International Journal of Computer and Electrical Engineering, Vol. 1, No. 2, June 2009
1793-8163

Gene Expression Analysis Using Clustering

Kumar Dhiraj and Santanu Kumar Rath

***Abstract*—Data Mining has become an important topic in effective analysis of gene expression data due to its wide application in the biomedical industry. In this paper, k-means clustering algorithm has been extensively studied for gene expression analysis. Since our purpose is to demonstrate the effectiveness of the k-means algorithm for a wide variety of data sets, Two pattern recognition data and thirteen microarray data sets with both overlapping and non-overlapping class boundaries were taken for studies, where the number of features/genes ranges from 4 to 7129 and number of sample ranges from 32 to 683. The number of clusters ranges from two to eleven. For pattern recognition, We use IRIS and WBCD data and for microarray data we use serum data (Iyer et. al.), yeast data (Cho et. al), leukemia data (Golub et. al), breast data (Golub et. al), Lymphoma data (Alizadeh et al.), lung cancer (Bhattacharjee et. al), and St. Jude leukemia data (Yeoh et. al). To identify common subtypes in independent disease data, four different types of breast data (Golub et. al) and four Diffused Large B-cell Lymphoma (DLBCL) data were used. Clustering error rate (or, clustering accuracy) is used as evaluation metrics to measure the performance of k-means algorithm.**

***Index Terms*— Bio-informatics, Cancer-Genomics, Clustering, Cluster validation, Data-mining, Gene-expression, K-means algorithm, Microarray.**

I. INTRODUCTION

DNA microarray [17] has become an important and widely used technology since it enables the possibility of examining the expressions of thousands of genes simultaneously in a single experiment. A key step in the analysis of gene expression data is the detection of gene groups that manifest similar expression patterns. The main algorithmic problem here is to cluster multi-conditions gene expression patterns. Basically, a cluster algorithm partitions entities into groups based on the given features of the entities, so that the clusters are homogeneous and well separated. Several algorithms require certain parameters for clustering, such as the number of clusters and cluster shapes.

In recent years, a number of simple search based clustering [21], [23], [25], [29] (e.g., PAM, UPGMA, CLICK, CAST, DBSCAN, DHC etc.) and complex search based clustering methods (e.g., evolutionary algorithms (EAs) [49], Genetic Algorithm (GA) [34~36], Simulated Annealing (SA) [50], [60-62] and Tabu search (TS) [47], [57~59] etc.) have been proposed. Clustering have application in "fields as diverse as VLSI design, image processing, neural networks, machine learning, job-shop scheduling methods [51~56]. A variety of clustering methods have been also proposed for the mining of gene expression data [18~22]. Although a number of clustering methods have been studied in the literature, they do not deal with clustering accuracy rather they generally deals with clustering validation metrics [21] to assess their performances. The problem here with existing approaches is how to decide about accuracy of a clustering algorithm even if a cluster validation index is optimal [21].

The rest of the paper is organized as follows: In section II, clustering is introduced. An overview of K-means clustering algorithm is described in section III. Section IV discusses review on variants of K-means clustering algorithm. Experimental work conducted to evaluate the performance of the K-means clustering algorithm is presented in section V. Section VI deals with the result and discussion. Conclusions and future works are given in section VII.

Manuscript received on January 15, 2009. This work was carried out at software engineering and Petri-net simulation Laboratory, National Institute Technology Rourkela, Orissa, 769008, INDIA.

Kumar Dhiraj is with the Computer Science and Engineering, National Institute Technology Rourkela, Orissa, 769008, INDIA. He is a Research scholar working in the area of Data mining and Bioinformatics. (Phone: +919853388520; fax: 0661-2464356;).

Santanu Kumar Rath is with the Computer Science and Engineering, National Institute Technology Rourkela, Orissa, 769008, INDIA. He is a senior professor in the NIT Rourkela INDIA.

II. CLUSTERING

In this section, we first define the clustering and how we can represent multivariate gene expression as a clustering problem.

The clustering problem is defined as the problem of classifying n objects into K clusters without any a priori knowledge. Let the set of n points be represented by the set S and the K clusters be represented by $\{C_1, C_2, \ldots, C_K\}$. Then

$$C_i \neq \phi \quad \text{for } i = 1,2,\ldots,K,$$
$$C_i \cap C_j = \emptyset \quad \text{for } i = 1,\ldots,K, \quad j = 1,\ldots,K \text{ and } i \neq j$$
$$\text{and } \bigcup_{i=1}^{K} C_i = S.$$

A. Problem Definition

The problem of multivariate gene expression clustering can be described briefly as follows. Given a set of genes with unique identifiers, a vector $E_{i,} = \{E_{i1}, E_{i2}, \ldots, E_{in}\}$ is associated with each gene i, where $E_{ij,}$ represents the response of gene i under condition j. The goal of gene expression clustering is to group genes based on similar expressions over all the conditions. That is, genes with similar corresponding vectors should be classified into the same cluster.

