Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 2124 – 2128

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Modeling Optimization and Computing (ICMOC-2012)

A Naïve Soft Computing based Approach for Gene Expression Data Analysis

Vinita Debayani Mishra[a], Shruti Mishra[b*], Debahuti Mishra[c], Sandeep Kumar Satapathy[d]

[a] *NIIS Institute of Business Adminstration, Bhubaneswar, Odisha, India*
[b,c,d] *Institute of Technical Education and Research, Siksha O Anusandhan Deemed to be University, Bhubaneswar, Odisha, India*

Abstract

In a gene expression microarray data set, there could be tens or hundreds of dimensions, each of which corresponds to an experimental condition. But handling large value data set is a time consuming task, so we have introduced the fuzzy concept in order to discretize the dataset within the range 0 to 1. For our experiment, the gene expression data has been considered because of the common challenges. The first one is the curse of dimensionality. With increasing dimensionality, this data set becomes computationally intractable and therefore inapplicable in many real applications. Secondly, the specificity of similarities between points in a high dimensional data diminishes. It was proven that; for any point in a high dimensional data, the expected gap between the Euclidean distance to the closest neighbour and that to the farthest point shrinks as the dimensionality grows. In this paper, a key step in the analysis of gene expression data is the clustering using *k*-means and fuzzy *k*-means which group the genes that manifest similar expression patterns. From the experimental evaluation; it has been analyzed that fuzzy *k*-means shows better performance. The optimization techniques like Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) for clustering on gene expression data has been used because; many mathematical features that make them attractive for gene expression analysis including their flexibility in choosing a similarity function, when dealing with large data set. As compared to all the techniques it has been observed that the number of clusters is more and runtime of fuzzy *k*-means is less and it provides crisp boundary among the clusters; and the application of soft computing techniques shows better result than the traditional algorithms.

Keywords: *k*-means Clustering; Fuzzy *k*-means Clustering; Genetic Algorithm; Particle Swarm Optimization

1. Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help researchers and organisations focus on the most important information in their data warehouses. Given high dimensional data sets with hundreds of thousands or millions of entries, computing all pair wise similarities between objects is often intractable, and more efficient methods are called for. Clustering [1] large data sets are a ubiquitous task. Unsupervised

* Corresponding author. Tel.: +91-9776055678; fax: +91-674-2351880
E-mail address: shruti_m2129@yahoo.co.in

1877-7058 © 2012 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.06.255

clustering techniques have been applied to many important problems. By clustering patient records, health care trends are discovered. However, traditional clustering algorithms become computationally expensive when the data set to be clustered is large. There are three different ways in which the data set can be large: (1) there can be a large number of elements in the data set, (2) each element can have many features, and (3) there can be many clusters to discover. Handling the large dataset is a time consuming task so in order to reduce the time factor the dataset was discretized using fuzzy membership function within the range of 0 to 1. Fuzzy set theory proposed by Zadeh [2] was used to deal with cognitive uncertainty, including vagueness and ambiguity [3]. *k*- means [1] is a centroid-based popular cluster algorithm used in gene expression data analysis due to its high computational performance. *k*-means converge to a local optimum, and its result is subject to the initialization process, which randomly generates the initial clustering. A number of researchers have proposed GA [3] for clustering. PSO proposed by Eberhart and Kennedy [4] is a global optimization evolutionary algorithm that originates from the imitation of food-looking of birds. The objective of the PSO clustering algorithm is to discover the proper centroids of clusters for minimizing the intra-cluster distance as well as maximizing the distance between clusters. In this paper, an analysis have been done on the gene expression data set by using *k*-means and fuzzy *k*-means clustering. Also, some of the evolutionary algorithms like GA and PSO are implemented to find better result in terms of number of clusters. The layout of the paper is as follows: section 2 deals with related work based on clustering techniques and evolutionary algorithms. Section 3 gives the work plan model; section 4 describes the experimental evaluation and result analysis and finally section 5 gives the conclusion and future directions.

2. Related Work

Kerr *et al.* [5] proved that commonly used, agglomerative and partitive techniques are insufficiently powerful given the high dimensionality and nature of the data. Valarmathie *et al.* [6] proposed a clustering algorithm that deals with the data sets which are small in size. The efficiency can be measured by validating the resultant cluster using some clustering validation methods. Celis *et al.* [7] proposed powerful techniques that are available to analyze the global gene expression patterns of cultured cells and tissues obtained from normal and diseased subjects.

