madhu Yedla, Srinivasa Rao Pathakota, Srinivasa TM (2010) Enhancing K-Means Clustering Algorithm with Improved Initial Center. International Journal of Computer Science and Information Technologies 1: 121-125.
17. Schena Mark, RW Davis (2000) Technology Standards for Microarray Research. In Microarray Biochip Technology edited by Mark Shena 1-18.
18. Wei Zhong, Altun G, Harrison R, Tai PC, Yi Pan (2005) Improved K-Means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property. IEEE Transactions on Nanobioscience 4: 255-265.
19. Lu Y, Lu S, Fotouhi F, Deng Y, Brown SJ (2004) Incremental Genetic K-Means Algorithm and Its Application in Gene Expression Data Analysis. BMC Bioinformatics 5: 172.
20. Swathi M (2017) Clustering Enhancement Using Similarity Indexing to Reduce Entropy. Enliven: Bioinform 4: 001.
21. Anil K. Jain (2010) Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters 31: 651-666.
22. Dunn JC (1973) A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics 3: 32-57.
23. Swathi M (2018) Enhancement of K-Mean Clustering for Genomics of Drugs. Enliven: J Genet Mol Cell Biol 5: 001.
24. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, et al. (2002) An Efficient K-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis & Machine Intelligence 7: 881-892.

25. Siddheswar Ray, Rose H. Turi (1999) Determination of Number of Clusters in K-Means Clustering and Application in Color Image Segmentation. In Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques 137-143.
26. Wilpon J, Rabiner LR (1985) A Modified K-Means Clustering Algorithm for Use in Isolated Work Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 33: 587-594.
27. Swathi M (2017) Drug Prediction of Cancer Genes Using SVM. Enliven: Pharmacovigilance and Drug Safety 4: 001.
28. Mac Queen J (1967) Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1: 281-297.