In next section we discuss about k-means Clustering algorithm and how we can use this for multivariate gene expression analysis.

III. K-MEANS CLUSTERING ALGORITHM

The K-means algorithm [23], one of the most widely used clustering techniques. The steps of the K-means algorithm are described in brief as follows:

Step 1: Choose K initial cluster centers $\{Z_1, Z_2, \ldots, Z_K\}$ randomly from the n points $\{X_1, X_2, \ldots, X_n\}$.

Step 2: Assign point $\{X_i\}$, $i = \{1,2,\ldots,n\}$. to cluster $\{C_j\}$.

$j \in \{1,2,\ldots,K\}$, *iff*

$\|X_i - Z\| < \|X_i - Z_p\|$, $p = \{1,2,\ldots,K\}$, *and* $j \neq p$.

Ties are resolved arbitrarily.

Step 3: Compute new cluster centers $\{z_1^*, z_2^*, \ldots, z_k^*\}$ as follows:

$$z_i^* = \frac{1}{n} \sum_{x_j \in c_i} X_j \qquad i = 1,2,\ldots,K,$$

Where n_i= is the number of elements belonging to cluster C_i.

Step 4: If $z_i^* = Z_i$, i=1, …, K then terminate. Otherwise repeat from step 2.

Note that in case the process does not terminate at step 4 normally, then it is executed for a maximum "*fixed number of iterations*".

IV. REVIEW ON VARIANTS OF K-MEANS CLUSTERING ALGORITHM

The K-means algorithm is the best-known squared error-based clustering algorithm [26], [27]. It is very simple and can be easily implemented in solving many practical problems [18~22], [51~56]. It can work very well for compact and hyper- spherical clusters. The time complexity of K-means is *O (N*K*d)* and space complexity is *O (N + K)*. Since K and d are usually much less than K, K-means can be used to cluster large data sets. Parallel techniques for K-means are developed that can largely accelerate the algorithm [28]. The drawbacks of K-means are also well studied, and as a result, many variants of K-means have appeared in order to overcome these obstacles. We summarize some of the major disadvantages as follows:

1) There is no efficient and universal method for identifying the initial partitions in K-means clustering algorithm. The convergence centroids vary with different initial points and that may results in suboptimal solution. Studies like K-Harmonic means [46] and Tabu K-Harmonic means (Tabu-KHM) [47] clustering solves the initialization problem trapping to the local minima is still a problem of clustering. This particular limitation of K-means was also studied in ref. [26], [27], [29~33], [45].

2) The iteratively optimal procedure of K-means cannot guarantee convergence to a global optimum. The stochastic optimal techniques, like simulated annealing (SA), Tabu search and genetic algorithms can be clubbed with K-means to find the global optimum. K. Krishna and M. N. Murty proposed a novel hybrid genetic algorithm (GA) called, genetic K-means algorithm (GKA) [34] which finds a globally optimal partition of a given data into a specified number of clusters. This hybrid GA circumvent expensive crossover operations by using a classical gradient descent algorithm used in clustering viz., K-means algorithm. GKA defines two operators; first is K-means operator which was used as a search operator instead of crossover and second is a biased mutation operator which is specific to clustering problem called distance-based-mutation. Unlike K-means, GKA converges to a global optimal solution and this was proved by using finite Markov chain theorem in paper [34]. It is also shown in the paper [34] that GKA searches faster than some of the other evolutionary algorithms used for clustering. Minimization of the Total Within Cluster Variation (TWCV) is one of the important and tough tasks in partitional based clustering. This issue was well addressed and handled in GKA [34].

An improved version of GKA, Fast Genetic K-means Algorithm (FGKA) was proposed by Shiyong and Fotouhi in [35]. Experiments indicate that, while K-means algorithm might converge to a local optimum [24], both FGKA and GKA always converge to the global optimum eventually but FGKA runs much faster than GKA. As similar to SGA (Simple GA), FGKA starts with the initialization phase, which generates the initial population P0. But the way in which next generation population is explored in FGKA that makes FGKA different with SGA and other variants of GA like GKA. The population in the next generation Pi+1 is obtained by applying the following genetic operators sequentially: the selection, the mutation and the K-means operator on the current population Pi. The evolution takes place until the termination condition is reached. The initialization phase randomly generates the initial population P0 of Z solutions, which might end up with illegal strings. Illegal strings, however, are permitted in FGKA, but were considered as the most undesirable solutions by defining their TWCVs as $+\infty$ and assigning them with lower fitness values. The flexibility of allowing illegal strings in the evolution process avoids the overhead of illegal string elimination as in [34], and thus improves the time performance of the algorithm.

