

AN EFFICIENT GENE SELECTION FOR MICROARRAY DATA USING CLUSTERING ALGORITHM

¹G. BASKAR, ²Dr.P.PONMUTHURAMALINGAM

¹ PhD Research Scholar, ² Associate Professor & Head

Department of Computer Science
Government Arts College (Autonomous)
Coimbatore, Tamil Nadu, INDIA

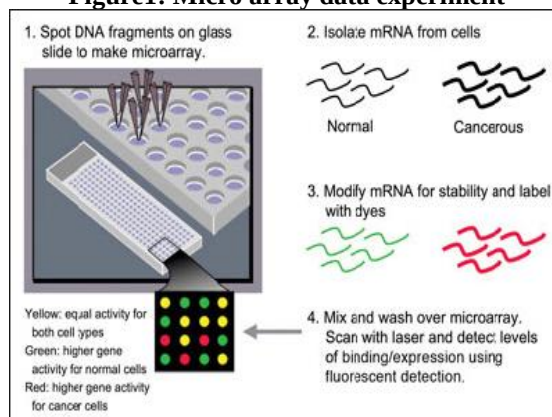
Abstract: -Microarray is an efficient method of gathering data that can be used for expressing the pattern of tens of thousands of genes in a few hours. In data mining two methods commonly applied to microarray data are clustering and classification, the process of data mining is concerned with extracting patterns from the data. In this research paper an analysis is made for data mining algorithm like K-Means, Fuzzy C-Means, and Fuzzy possibilistic algorithm, the experiment is based on microarray datasets. The result demonstrates that Fuzzy possibilistic clustering have much lower error when compared with other algorithms.

Key word: - Data Mining, Clustering, Microarray data, K-Means, Fuzzy

I. Introduction

Data mining in bioinformatics is the study and application of life science, the main reason is the sequence of human genome and the process helpful for impact on modern medicine and the data mining consist a record for large dataset. Data mining algorithm look for data in pattern and the problem applied to number of application area not only in bioinformatics like insurance, medical etc.

Figure1: Micro array data experiment



clustering is the most important unsupervised techniques in data mining which groups the similar data ,while in bioinformatics the similar samples and genes are grouped together and this paper is based on cancer micro array data for cancer diagnosis is an important problem in cancer genomics, gene expression micro array is used to identify candidate gene in various cancer studies. in 20th century scientist came with many ways to study on gene, mapping then sequence ,analysis. But scientist cant applied one at a time but living, for example around 20 thousand gene for an experiment it will take very long time to investigate the gene one at a time, it is great to study many gene one at time and what they could learn in one experiment. genomics is a way to study many genes it is the new word in 21st century one providing a tool for this task is microarray, using

II. Clustering Algorithm

A) K-Means

This is an unsupervised data mining clustering algorithm by (MacQueen, 1967), A well-known cluster problem is solved using this algorithm. It is one of the simplest approach in clustering method, and this is mostly used for clustering data , in this 'K' is look for different cluster. When applied on data set the algorithm break the data in to 'K' cluster if you specify K=5, then 'K' break the dataset in to five clusters.

Algorithm Steps

Step 1 Select 'K' point center.

Step 2 Distances must be calculated with respect to center.

Steps 3 Choose center point to the minimum.

Step 4 Recalculate the new cluster

Step 5 until no change in centroid otherwise repeat step 3.

B) Fuzzy C-Means clustering algorithm:

The algorithm is related to a form of set theory and logic in which predicates may have degrees of applicability, rather than simply being true or false.

Fuzzy logic is rather based on 0 and 1, the idea is said by DR. Lotfi Zadeh in 1960 when he worked on a problem in natural language.

The fuzzy C-Means which minimized intra cluster variation, it has the same drawback like K-Means based on local minimum and the result is with respect to the weight chosen.

Algorithm Steps

Step 1 The number of clusters must be initialized before selection.

Step 2 Before choosing the cluster the point must be assigned.

Step 3 Repeat until it gets converged (i.e. the speed approaches its limit) the coefficient must be changed between each and every convergence.

Step 4 Use formulas for selecting cluster center.

Step 5 Elimination of cluster.

C) Fuzzy possibilistic clustering:

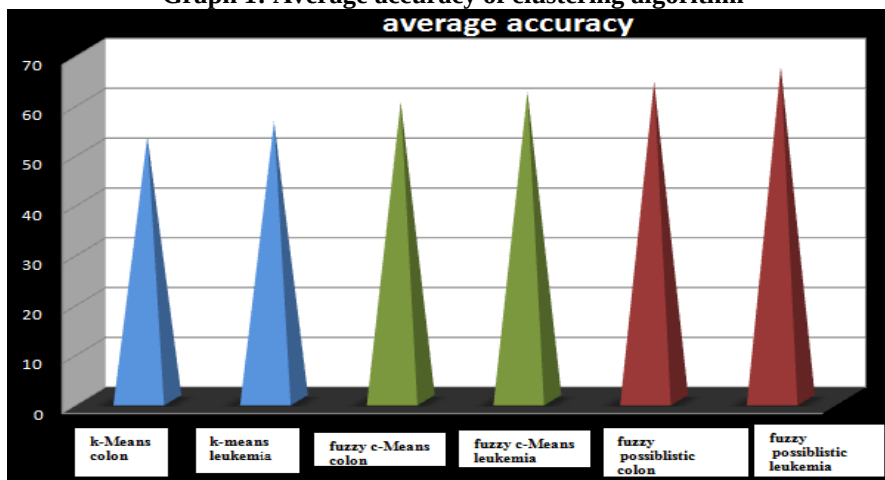
A fuzzy possibilistic algorithm keeps improving the outlier and noise in fuzzy c-means and it will improve the accuracy, this algorithm initializes cluster center and generates coincident clusters.