3. Proposed Model

Fig.1 depicts the proposed model and shows the generalized model relating to the fuzzification and pre-processing of the dataset, implementing clustering techniques to generate the number of clusters using some of the evolutionary algorithms (like GA and PSO) to get better results.

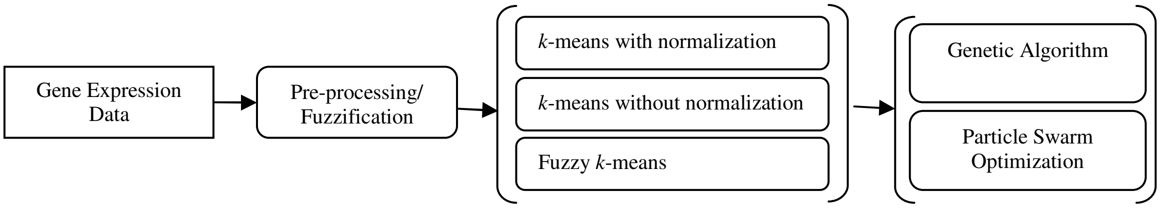

Gene Expression Data
Pre-processing/ Fuzzification
k-means with normalization
k-means without normalization
Fuzzy *k*-means
Genetic Algorithm
Particle Swarm Optimization

Fig. 1. Proposed Work flow Model

4. Experimental Evaluation and Result Analysis

Leukaemia data set [9] with 72 genes and 3859 samples has been selected for the experimental analysis. Some of the clustering techniques have been used along with certain evolutionary algorithms like GA and PSO. The experimental evaluation is divided into following steps:

Step I: Collection of leukemia data set [9].

Step II: Pre-processing/ fuzzification of data set

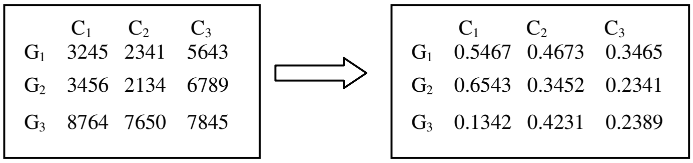
Data pre-processing is an essential procedure for handling the gene expression data matrix. One of the most commonly utilized normalization approaches is the *z*-score method shown in (1).

$$z = (x - mean)/sd \tag{1}$$

Where, *mean* is the mean of the sample (or population) and *sd* is the standard deviation of the sample or the population. *x* is the raw score. Pre-processing tools help to optimize the feature selection process, which leads to an increase in classification accuracy.

Fuzzification of the dataset: Here, the Gaussian membership was used to fuzzify the dataset as shown below in fig 2 (Gaussian membership is shown in (2)).

$$Y = e^{\frac{(-(x-mean))^2}{2*variance}} \tag{2}$$



	C_1	C_2	C_3
G_1	3245	2341	5643
G_2	3456	2134	6789
G_3	8764	7650	7845

	C_1	C_2	C_3
G_1	0.5467	0.4673	0.3465
G_2	0.6543	0.3452	0.2341
G_3	0.1342	0.4231	0.2389

Fig. 2. Example of a data matrix and the fuzzified matrix set of 3 genes and 3 conditions

Step III: *Clustering the data set using k- means algorithm*

The *k*-means clustering partitions data into *k* clusters such that objects in the same cluster are more similar to each other i.e., the clusters are internally similar, but externally dissimilar. The goal is to divide the objects into *k* clusters such that some metric relative to the centroids of the clusters is minimized. The algorithm first assigns each object to a cluster that has the closest centroid, and then sets initial positions for the cluster centroids, that is, when all objects have been assigned, recalculate the positions of the *k* centroids. This procedure is continued until the optimum assignment of objects to clusters is found.

Step IV: *Fuzzy k-means clustering*

The fuzzy data values are used for the fuzzy *k*- means clustering of all datasets. The degree of fuzziness differed markedly between the datasets, being low for datasets with clear clustering and small variations.

Step V: *Use of Genetic Algorithm (GA)*

Most of the clustering methods failed to find the correct number of clusters for all data sets. So in order to get crisp cluster we have implemented simple genetic algorithm for clustering and its performance was evaluated on real data sets and in comparison with other clustering algorithm described above. The result shows that none of the chosen algorithms is clearly superior to the others in terms of ability to identify classes of truly functionally related genes in the given data sets. However, GA seems to be competitive with all of the implemented algorithms and well suited for use in conjunction with the data driven internal validation measures.