Incremental Genetic K-means Algorithm (IGKA) [36] was an extension to previously proposed clustering algorithm, the FGKA. IGKA outperforms FGKA when the mutation probability was small. The main idea of IGKA was to calculate the objective value TWCV and to cluster centroids incrementally whenever the mutation probability was small. IGKA inherits the salient feature of FGKA of always converging to the global optimum.

Improved version of K-means which is combined with TS and SA has been discussed in paper [47] and [50] respectively.

3) K-means algorithm is also sensitive to outliers and noise. Even if an object is quite far away from the cluster centroid, it is still forced into a cluster and, thus, distorts the cluster shapes. Ref. [29] and [32], Both considers the effects of outlier in clustering procedure. Ref. [32] gets rid of clusters with few objects. The splitting operation of [32] eliminates the possibility of elongated clusters typical of K-means. Ref.

[29] utilizes real data points (medoids) as the cluster prototypes and avoids the effect of outliers. In Ref. [37], a K-medoids algorithm is presented. In this cluster centroid is searched using the discrete 1-medians.

4) The K-means algorithm is limited to numerical variables. In paper [29], [38] and [39] different variants of K-means algorithm has been used which can be applied to categorical data. The Proposed K-medoids [37] and K-modes algorithm operates in a similar way as K -means.

5) Determining the optimal number of clusters in a set of data is prerequisite to K-means clustering algorithm. Ref. paper [48] addresses this problem. In this paper an algorithm called G-means is introduced to find k. G-means algorithm is described as follows: The G-means algorithm starts with a small number of k-means centers, and grows the number of centers. Each iteration of the algorithm splits into two centers whose data appear not to come from a Gaussian distribution. Between each round of splitting, we run k-means on the entire dataset and all the centers to refine the current solution. We can initialize with just k=1, or we can choose some larger value of k if we have some prior knowledge about the range of k. It is important to note in the paper [48] that they assume each cluster to be Gaussian distributed, and the only other "intuitive parameter" is "the standard statistical significance level".

Several recent advances on K-means and other squared-error based clustering algorithms with their applications can be found in [40~45].

The rest of this paper focuses on my own experiment to application of K-means clustering algorithm to pattern recognition data as well as gene expression data.

V. Experimental work

In this section, we describe the datasets used for assessing the performance the k-means algorithm, which are listed in Table 1, together with some of their relevant characteristics, such as number of classes, number of features/genes, and number of items/samples. The first two datasets represent pattern recognition data, while the other represents gene-expression microarray data. The source of the data has been also given in Table I.

B. Pattern recognition data

1. Iris data

The IRIS data set [1] is a well known and well used benchmark data set used in the machine learning community. Since the data are labeled, this data has been extensively used for Classification purpose in previous work. For clustering algorithm we have not used the labeled information available to iris data. The size of the data is [150x4].The characteristic of this data is it's having some overlap between classes 2 and 3. Since the number of classes in this data is three therefore the value of the K chosen to be 3. The original data can be obtained from UCI repository websites: http://archive.ics.uci.edu/ml/datasets/Iris).

1. Wisconsin Breast Cancer Data

Breast cancer is one of the most common cancers in women and a frequent cause of death in the 35-55 year age group. The presence of a breast mass is an alert sign, but it does not always indicate a malignant cancer. It contains 699 instances of cytological analysis of fine needle aspiration from breast tumors.

Each case comprises 11 attributes: a case ID, cytology data (normalized, with values in the range 1-10) and a benign/malignant attribute. The number of benign instances is 458 and the number of malignant instances is 241. We removed sixteen instances of cases (14 benign, 2 malignant) with missing values from the data set [2], [16].

C. Gene-expression microarray data

The gene-expression datasets and a very short description of their content are given in Table I. Further biological details about these data sets can be found in the referenced papers. Most data were processed on the Human Genome U95 Affymetrix c microarrays. The leukemia dataset is from the previous-generation Human Genome HU6800 Affymetrix c microarray.

2. Iyer data (serum data)

This data set is described and used in [3], [4], [8], [14]. It can be downloaded from: www.sciencemag.org/feature/data/984559.shl and corresponds to the selection of 517 genes whose expression varies in response to serum concentration in human fibroblasts and is classified into 11 groups.

3. Cho data (yeast data)

In this data set [5~8], [14] the expression pro-files of 6200 yeast genes were measured every 10 min during two cell cycles in 17 hybridization experiments [5]. Tavazoie et al. [6] used the yeast data of 2945genes. He selected the data after excluding time points 90 and 100 min. from Yeast data. We use 386 genes from Cho data for analysis of experimental work [14]. Total no. of classes for this data is 5.

