Feature Selection for Micro Array Cancer Classification using Clustering Algorithms

K.Prema
Research Scholar
Department of Computer Science
Chikkanna Govt arts college

Dr.A.Kumar kombaiya
Assistant professor
Department of Computer Science
Chikkanna Govt arts college

***Abstract*— Feature selection has in recent times attracted burly interest in knowledge discovery from high-dimensional data. Classification is a data mining technique used to predict group membership for data instances; microarray is the technology which allows researchers to congregate information on various gene expressions. Gene selection for cancer classification is one of the mainly important topics in biomedical field. In this paper, we compare different clustering algorithm for microarray cancer classification techniques based on data mining methodology for perform both accuracy and stability measurement.**

Keywords — *Feature selection, micro array, stability, clustering algorithms..*

I. INTRODUCTION

Mining biological data is an emerging area of intersection between bioinformatics and data mining .bioinformatics have taken a computational approach to understanding biological phenomena. Because these phenomena are typically characterized by large and increasing amount of data, diver and unusual data type, and complex relationship, interpreting biological data requires novel approaches that include multiple tools, new algorithms, resources, etc. in an integrated fashion. Data mining has focused on extracting useful information from large data base, focusing on scalable, robust algorithms and their implementations.[23] At the end of the 1980's a new discipline, named data mining, emerged. The introduction of new technologies such as computers, satellites, new mass storage media and many others have lead to an exponential growth of collected data. Traditional data analysis techniques often fail to process large amounts of -often noisy- data efficiently, in an exploratory fashion. Recently, the collection of biological data has been increasing at explosive rates due to improvements of existing technologies and the introduction of new ones such as the microarrays. These technological advances have assisted the conduct of large scale experiments and research programs. An important issue in data analysis is feature selection. In gene expression analysis the features are the genes. Gene selection is a process of finding the genes most strongly related to a particular class. One benefit provided by this process is the reduction of the foresaid dimensionality of dataset. Moreover, a large number of genes are irrelevant when classification is applied. The danger of overshadowing the contribution of relevant genes is reduced when gene selection is applied. Clustering is the far most used method in gene expression analysis. Tibshirani et al. (1999) and Aas (2001) provide a classification of clustering methods in two categories: one-way clustering and two-way clustering. Methods of the first category are used to group either genes with similar behavior or samples with similar gene expressions. Two-way clustering methods are used to simultaneously cluster genes and samples. Hierarchical clustering is currently the most frequently applied method in gene expression analysis. An important issue concerning the application of clustering methods in microarray data is the assessment of cluster quality. Many techniques such as bootstrap, repeated measurements, mixture model-based approaches, sub-sampling and others have been proposed to deal with the cluster reliability assessment.

II. RELATED WORKS

In[1] Anjan Goswami, Ruoming Jin, Gagan Agrawal present a new algorithm which typically requires only one or a small number of passes on the entire dataset, and provably produces the same cluster centers as reported by the original k-means algorithm. The algorithm uses sampling to create initial cluster centers, and then takes one or more passes over the entire dataset to adjust these cluster centers and provide theoretical analysis that the cluster centers thus reported are the same as the ones computed by the original k-means algorithm. Experimental results from a number of real and synthetic datasets as compared to k-means. In[3] Greg Hamerly propose a new acceleration for exact k-means that gives the same answer, but is much faster in practice. Like Elkan's accelerated algorithm, The algorithm avoids distance computations using distance bounds and the triangle inequality and the algorithm uses one novel lower bound for point-center distances, which allows it to eliminate the innermost k-means loop 80% of the time or more in our experiments. On datasets of low and medium dimension (e.g. up to 50 dimensions), the algorithm is much faster than other methods, including methods based on low-dimensional indexes, such as k-d trees. Other advantages are that it is very simple to implement and it has a very small memory overhead, much smaller than other accelerated algorithms.

