Gene expression

An improved algorithm for clustering gene expression data

Sanghamitra Bandyopadhyay[1], Anirban Mukhopadhyay[2,*] and Ujjwal Maulik[3]

[1]Machine Intelligence Unit, Indian Statistical Institute, Kolkata-700108, [2]Department of Computer Science & Engineering, University of Kalyani, Kalyani-741235 and [3]Department of Computer Science & Engineering, Jadavpur University, Kolkata 700032, India

Received on April 20, 2007; accepted on August 10, 2007

Advance Access publication August 25, 2007

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Recent advancements in microarray technology allows simultaneous monitoring of the expression levels of a large number of genes over different time points. Clustering is an important tool for analyzing such microarray data, typical properties of which are its inherent uncertainty, noise and imprecision. In this article, a two-stage clustering algorithm, which employs a recently proposed variable string length genetic scheme and a multiobjective genetic clustering algorithm, is proposed. It is based on the novel concept of points having significant membership to multiple classes. An iterated version of the well-known Fuzzy C-Means is also utilized for clustering.

Results: The significant superiority of the proposed two-stage clustering algorithm as compared to the average linkage method, Self Organizing Map (SOM) and a recently developed weighted Chinese restaurant-based clustering method (CRC), widely used methods for clustering gene expression data, is established on a variety of artificial and publicly available real life data sets. The biological relevance of the clustering solutions are also analyzed.

Contact: anirbanbuba@yahoo.com

Supplementary information: The processed and normalized data sets, supplementary figures, tables and other related materials are available at http://d.1asphost.com/anirbanmukhopadhyay/simmts.html

1 INTRODUCTION

The advent of microarray technologies have now made it possible to have a global and simultaneous view of the expression levels for many thousands of genes over different time points during different biological processes (Sharan *et al.*, 2003). Clustering is a primary approach to analyze such large amount of data. Clustering (Jain and Dubes, 1988; Maulik and Bandyopadhyay, 2002) is an unsupervised exploratory pattern classification technique which partitions the input space into K regions $\{C_1, C_2, \ldots, C_K\}$ on the basis of some similarity/dissimilarity metric, where the value of K may or may not be known a priori. In partitional clustering, the aim is to produce a $K \times n$ partition matrix $U(X)$ of the given data set X, consisting of n objects, $X = \{x_1, x_2, \ldots, x_n\}$. The partition matrix may be represented as $U = [u_{kj}]$, $k = 1, \ldots, K$ and $j = 1, \ldots, n$, where u_{kj} is the membership of pattern x_j to the kth cluster. For fuzzy clustering of the data, $0 < u_{kj} < 1$, i.e. u_{kj} denotes the probability of belongingness of pattern x_j to the kth cluster. Fuzzy C-Means (FCM) (Bezdek, 1981; Pakhira *et al.*, 2005) and its variants are widely used techniques used for microarray data clustering (Tomida *et al.*, 2002).

In general, it has been observed that the performance of clustering algorithms degrade with more and more overlaps among clusters in a data set. This is because in such situations several points have significant belongingness to more than one cluster, creating confusion regarding the overall cluster assignments. It may be beneficial if these points are first identified and excluded from consideration while clustering the data set. They could, thereafter, be assigned to one of the clusters using some similarity-based criterion. Motivated by these observation, a two-stage clustering algorithm is proposed in this article, which utilizes a novel concept of points having significant multi-class membership (SiMM). The proposed method is referred to as SiMM-TS (two-stage) clustering. The first stage deals with the identification of the number of clusters and the SiMM points and is based on the result of the application of a variable string length genetic algorithm (VGA)-based method (Maulik and Bandyopadhyay, 2003). In the second stage, a re-clustering of the data set is performed after excluding the SiMM points using a multiobjective genetic (MOGA) clustering (Bandyopadhyay *et al.*, 2007). These points are thereafter assigned to one of the clusters using the nearest neighbor criterion.