4. Leukemia(Golub experiment)

The Leukemia dataset belongs to two types of Leukemia cancers [14], [23]: Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). It consists of 72 samples of 7129 gene expressions each. The data has 47 samples belong to ALL cancer class and 25 samples belong to AML cancer class.

5. Lymphoma (Alizadeh et al. Experiment)

In this dataset, microarray was utilized to conduct a systematic characterization of gene expression patterns in the three most prevalent adult lymphoid malignancies: DLBCL [9~11] (47 samples), FL (9 samples) and CLL (11 samples). In addition, 29 non-lymphoma (normal) samples are also involved because of the suspected correlation between them and the three malignances. Each sample consists of gene expression values of 4026 genes. We use DLBCL A, DLBCL B, DLBCL C, DLBCL D to analyze our experimental work. All these Data can be downloaded from www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.

6. Breast data

The microarray breast data [9], [15] used in this paper can be downloaded from www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. This data is of dimension [98x1213] and total number of classes for this data is three.

7. Lung Cancer

This microarray datasets [9], [12] is of dimension [197x581]. It includes 4 known classes: 139 adenocarcinomas (AD), 21 squamous cell carcinomas (SQ), 20 carcinoids (COID), and 17 normal lungs (NL). The AD class is highly heterogeneous, and substructure is known to exist, although not well understood [4].

8. St. Jude Leukemia data

This datasets [9], [13] is of diagnostic bone marrow samples from pediatric acute leukemia patients corresponding to 6 prognostically important leukemia subtypes: 43 T-lineage ALL; 27 E2A-PBX1, 15 BCR-ABL, 79 TEL-AML1, and 20 MLL rearrangements; and 64 "hyperdiploid > 50" chromosomes.

VI. Result and Discussions

Table 1 represents summary of k-means clustering algorithm result for fifteen datasets (fig. 1). It consists some of the relevant characteristics, such as number of classes, number of features/genes and the number of item samples. These datasets are having both overlapping and non-overlapping class boundaries, where the number of features/genes ranges from 4 to 7129 and number of sample ranges from 32 to 683. The number of cluster ranges from 2 to 11. Out of fifteen datasets maximum accuracy was reported for IRIS data while least accuracy data was reported for DLBCL D. As far as Gene Expression data is concerned maximum accuracy was achieved in case of St. Jude Leukemia data.

Table 2 represents the result of IRIS data. We have achieved accuracy up to 88.67%. The total count error in this case is 17. Note that we have achieved 100% accuracy for cluster 1.

Table 3 represents the result of WBCD data. 62.81% accuracy we have achieved in this case.

Table 4 represents the result of Yeast data (Cho data). We have achieved accuracy up to 60.88%. Note that cluster 4 is having accuracy approx. 12%. The reason for this is the data belonging to this cluster are very overlapping in nature with other clusters.

Table 5 represents the result of Leukemia data. We have achieved accuracy up to 59.72%. The Golub et al.'s microarray data set is very challenging because the appearance of the two types of acute leukemia is highly similar in nature. This was the reason we have not much achieved more accuracy in this case. One probable solution to deal with this problem is that we can use dimensionality reduction techniques to reduce the number of feature.

Table 6 represents the result of Serum data (Iyer data). We have achieved accuracy up to 51.84%.

Table 7 to Table 10 represents the result of subtypes of Breast data. We have achieved maximum accuracy for Breast Multi data A (79.61%) whereas the least accuracy for Breast data B (53.06%). The reason of the less accuracy could be probably Breast data B is more overlapping in nature and is having nonlinear structure.

Table 11 to Table 14 represents the result of subtypes of DLBCL data (Diffused Large B-cell Lymphoma). DLBCL D is of highly overlapping nature and that's why we have achieved least accuracy 42.64% in this case. The data of DLBCL B is of highly distinctively separated in nature compare to other DLBCL(A, C, D) and that is the reason we have achieved higher accuracy in case of DLBCL B.

Table 15 represents the result of Lung Cancer. We have achieved accuracy up to 72.08%. In this, cluster 2 and cluster 4 are highly separable in nature compare to cluster 1 and cluster 3. We achieve approx. 95% accuracy for cluster2 and cluster 4 whereas for cluster3 we got least (47%).

Table 16 represents the result of St. Jude Leukemia data. We have achieved accuracy up to 85.08%. The data in this case is of highly separable in nature. We achieve 100% accuracy for cluster and least one we got for cluster 5.

VII. Conclusion and Future Work

Clustering is an efficient way of analyzing information from microarray data and K-means is a basic method for it. K-means can be very easily applied to Microarray data. Depending on the nature and complexity of the data performance of K-means varies. We achieve maximum accuracy for IRIS data where as lowest for DLBCL D.

K-means has some serious drawbacks. Many papers have presented in past to improve K-Means. In the future we are planning to study K-Means clustering with other heuristic based search methods like SA and GA or some others.