In [9] Yanchang Zhao, Chengqi Zhang, Shichao Zhang, and Lianwei Zhao proposed Subspace clustering is a challenging task in the field of data mining. Traditional distance measures fail to differentiate the furthest point from the nearest point in very high dimensional data space. To tackle the minimal subspace distance which measures the similarity between two points in the subspace are nearest to each other. It can discover subspace clusters implicitly when measuring the similarities between points and use the new similarity measure to improve traditional k-means algorithm for discovering clusters in subspaces. By clustering with low-dimensional minimal subspace distance first, the clusters in low-dimensional subspaces are detected. Then gradually increasing the dimension of minimal subspace distance, the clusters get refined in higher dimensional subspaces. The experiments on both synthetic data and real data show the effectiveness of the proposed similarity measure and algorithm. improve the new techniques and discover sub space clustering the distance measured to the nearest point of data.

III. METHODOLOGY

The objective of this paper is to provide a comparative evaluation of clustering algorithms for its accuracy and convergence rate with real data set colon and leukemia dataset. The clustering algorithms K-Means, Fuzzy C-Means, Modify Fuzzy C-Means, Fuzzy Possibilistic C-Means, Modify Fuzzy Possibilistic C-Means, Kernel-based Fuzzy C-Means, and Modify Kernel-based Fuzzy C-Means, performance are evaluate based on the time, space and accuracy. Two different cancer datasets to make a study of k-family algorithms, the leukemia data set is a collection of gene expression measurements from 72 leukemia (composed of 62 bone marrow and 10 peripheral bloods) samples report by Golub. It contains an initial training set composed of 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloblastic leukemia (AML).here two variants of leukemia dataset one with 50-genes and another one with 3859-Genes. The colon dataset is a collection of gene expression measurements from 62 colon biopsy sample reported by alon. It contains 22 normal and 40 colon cancer samples. The dataset consists of 2000 genes. The dataset has taken from uci repository.[23]

Modify Kernel based Fuzzy C-Means

Step 1: Fix c, t_{max}, $m > 1$ and $\varepsilon > 0$ for some positive constant;

Step 2: Initialize the membership u^0_{ik}

$$J_m = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^m \|X_k - V_i\|^2$$

Step 3: For $t = 1, 2\ldots, t_{max}$, do:

(a) Update all prototypes V^t_i

(b) Update all memberships U^t_{ik}

(c) Compute $E^T = \max_{i,k}|u^t_{ik} - u^{t-1}_{ik}|$, If $ET \le \varepsilon$,

$$U \in \{u_{ik} \in [0,1] | \sum_{i=1}^{c} u_{ik} = 1 \forall k \quad \text{and} \quad 0 < \sum_{k=1}^{N} u_{ik} < N, \forall i\}$$

Step 4: Stop: else t=t+1.

They are both defined over $R^n \times R^n$. Obviously, due to the detail that the know kernel functions requirement to solve the problems in the kernel space only by means of kernel functions, i.e., the inner creation of the transform function ϕ. Typically this is called "kernel trick". There are two kinds of KFCM. When the prototypes $o_i$areformed in the kernel space, this kind of KFCM is mentioned as KFCM-k (whereK standup for the kernel space). The main function of KFCM-K

$$Q = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \left\|\varphi(x_j) - o_i\right\|^2$$

$$u_{ij} = 1 \Big/ \sum_{h=1}^{c}\left(d\varphi 2_{ij}/d\varphi 2_{ij}\right) 1/(m-1) \quad (3.26)$$

Where

$$d\varphi_{ij}^2 = k(x_j, x_j) - \frac{2\sum_{h=1}^{n} u_{ih}^m k(x_h, x_j)}{\sum_{h=1}^{n} u_{ih}^m} + \frac{\sum_{h=1}^{n}\sum_{l=1}^{n} u_{ih}^m u_{il}^m k(x_h, x_l)}{(\sum_{h=1}^{n} u_{ih}^m)^2}$$

Another type of KFCM limitations that the prototypes in the kernel space is basically mapped from the unique data space otherwise the feature space. That is, the function is defined as

$$Q = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \left\|\varphi(x_j) - \varphi(o_i)\right\|^2$$

