Enliven Archive
Intensifying Thoughts

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

ISSN: 2379-5700

Clustering Gene Expression Data Using an Advanced Harmony Search-K Means Hybrid Algorithm

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 Sepetember 2018
Accepted Date: 21st November 2018
Published Date: 23rd November 2018

Citation: Titus J (2018) Clustering Gene Expression Data Using an Advanced Harmony Search-K Means Hybrid Algorithm. Enliven: J Genet Mol Cell Biol 5(2): 006.

There are massive volumes of biological data which has accumulated at varied data sources due to recent developments in bioinformatics research. Furthermore, scientists can now conduct a simultaneous analysis of the substantial number of genes across diverse samples using the DNA microarray technology [1]. Additionally, such fields as drug development, disease diagnosis, comparative genomics, and functional genomics have greatly benefitted from the vast amount of expression data that is derived from clustering of microarray data that discloses gene expression patterns which are hidden [2]. The k-means clustering algorithm is commonly relied on to facilitate several practical applications [3]. Nonetheless, the original k-means algorithm has numerous limitations such as it is costly to perform while using it to perform computations and usually generates ideal solutions from the arbitrary choice of the first centroids [4].

Different researchers have suggested numerous techniques to enhance the performance of the k- means algorithm [5]. They developed a harmony search which can be considered as a meta-heuristic optimization algorithm that looks for the whole solution space [6]. The precision levels of the clustering achieved by the current algorithms usually restrict their application in several critical life science applications [7]. In this case, the researchers sought to improve the clustering of gene expression data by recommending a New Harmony Search-K means Hybrid (HSKH) algorithm [8]. The findings from the experiments done indicate that the new algorithm yields clusters with superior precision as compared to the current algorithms [9].

Large quantities of data continue to pile up in different life science and biological databases owing to the rapid developments being witnessed in methods used to collect scientific data [10]. The introduction of DNA microarrays enables scientist to concurrently monitor expression levels of many genes across a host of thousands of samples [11]. By exploring the gene expression data to reveal the hidden patterns, it provides a discrete opportunity to enable people to understand the significance of the biological processes [12]. The data collected can successfully be applied in to conduct diagnosis and classify different diseases. Nonetheless, the intricacy of the massive volume of the genomic data usually leads to some problems related to analyzing the big data banks to discover the hidden patterns [13]. Therefore, researchers have adopted cluster analysis in an attempt to resolve the issue [14]. The process is among the main data mining techniques which are useful with regards to exploring dataset items to establish the natural grouping. Clustering entails dividing a specific set of objects into separate groups [15]. The process is done in a manner that the objects contained in a group are similar while the objects clustered in different group vary significantly based on their characteristics [16].

Microarray Gene Expression (GE) data clustering greatly assists researchers to create a good understanding of the cellular processes, gene regulation, as well as gene functioning [17]. The genes contained in similar group usually demonstrate same expression patterns; besides, they are susceptible to co-regulation [18]. Additional early diagnosis of diseases can be possible since unknown samples can be effectively be classified. Clustering of tissues samples obtained from the healthy and sick people for instances the cancer patients on the basis of expression patterns' similarity can assist to successfully classify the unknown samples which ultimately can help to perform timely diagnosis of diseases [19].

In many practical applications, the k-means algorithm can successfully be applied. Nonetheless, with regards to large sets of data, the original k-means algorithms usually have very high computational complexity. Additionally, the algorithm mostly the lead to the attainment of locally ideal solutions and leads to diverse kind of clusters dependent on the arbitrary selection of the first centroid. The numerous techniques have been suggested to improve the k-means clustering algorithm's performance [20]. Nonetheless, the majority of the users have failed to accept them widely. Additionally, clustering can be considered as an optimization challenge that helps to lower the inter-cluster and optimize the intra-cluster similarities among the numerous data points [21]. Researchers are keen on obtaining globally optimum solutions;

therefore, they are focused on developing a hybrid algorithm that combines the harmony search algorithm [22].

Search optimization which is a meta-heuristic technique together with the k-means algorithm [23]. The new and efficient method used to perform clustering of the microarray gene expression data as proposed in the article is known as the Harmony Search-K means Hybrid (HSKH) algorithm [24]. The technique is intended to assist researchers to effectively manipulate the microarray gene expression data to facilitate retrieval of important information or knowledge. Drug development and disease diagnosis require accurate clustering, but the current conventional clustering methods usually do not produce adequately excellent outcomes [26].

Figure 1: Comparing the performance of the algorithm (Source: Nazeer, KA Abdul, Sebastian MP, SD Madhu Kumar (2013) A novel harmony search-K means hybrid algorithm for clustering gene expression data. Bioinformation 9: 84)