References

[1] E. Anderson, "The IRISes of the Gaspe Penisula, " Bulletin of the American IRIS society, vol. 59, 1939, pp. 2-5.
[2] http://archive.ics.uci.edu/ml/datasets
[3] http://www.sciencemag.org/feature/data/984559.shl
[4] Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F, Trent, J. M., Staudt, L. M., Hudson Jr, J., Bogosk, M. S. et al., "The transcriptional program in the response of human fibroblast to serum", Science, 283, 1999, pp. 83–87.
[5] Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. and Davis, R. W., "A genome-wide transcriptional analysis of the mitotic cell cycle", Mol. Cell., 2, 1998, pp. 65–73.
[6] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R.J. and Church, G. M., "Systematic determination of genetic network architecture", Nat. Genet., 22, 1999, pp. 281–285.
[7] Doulaye, Dembele and Philippe Kastner, "Fuzzy C-means method for clustering microarray data", Bioinformatics, vol. 9 no. 8, 2003 pp. 973-980.
[8] http://www.cse.buffalo.edu/faculty/azhang/Teaching/index.html
[9] http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
[10] A. A. Alizadeh et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, " Nature, vol. 43, 2000, pp. 503-511.
[11] Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, , and E. Lander:, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression". Science 286(5439), 1999, 531–537.
[12] Bhattacharjee, A., W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson, 'Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct

Adenocarcinomas Sub-classes". Proceedings of the National Academy of Sciences 98(24), 2001, pp. 3790–13795.

[13] Yeoh, E. J., M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.-H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling". Cancer Cell 1(2), 2002.

[14] http://www-igbmc.u-strasbg.fr/projets/fcm

[15] Y. Hoshida, J. P. Brunet, P. Tamayo, T. R. Golub, J. P. Mesirov, "Subclass mapping: identifying common subtypes in independent disease data sets", PLoS ONE, Vol. 2, No. 11, 2007.

[16] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[17] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent, "Use of a cDNA microarray to analyze gene expression patterns in human cancer, " Nature Genetics, Vol. 14, 1996, pp. 457-460.

[18] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, " in Proceedings of National Academy of Science, Vol. 96, 1999, pp. 6745-6750.

[19] A. Ben-Dor and Z. Yakhini, "Clustering gene expression patterns, " Journal of Computational Biology, Vol. 6, 1999, pp. 281-297.

[20] M. B. Eissen, P. T. Spellman, P. O. Brown, and D. Botstein, "Clustering analysis and display of genome wide expression patterns, " in Proceedings of the National Academy of Sciences, Vol. 95, 1998, pp. 14863-14868.

[21] D Jiang, C Tang, A Zhang, " Cluster analysis for gene expression data: a survey", Knowledge and Data Engineering, IEEE Transactions on, Vol. 16, No. 11, 2004, pp. 1370-1386.

[22] G. P. Shapiro, T. Khabaza and S. Ramaswamy, Capturing best practice for microarray gene expression data analysis", SIGKDD '03, August 24-27, 2003.

[23] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[24] S. Z. Selim, M.A. Ismail, K-means type algorithms: a generalized convergence theorem and characterization of local optimality, IEEE Trans. Pattern Anal. Mach. Intell. 6, 1984 pp. 81-87.

[25] H. Spath, Cluster Analysis Algorithms, Ellis Horwood, Chichester, UK, 1989.

[26] E. Forgy, "Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications, " *Biometrics*, vol. 21, 1965, pp. 768–780.

[27] J. MacQueen, "Some methods for classification and analysis of multivariate observations, " in *Proc. 5th Berkeley Symp.*, vol. 1, 1967, pp. 281–297.

[28] K. Stoffel and A. Belkoniene, "Parallel K-means clustering for large data sets, " in *Proc. EuroPar'99 Parallel Processing*, 1999, pp. 1451–1454.

[29] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*: Wiley, 1990.

[30] J. Peña, J. Lozano, and P. Larrañaga, "An empirical comparison of four initialization methods for the K-means algorithm, " *Pattern Recognit Lett.*, vol. 20, 1999, pp. 1027–1040.

[31] A. Likas, N. Vlassis, and J. Verbeek, "The global K-means clustering algorithm, " *Pattern Recognit.*, vol. 36, no. 2, 2003, pp. 451–461.

[32] G. Ball and D. Hall, "A clustering technique for summarizing multivariate data, " *Behav. Sci.*, vol. 12, 1967, pp. 153–155.

[33] P. Bradley and U. Fayyad, "Refining initial points for K-means clustering, " in *Proc. 15th Int. Conf. Machine Learning*, 1998, pp. 91–99.