This type of KFCM is mentioned as KFCM-F (with F referredto feature space/data space). Naturally, only k(x, y) = exp $(-\|x - y\|^2/r^2)$Gaussian kernel is applied in KFCM, and sincek(x, x) = 1 for Gaussian kernel

$$\left\|\varphi(x_j) - \varphi(o_i)\right\|^2 = <\varphi(x_j), \varphi(x_j)> + <\varphi(o_i), \varphi(o_i)> -2 <\varphi(x_j), \varphi(o_i)>$$

$$= k(x_j, x_j) + k(o_i, o_i) - 2k(x_j, o_i)$$

$$= 2(1 - k(x_j, o_i))$$

Here,$K(X_j,O_i)$ can be considered as a robust distance measurement derived in the kernel space. For these KFCMF applying Gaussian kernels, iteratively update the prototypes and memberships as

$$= \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \left(-k(x_j, o_i)\right)$$

IV. RESULTS AND DISCUSSION

The performance of K-Means, Fuzzy C-Means, Modify Fuzzy C-Means, Fuzzy Possibilistic C-Means, Modify Fuzzy Possibilistic C-Means, Kernel Based Fuzzy C-Means, and Modify Kernel Based Fuzzy C-Means is comparable with respect to time, memory space, correctly classified, and

average accuracy for colon, leukemia data sets and they are listed in table1 and table 2

Table 1 Colon Data set

Result over Colon data set					
S.No	**Clustering algorithm**	**Time (Sec)**	**Memory Space (MB)**	**Correctly Classified**	**Average Accuracy**
1	K-Means	0.2406	1270	33	53
2	Fuzzy C-Means	0.8990	1294	40	60
3	Modify Fuzzy C-Means	0.1996	1347	43	64
4	Fuzzy Possibilistic C-Means	0.0513	1348	45	66
5	Modify Fuzzy Possibilistic C-Means	0.0541	1353	50	70
6	Kernel Based Fuzzy C-Means	0.6871	1340	56	78
7	Modify Kernel Based Fuzzy C-Means	0.3180	1405	61	84