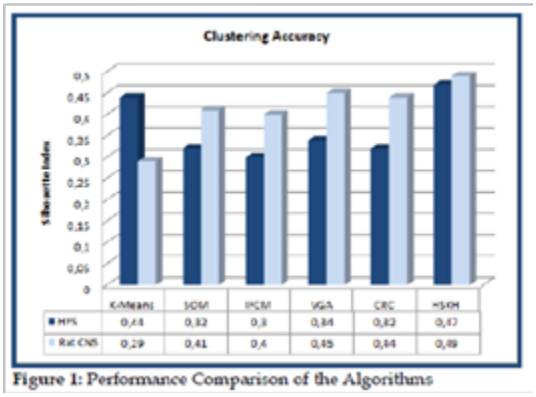

Figure 1: Performance Comparison of the Algorithms

References

1. Nazeer KA, Sebastian M, Kumar SM (2013) A Novel Harmony Search-K Means Hybrid Algorithm for Clustering Gene Expression Data. Bioinformation 9: 84-88.
2. Weiguo S, Liu X (2004) A Hybrid Algorithm for K-Medoid Clustering of Large Data Sets. Proceedings of the 2004 Congress on Evolutionary Computation 1: 77-82.
3. Ben-Dor A, Chor B, Karp R, Yakhini Z (2003) Discovering Local Structure In Gene Expression Data: The Order-Preserving Submatrix Problem. J Comput Biol 10: 373-384.
4. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, et al. (2000) Tissue Classification with Gene Expression Profiles. J Comput Biol 7: 559-583.
5. Slonim DK (2002) From Patterns to Pathways: Gene Expression Data Analysis Comes of Age. Nat Gent 32: 502-508.
6. Daxin Jiang, Chun Tang, Aidong Zhang (2004) Cluster Analysis for Gene Expression Data: A Survey. IEEE Transactions on Knowledge and Data Engineering 16: 1370-1386.
7. 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.
8. Jiang D, Pei J, Zhang A (2005) An Interactive Approach to Mining Gene Expression Data. IEEE Transactions on Knowledge and Data Engineering 2: 1363-1378.
9. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, et al. (2001) Gene-Expression Profiles in Hereditary Breast Cancer. N Engl J Med 344: 539-548.
10. de Souto MC, Costa IG, de Araujo DS, Ludermir TB, Schliep A (2008) Clustering Cancer Gene Expression Data: A Comparative Study. BMC Bioinformatics 9: 497.
11. de Souto MC, Jaskowiak PA, Costa IG (2015) Impact of Missing Data Imputation Methods on Gene Expression Clustering and Classification. BMC Bioinformatics16: 64.
12. Nascimento Andre CA, Ricardo Prudencio BC, Marcilio De Souto BC, Ivan Costa G (2009) Mining Rules for the Automatic Selection Process of Clustering Methods Applied to Cancer Gene Expression Data. In International Conference on Artificial Neural Networks 20-29.
13. Katti Faceli, Andre de Carvalho CPLF, Marcilio de Souto CP (2008) Cluster Ensemble and Multi-Objective Clustering Methods. In Pattern Recognition Technologies and Applications: Recent Advances, 325-343.
14. Mabbott Neil A, J Kenneth Baillie, Helen Brown, Tom C Freeman , David A Hume (2013) An Expression Atlas of Human Primary Cells: Inference of Gene Function from Co-Expression Networks. BMC Genomics 14: 632.
15. Forsati Rana, Mehrdad Mahdavi, Mehrnoush Shamsfard, Mohammad Reza Meybodi. (2013) Efficient Stochastic Algorithms for Document Clustering. Information Sciences 220: 269-291.
16. Senthilnath J, Sushant Kulkarni, Raghuram DR, Meghana Sudhindra, Omkar SN, (2016) A Novel Harmony Search-Based Approach for Clustering Problems. International Journal of Swarm Intelligence 2: 66-86.
17. Arunanand TA, Abdul Nazeer KA, Mathew J. Palakal, Meeta Pradhan (2014) A Nature-Inspired Hybrid Fuzzy C-Means Algorithm for Better Clustering of Biological Data Sets. In 2014 International Conference on Data Science & Engineering (ICDSE) 76-82.
18. Peker Musa (2016) A Decision Support System to Improve Medical Diagnosis Using a Combination of K-Medoids Clustering Based Attribute Weighting and SVM. Journal of Medical Systems 40: 116.
19. Jian W, Yuan W, Cheng D (2015) Hybrid Genetic–Particle Swarm Algorithm: An Efficient Method for Fast Optimization of Atomic Clusters. Computational and Theoretical Chemistry 1059: 12-17.
20. Mulay Preeti (2016) Threshold Computation to Discover Cluster Structure: A New Approach. International Journal of Electrical and Computer Engineering (IJECE) 6: 275-282.
21. Swathi M (2018) Enhancement of K-Mean Clustering for Genomics of Drugs. Enliven: J Genet Mol Cell Biol 5: 001.
22. Ben-Dor A, Shamir R, Yakhini Z (1999) Clustering Gene Expression Patterns. J Comput Bio 6: 281-297.
23. 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.
24. Anil K. Jain (2010) Data Clustering: 50 Years Beyond K-means. Pattern Recognition Letters 31: 651-666.

25. Swathi M (2017) Drug Prediction of Cancer Genes Using SVM. Enliven: Pharmacovigilance and Drug Safety 4: 001.
26. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S (1999) Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods And Application To Hematopoietic Differentiation. Proceedings of the National Academy of Sciences 96: 2907-2912.
27. Kerr G, Ruskin HJ, Crane M, Doolan P (2008) Techniques for Clustering Gene Expression Data. Comput Biol Medicine 38: 283-293.
28. Swathi M (2017) Clustering Enhancement Using Similarity Indexing to Reduce Entropy. Enliven: Bioinform 4: 001.