[34] K. Krishna and M. N. Murty, "Genetic K-Means Algorithm", IEEE Transaction On Systems, Man, And Cybernetics—Part B: CYBERNETICS, Vol. 29, No. 3, June 1999

[35] Yi. Lu. Shiyong and Lu. Farshad Fotouhi, "FGKA: A Fast Genetic K-means Clustering Algorithm", SAC'04 Nicosia, Cyprus., March 2004 ACM 1-58113-812-1/03/04.

[36] Yi. Lu. Shiyong, Farshad Fotouhil, Youping Deng, d. Susan, J. Brown, " an Incremental genetic K-means algorithm and its application in gene expression data analysis", BMC Bioinformatics 2004

[37] V. Estivill-Castro and J. Yang, "A fast and robust general purpose clustering algorithm, " in *Proc. 6th Pacific Rim Int. Conf. Artificial Intelligence (PRICAI'00)*, R. Mizoguchi and J. Slaney, Eds., Melbourne, Australia, 2000, pp. 208–218.

[38] S. Gupata, K. Rao, and V. Bhatnagar, "K-means clustering algorithm for categorical attributes, " in *Proc. 1st Int. Conf. Data Warehousing and Knowledge Discovery (DaWaK'99)*, Florence, Italy, 1999, pp. 203–208.

[39] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values, " *Data Mining Knowl. Discov.*, vol. 2, , 1998, pp. 283–304.

[40] P. Hansen and N. Mladenoviæ, "J-means: A new local search heuristic for minimum sum of squares clustering, " *Pattern Recognit.*, vol. 34, 2001, pp. 405–413.

[41] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu, "An efficient K-means clustering algorithm: Analysis and implementation, " *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, 2000, pp. 881–892.

[42] G. Patanè and M. Russo, "The enhanced-LBG algorithm, " Neural Netw., vol. 14, no. 9, , 2001, pp. 1219–1237.

[43] -----, "Fully automatic clustering system, " IEEE Trans. Neural Netw., vol. 13, no. 6, 2002, pp. 1285–1298.

[44] M. Su and C. Chou, "A modified version of the K-means algorithm with a distance based on cluster symmetry, " IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, 2001, pp. 674–680.

[45] K. Wagstaff, S. Rogers, and S. Schroedl, "Constrained K-means clustering with background knowledge, " in Proc. 8th Int. Conf. Machine Learning, 2001, pp. 577–584.

[46] Zhang, B., "Generalized k-harmonic means. dynamic weighting of data in unsupervised learning", In Proceedings of the 1st SIAM ICDM, Chicago, IL, 2001.

[47] Zulal Gungor and Alper Unler, "K-Harmonic means data clustering with tabu-search method", Elsevier Applied Mathematical Modelling vol. 32, 2008, pp. 1115–1125.

[48] Greg Hamerly, Charles Elkan, "Learning the *k* in *k*-means. In proceedings of the seventeenth annual conference on neural information processing systems (NIPS), December 2003, pp. 281-288.

[49] D. Fogel, "An introduction to simulated evolutionary optimization, " IEEE Trans. Neural Netw., vol. 5, no. 1, pp. 3–14, Jan. 1994.

[50] S. Kirkpatrick, C. Gelatt, and M. Vecchi, "Optimization by simulated annealing, " Science, vol. 220, no. 4598, pp. 671–680, 1983.

[51] S.K. Pal, D. Bhandari, Selection of optimal set of weights in a layered network using genetic algorithms, Inform. Sci. 80 1994, pp. 213-234.

[52] S.K. Pal, D. Bhandari, M.K. Kundu, Genetic algorithms for optimal image enhancement, Pattern Recognition Lett. 15 1994, pp. 261-271.

[53] D. Whitley, T. Starkweather, C. Bogart, Genetic algorithms and neural networks: optimizing connections and connectivity, Parallel Comput. 14, 1990, pp. 347-361.

[54] R.K. Belew, J.B. Booker (Eds.), Proceedings of the Fourth International Conference on Genetic Algorithms, Morgan Kaufmann, San Mateo, 1991.

[55] S. Forrest (Ed.), Proceedings of the Fifth International Conference Genetic Algorithms, Morgan Kaufmann, San Mateo, 1993

[56] L.J. Eshelman (Ed.), Proceedings of the Sixth International Conference Genetic Algorithms, Morgan Kaufmann, San Mateo,

[57] K. Al-Sultan, "A Tabu search approach to the clustering problem, " Pattern Recognit., vol. 28, no. 9, 1995, pp. 1443–1451.

[58] C. Sung and H. Jin, "A Tabu-search-based heuristic for clustering, " Pattern Recognit., vol. 33, 2000, pp. 849–858.

[59] M. Delgado, A. Skármeta, and H. Barberá, "A Tabu search approach to the fuzzy clustering problem, " in Proc. 6th IEEE Int. Conf. Fuzzy Systems, vol. 1, 1997, pp. 125–130.