III. Data Set Description

In this paper cancer data sets are used which are namely colon and leukemia data sets. Taken from <http://cs.binghamton.edu/~lyu/KDD08/data/>. The colon data set consists of 2000 genes and leukemia 7129.

IV. Result over the Data Set

Graph 1: Average accuracy of clustering algorithm
average accuracy



V. Conclusion and Future Work

Microarray data for colon and leukemia are tested with clustering algorithms, and the results show that the fuzzy possibilistic algorithms have more accuracy when compared with both cancer datasets and in the future it can be improved with a modified fuzzy C-Means algorithm.

VI. References

- [1] John, G.H., Kohavi, R., Pfleger, K.: (1994) Irrelevant Feature and The Subset Selection Problem. In: Proceedings of the Eleventh International Conference on Machine Learning, pp. 121–129
- [2] Liu, H., Yu, L.: Toward Integrating Feature Selection Algorithms for Classification and Clustering. (2005) IEEE Transactions on Knowledge and Data Engineering 17(4), 491–502
- [3] Ding, C., Peng, H.: Minimum Redundancy Feature Selection from Microarray Gene Expression Data. In: Proceedings of the Computational Systems Bioinformatics Conference (CSB 2003), pp. 523–529 (2003)
- [4] Yu, L., Liu, H.: Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research 5, 1205–1224 (2004)
- [5] Pepe, M.S., Etzioni, R., Feng, Z., et al.: Phases of Biomarker Development for Early Detection of Cancer. J. Natl. Cancer Inst. 93, 1054–1060 (2001)
- [6] T. Li, C. Zhang, and M. Ogihara, “A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression,” *Bioinformatics*, vol. 20, pp. 2429–2437, 2004

- [7] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 17, no. 4, pp. 491–502, 2005.
- [8] M. Wasikowski and X. Chen, "Combating the small sample class imbalance problem using feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1388–1400, 2010.
- [9] C. A. Davis, F. Gerick, V. Hintermair, et al., "Reliable gene signatures for microarray classification: assessment of stability and performance," *Bioinformatics*, vol. 22, pp. 2356–2363, 2006.
- [10] C. A. Davis, F. Gerick, V. Hintermair, et al., "Reliable gene signatures for microarray classification: assessment of stability and performance," *Bioinformatics*, vol. 22, pp. 2356–2363, 2006.
- [11] J.A. Lozano, J.M. Pena, P. Larranaga, An empirical comparison of four initialization methods for the k-means algorithm, *Lett.* 20 (1999) 1027–1040.
- [12] Nielsen T.O, West R.B, Linn S.C, et al. Molecular characterization of soft tissue tumours: a gene expression study. *Lancet* 2002
- [13] G.Baskar, Dr.P.ponmuthuramalingam "a comparative study and analysis for microarray gene expression data using clustering techniques" June 2013
- [14] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene
- [15] G. Baskar, P. Ponmuthuramali Analysis of Gene Expression Microarray Dataset for Feature Selection: National Conference on Communication Technologies & its impact on Next Generation Computing CTNGC 2012 Proceedings published by International Journal of Computer Applications® (IJCA)
- [16] Y. Saeys, I. Inza, and P. Larranaga, "A Review of Feature Selection Techniques in Bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
- [17] I.H. Witten and E. Frank, *Data Mining - Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 2005.
- [18] Pawan Lingras, Chad West. Interval set Clustering of Web users with Rough k-Means, submitted to the *Journal of Intelligent Information System* in 2002.
- [19] Yeung K.Y, Haynor D.R, Ruzzo W.L. Validating clustering for gene expression data. *Bioinformatics*. 2001.
- [20] Iliak, A., Vlassis, M. & Verbeek, J. (2003), the global kmeans clustering algorithm, *pattern*

Authors profiles

[1] **G.Baskar** received his Master's degree in Information Technology in K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu India in 2008 and M.Phil Degree in Computer Science from Bharathiar University, Coimbatore, Tamil Nadu, India in 2010, and He is currently working towards the PhD degree in Department of Computer Science, Government Arts College, Coimbatore, Tamil Nadu, INDIA in 2011. His area of interest includes Data mining, bioinformatics.



[2] **P.Ponmuthuramalingam** received his Masters Degree in Computer Science from Alagappa University, Karaikudi in 1988 and the Ph.D. in Computer Science from Bharathiar University, Coimbatore. He is working as Associate Professor and Head in Department of Computer Science, Government Arts College (Autonomous), Coimbatore. His research interest includes Text mining, Semantic Web, Network Security and Parallel Algorithm