Step-VI: *Use of Particle Swarm Optimization (PSO)*

PSO is a population-based stochastic optimization technique. In PSO, each single candidate solution is 'an individual bird of the flock', that is, a particle in the search space. Each particle makes use of its individual memory and knowledge gained by the swarm as a whole to find the best solution. PSO is based on the idea of collaborative behaviour and swarming in biological populations. The computation time used in PSO is less than in GAs. The parameters used in PSO are also fewer. However, if the proper parameter values are set, the results can easily be optimized. The results obtained after implementing the GA and PSO algorithm yields much better desired result as compared *k*-means without normalization, *k* means with normalization, fuzzy *k* means as shown in table 1 and fig 3.

Table. 1. Comparison based on the parameters mentioned

Algorithms	Run time (milliseconds)	Number of Clusters obtained
k-Means without Normalization	2934	76
k-Means with Normalization	2786	98
Fuzzy *k*-Means	2590	120
GA based Clustering	2279	133
PSO based Clustering	2025	148

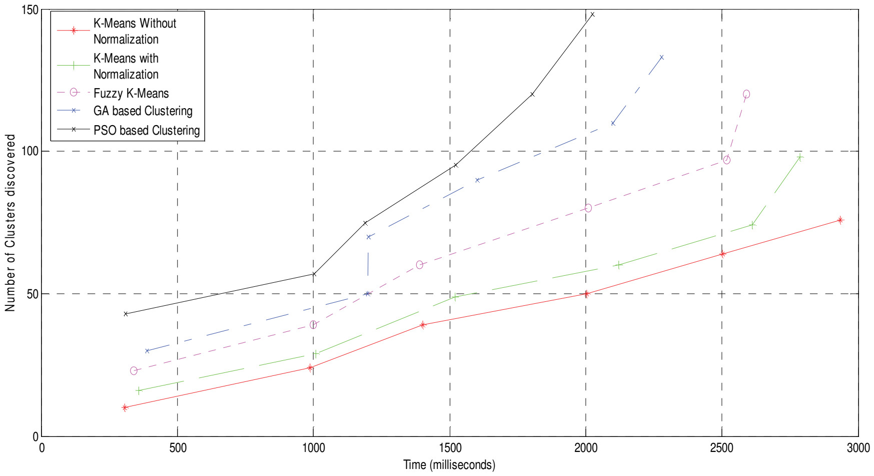


Fig.3. Performance comparison of all the clustering methods

5. Conclusion

The unique advantages of GA and PSO makes the technique a valuable tool for gene expression analysis, its flexibility can be used to reveal more complex correlations between gene expression patterns, promoting refined hypotheses of the role and regulation of gene expression changes. By analyzing the above algorithms at the end we conclude that the result obtained by the evolutionary algorithms provides good number of clusters.

References

[1] Berkhin P. Survey of clustering data mining techniques; *Accrue Software Research Paper*; 2002

[2] Zadeh LA. Fuzzy sets. *Information Control*; 1965: 8: 3: 338-353

[3] Yuan Y, Shaw MJ. Induction of fuzzy decision trees. *Fuzzy Sets and Systems*; 1995: 69:125-139

[4] Goldberg DE, Korb B and Deb K. Messy genetic algorithms: Motivation, analysis, and first results; *Complex Systems*; 1989: 5: 3: 493–530.

[5] Eberhart RC, Kennedy J. A new Optimizer using Particle Swarm theory. *Proc. of the Sixth Int. Symposium on Micro Machine and Human Science*; 1995: 39-43.

[6] Kerr G, Ruskin HJ, Crane M and Doolan P. Techniques for clustering gene expression data; Computers in Biology and Medicine; 2008: 38(3): 283-93.

[7] Shanthi I, Valarmathi ML. Comparison of Fuzzy C-Mean Clustering and k-Means Clustering for SAR Image Despeckling using Edge Detection. European Journal of Scientific Research; 2011: 62: 3: 426- 437.

[8] Celisa JE, Mogens K, Gromovaa I. *FEBS Letters;* 2000*:* 480: 2-16

[9] *http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43*