[60] D. Brown and C. Huntley, "A practical application of simulated annealing to clustering, " Pattern Recognit., vol. 25, no. 4, 1992, pp. 401–412.

[61] S. Selim and K. Alsultan, "A simulated annealing algorithm for the clustering problems, " Pattern Recognit., vol. 24, no. 10, 1991, pp. 1003–1008.

[62] S. Chu and J. Roddick, "A clustering algorithm using the Tabu search approach with simulated annealing, " in Data Mining II—Proceedingsof Second International Conference on Data Mining Methods and Databases, N. Ebecken and C. Brebbia, Eds, Cambridge, U.K., 2000, pp. 515–523.

TABLE 1: COMPARISON OF RESULTS FOR ALL FIFTEEN DATASETS

Datasets	Primary source	Secondary source	Dimension	# of cluster	K-means			
					# Error	Error (%)	# correct	Accuracy (%)
IRIS	[1]	[2]	[150x4]	3	17	11.33	133	88.67
WBCD	[16]	[2]	[683x9]	2	254	37.19	429	62.81
Iyer data/Serum data	[4]	[3], [8], [14]	[517x12]	11	249	48.1624	268	51.84
Cho data (yeast data)	[5],[6]	[7], [8], [14]	[386x16]	5	151	39.1191	235	60.88
Lung Cancer	[12]	[9]	[197x581]	4	55	27.92	142	72.0812
Leukemia (Golub Experiment)	[23]	[8], [14]	[72x7129]	2	29	40.28	43	59.72
Breast data A	[15]	[9]	[98x1213]	3	27	27.55	71	72.44
Breast data B	[15]	[9]	[49x1024]	4	23	46.93	26	53.0612
Breast Multi data A	[15]	[9]	[103x5565]	4	21	20.39	82	79.61
Breast Multi data B	[15]	[9]	[32x5565]	4	15	48.88	17	53.125
DLBCL A	[10],[11]	[9]	[141x661]	3	66	46.8085	75	53.191
DLBCL B	[10],[11]	[9]	[180x661]	3	40	22.22	140	77.78
DLBCL C	[10],[11]	[9]	[58x1772]	4	28	48.28	30	51.7241
DLBCL D	[10],[11]	[9]	[129x3795]	4	74	57.36	55	42.64
St. Jude Leukemia data	[13]	[9]	[248x985]	6	37	14.91	211	85.08

Note: Colored index indicates the reference from where we have downloaded data

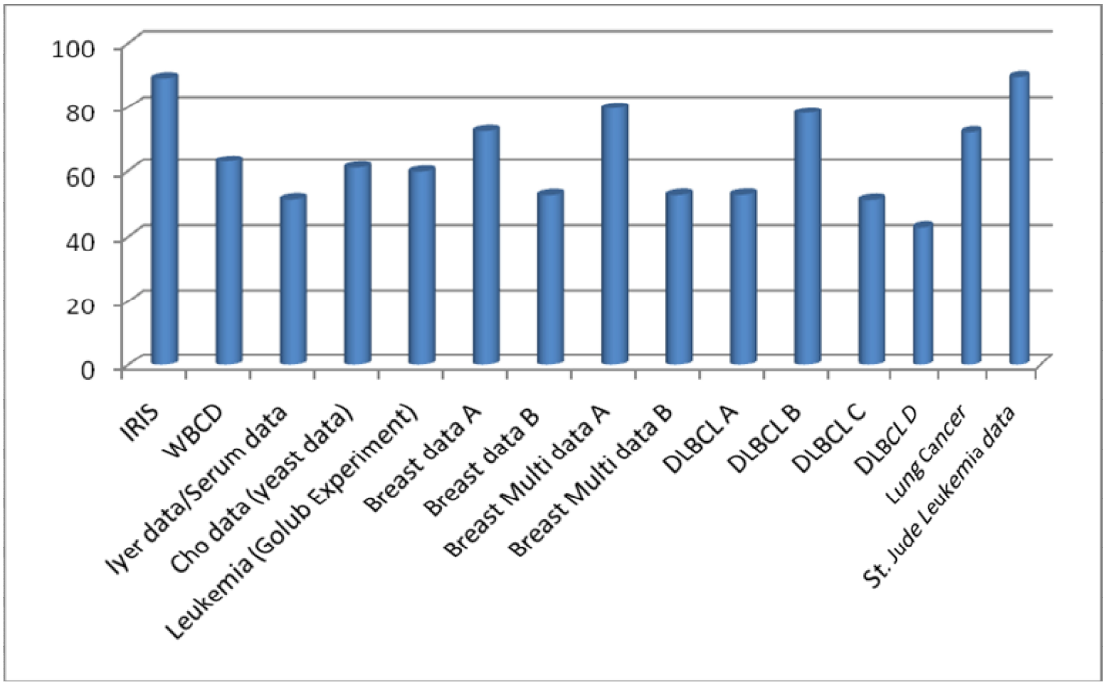

Fig. 1. Clustering accuracy for k-means for pattern recognition as well as microarray data, Horizontal axis represents different Datasets whereas vertical axis represents accuracy in percentage.