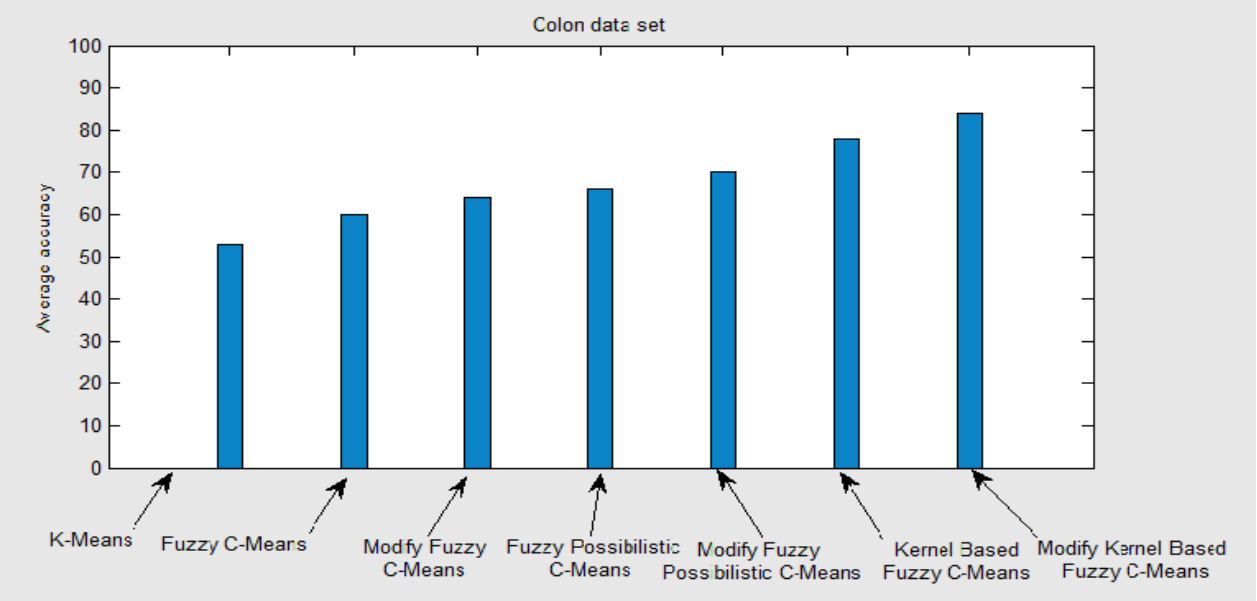

Fig. 1. Graph Average accuracy of Colon Datasets

Table 2 Leukemia Data set

Result over Leukemia data set					
S.No	**Clustering algorithm**	**Time (Sec)**	**Memory Space (MB)**	**Correctly Classified**	**Average Accuracy**
1	K-Means	0.7577	1283	35	56
2	Fuzzy C-Means	0.1888	1312	41	62
3	Modify Fuzzy C-Means	0.6793	1367	45	67
4	Fuzzy Possibilistic C-Means	0.0076	1371	46	68
5	Modify Fuzzy Possibilistic C-Means	0.0119	1284	51	73
6	Kernel Based Fuzzy C-Means	0.6730	1339	60	80
7	Modify Kernel Based Fuzzy C-Means	0.3259	1424	64	88

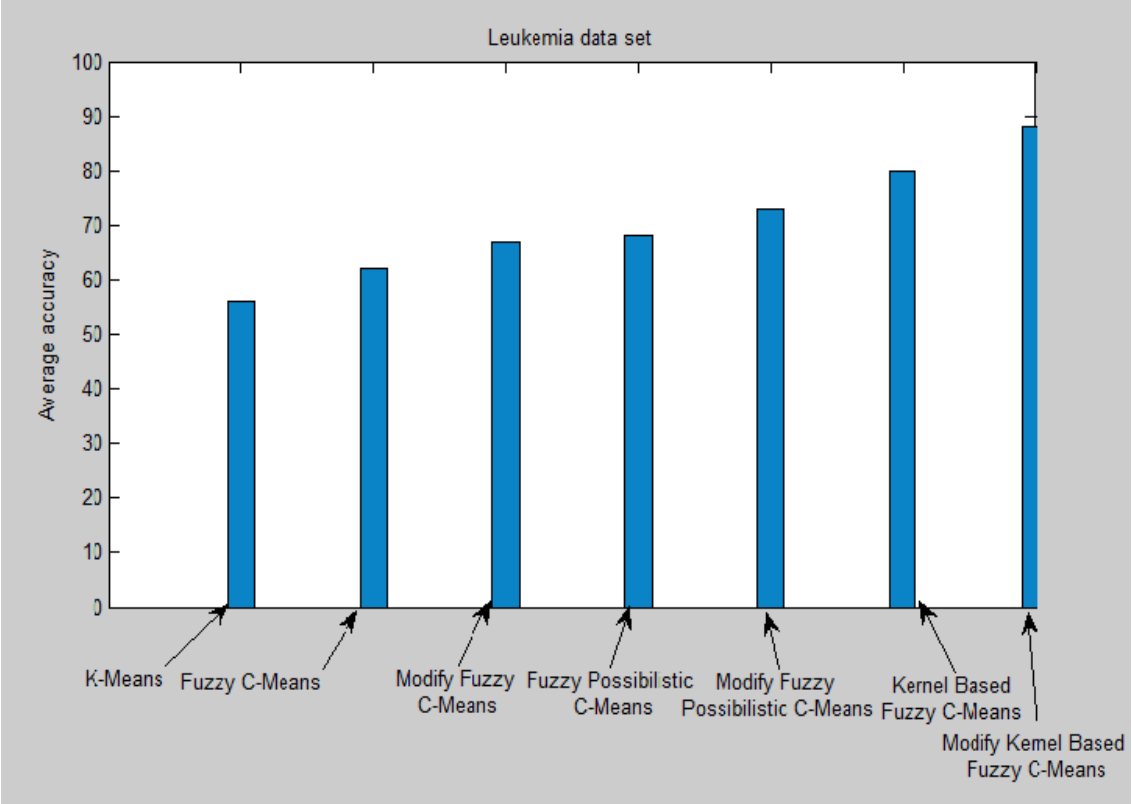

Fig. 2. Graph Average accuracy of Leukemia Data set

The results show the average accuracy of the clustering algorithm for cancer datasets(colon, leukemia).Cluster Algorithm's comparison performance evaluated in this paper shows that the the performance of K-Means, Fuzzy C-Means, Modify Fuzzy C-Means,Fuzzy Possibilistic C-Means, Modify Fuzzy Possibilistic C-Means, Kernel Based Fuzzy C-Means, and Modify Kernel Based Fuzzy C-Means is comparable. Modify Kernel Based Fuzzy C-Means has several advantages over related techniques, and Modify Kernel Based Fuzzy C-Means found clusters with much lower error when compared with other methods and has good accuracy.

V. CONCLUSION

This research is concerned with the study and analysis of data mining techniques for clustering and classification algorithm based on stable gene selection using microarray data. The average accuracy of the clustering algorithm for cancer datasets (colon, leukemia, prostate and lung) evaluated in this thesis work shows that the accuracy of these algorithms is good for the lung data set . However, the performance of K-Means, Fuzzy C-Means, Modify Fuzzy C-Means, Fuzzy Possibilistic C-Means, Modify Fuzzy Possibilistic C-Means, Kernel-based Fuzzy C-Means, and Modify Kernel-based Fuzzy C-Means is comparable. Modify Kernel-based Fuzzy C-Means has several advantages over related techniques, and Modify Kernel-based Fuzzy C-Means found clusters with much lower error when compared with other methods. The gene selection accuracy of clustering algorithm is not optimal. So, the research focuses on classification, using a feature selection algorithm

References

[1] "Anjan Goswami. Department of Computer Science and Engineering" Fast and Exact Out of-Core and Distributed K-Means Clustering 2001

[2] Brendan J. Frey and Delbert Dueck. Clustering by Passing Messages between Data Points. Science 315, 972 (2007).

[3] Greg Hamerly "Making k-means even faster"2010 academic.research.microsoft.com

[4] Guha, S., Rastogi, R., and Shim K. CURE: An Efficient Clustering Algorithm for Large Databases. In Proceedings of the ACM SIGMOD Conference (1998).

[5] Harleen Kaur and Siri Krishan Wasan, " Empirical Study on Applications of Data Mining Techniques in Healthcare" Journal of Computer Science 2 (2): 194-200, 2006, ISSN 1549-3636, © 2006 Science Publications.

[6] Sukjoon Yoon, Young Yang, Jiwon Choi and Seong, "Large scale data mining approach for gene-specific standardization of microarray gene expression data", Vol.22 no 23 2006, Pages 2898-2904.

[7] Sunita Jahirabadkar, and Parag Kulkarni, "ISC – Intelligent Subspace Clustering,

[8] A Density based Clustering approach for High Dimensional Dataset" 2009.

[9] Wagsta K, Cardie C, Rogers S, Schroedl S: Constrained K-means Clustering with Background Knowledge. Proceedings of 18th International Conference on Machine Learning (ICML-01) 2001:577-584.

[10] Yanchang Zhao, Chengqi Zhang, Shichao Zhang, and Lianwei Zhao "k-means algorithm for discovering clusters in subspaces" 2006

[11] Yeung K, Medvedovic M, Bumgarner R: Clustering gene-expression data with repeated Measurements Genome Biology 2003.

[12] Yeung K.Y, Haynor D.R, Ruzzo W.L. Validating clustering for gene expression data. Bioinformatics. 2001.

[13] Zhang Y. , Mao J. and Xiong Z.: An efficient Clustering algorithm, In Proceedings of Second International Conference on Machine Learning and Cyber netics, November 2003.

[14] Zhenjie Zhang, Bing Tian Dai, Anthony K.H. Tung On the Lower Bound of Local Optimums in K-Means Algorithm200