Table 2: Results for Iris Data

	Cluster 1	Cluster 2	Cluster 3	Total
The right number of data point	50	50	50	150
The number of data point wrongly clustered	0	4	13	17
The number of data point correctly clustered	50	46	37	133
Accuracy(%)	100	92	74	88.67

Table 3: Results for WBCD Data

	Cluster 1	Cluster 2	Total
The right number of data point	444	239	683
The number of data point wrongly clustered	142	112	254
The number of data point correctly clustered	302	127	37.19
Accuracy(%)	68.01	53.14	62.81

Table 4: Results for Cho Data

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
The right number of data point	67	135	75	54	55	386
The number of data point wrongly clustered	30	25	28	47	21	151
The number of data point correctly clustered	37	110	47	7	34	235

Accuracy(%)	55.22	81.48	62.67	12.96	61.82	60.88

Table5: Results for Leukemia Data (Golub Experiment)

	Cluster 1	Cluster 2	Total
The right number of data point	47	25	72
The number of data point wrongly clustered	15	14	29
The number of data point correctly clustered	32	11	43
Accuracy (%)	68.09	44	59.72

Table 6: Results for Serum data (Iyer data)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	Total
The right number of data point	33	100	145	34	43	7	34	14	63	19	25	517
The number of data point wrongly clustered	17	84	7	34	31	6	17	1	28	12	12	249
The number of data point correctly clustered	16	16	138	0	12	1	17	13	35	7	13	268
Accuracy(%)	48.48	16	95.17	0	27.91	14.29	50	92.86	55.56	36.84	52	51.84

Table 7: Results for Breast data A

	Cluster 1	Cluster 2	Cluster 3	Total
The right number of data point	11	51	36	98
The number of data point wrongly clustered	1	22	4	27
The number of data point correctly clustered	10	29	32	71
Accuracy (%)	90.91	56.86	88.89	72.44

Table 8: Results for Breast data B

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
The right number of data point	12	11	7	19	49
The number of data point wrongly clustered	0	5	5	13	23
The number of data point correctly clustered	12	6	2	6	26
Accuracy (%)	100	54.54	28.57	31.58	53.06

Table 9: Results for Breast Multi data A

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
The right number of data point	26	26	28	23	103
The number of data point wrongly clustered	2	1	18	0	21
The number of data point correctly clustered	24	25	10	23	82

Accuracy (%)	92.31	96.15	35.71	100	79.61

Table 10: Results for Breast Multi data B

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
The right number of data point	5	9	7	11	32
The number of data point wrongly clustered	3	2	3	7	15
The number of data point correctly clustered	2	7	4	4	17
Accuracy (%)	40	77.78	57.14	36.36	53.13

Table 11: Results for DLBCL A

	Cluster 1	Cluster 2	Cluster 3	Total
The right number of data point	49	50	42	141
The number of data point wrongly clustered	26	22	18	66
The number of data point correctly clustered	23	28	24	75
Accuracy (%)	46.94	56	57.14	53.19

Table 12: Results for DLBCL B

	Cluster 1	Cluster 2	Cluster 3	Total
The right number of data point	42	51	87	180
The number of data point wrongly clustered	18	7	15	40
The number of data point correctly clustered	24	44	72	140
Accuracy (%)	57.14	86.27	82.76	77.78

Table 13: Results for DLBCL C

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
The right number of data point	17	16	13	12	58
The number of data point wrongly clustered	1	9	12	6	28
The number of data point correctly clustered	16	7	1	6	30
Accuracy (%)	94.11	43.75	7.69	50	51.72

Table 14: Results for DLBCL D

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
The right number of data point	19	37	24	49	129
The number of data point wrongly clustered	13	28	13	20	74
The number of data point correctly clustered	6	9	11	29	55
Accuracy (%)	31.58	24.32	45.83	59.18	42.64

Table 15: Results for Lung Cancer

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
The right number of data point	139	17	21	20	197
The number of data point wrongly clustered	42	1	11	1	55
The number of data point correctly clustered	97	16	10	19	142
Accuracy (%)	69.78	94.11	47.62	95	72.08

Table 16: Results for St. Jude Leukemia data

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Total
The right number of data point	15	27	64	20	43	79	248
The number of data point wrongly clustered	15	0	3	4	14	1	37
The number of data point correctly clustered	0	27	61	16	29	78	211
Accuracy (%)	0	100	95.31	80	67.44	98.73	85.08