The effectiveness of using the VGA-MOGA combination in SiMM-TS is first established by experimenting with other clustering methods in the two stages for an artificial data. This includes an iterated FCM (IFCM) method, described later, in the first stage and methods like FCM (yielding combinations like VGA-FCM, IFCM-FCM) and simple genetic technique (Maulik and Bandyopadhyay, 2000) (yielding combinations like VGA-SGA, IFCM-SGA) in the second stage. Another combination tested is IFCM-MOGA (refer to Table 1). Thereafter, the superiority of SiMM-TS clustering method, as compared to two well-known methods for clustering gene expression data, namely average linkage (Jain and Dubes, 1988) and Self Organizing Map (SOM) (Tamayo *et al.*, 1999), and a recently proposed weighted Chinese restaurant clustering (CRC) scheme (Qin, 2006), is established for three real life gene

*To whom correspondence should be addressed.

expression data sets, namely, Yeast Sporulation, Human Fibroblasts Serum and Rat CNS data sets. Statistical tests have been carried out to establish that the proposed technique produces results that are statistically significant and do not come by chance. Biological interpretations have also been given for the clustering solutions.

2 THE PROPOSED SiMM-TS CLUSTERING TECHNIQUE

In this section, the proposed SiMM-TS clustering algorithm is described. First, the number of clusters and the corresponding fuzzy partition matrix are determined using a VGA-based method. Then, the technique for identifying the SiMM points has been discussed. Thereafter, a MOGA-based algorithm is used in the second stage of SiMM-TS algorithm that utilizes the knowledge of the identified SiMM points.

2.1 SiMM-TS clustering: the first stage

A variable string length GA (VGA)-based clustering technique has been proposed in Maulik and Bandyopadhyay (2003), where real valued encoding of cluster centers is used. The algorithm automatically evolves the number of clusters as well as the partitioning and minimizes the Xie–Beni (XB) (Xie and Beni, 1991) cluster validity index. Let $\{z_1, z_2, \ldots, z_K\}$ be the set of K cluster centers encoded in a chromosome. The XB index is defined as a function of the ratio of the total variation σ $(= \sum_{i=1}^{K} \sum_{k=1}^{n} u_{ik}^2 D^2(z_i, x_k))$ to the minimum separation sep $(= \min_{i \neq j}\{D^2(z_i, z_j)\})$ of the clusters, i.e.

$$\mathrm{XB} = \frac{\sigma}{n \times \mathrm{sep}} = \frac{\sum_{i=1}^{K}(\sum_{k=1}^{n} u_{ik}^2 \; D^2(z_i, x_k))}{n \times (\min_{i \neq j}\{D^2(z_i, z_j)\})}, \tag{1}$$

where n is the number of data objects, K represents the number of clusters, $U = [u_{kj}]$ is the fuzzy membership matrix (partition matrix). Here, $D(\;)$ is a distance function for finding the distance between two data points. In this article, Pearson Correlation based distance measure (Eisen *et al.*, 1998) is used to compute the distance between two gene vectors. Note that for compact and well-separated clusters, σ should be low while sep should be high, thereby yielding lower values of the XB index. The objective is therefore to minimize the XB index for achieving proper clustering. Since the number of clusters is considered to be variable, the string lengths of different chromosomes in the same population are allowed to vary (Maulik and Bandyopadhyay, 2003). Elitism is incorporated in order to keep track of the best chromosome obtained so far. This algorithm, utilized in the first stage of SiMM-TS clustering, has been run for a fixed number of generations. The best chromosome (in terms of XB index) of the last generation provides both the number of clusters and the corresponding fuzzy partition matrix $U(X, K)$.

2.2 Identification of SiMM points

The matrix $U(X, K)$ thus produced is used to find out the points which have significant multi-class membership (SiMM), i.e. the points which are situated at the overlapping regions of two or more clusters, and hence cannot be assigned to any cluster with a reasonable amount of certainty. Let us assume that a particular point $x_j \in X$ has the highest membership value in cluster C_q, and next highest membership value in cluster C_r, or, $u_{qj} \geq u_{rj} \geq u_{kj}$ where $k = 1, \ldots, K$, and $k \neq q$, and $k \neq r$. Suppose the difference in the membership values u_{qj} and u_{rj} is δ, i.e. $\delta = u_{qj} - u_{rj}$. Let $\mathcal{B}$ be the set of points lying on the overlapping regions of two or more clusters (SiMM points) and $Prob(x_j \in \mathcal{B})$ denotes the probability of x_j belonging to $\mathcal{B}$. Evidently, as δ increases, x_j can be assigned more confidently to cluster C_q and hence less confidently to $\mathcal{B}$. Therefore,

$$Prob(x_j \in \mathcal{B}) \propto \frac{1}{\delta}. \tag{2}$$

For each point in the data set X, the value δ is calculated. Now the data points are sorted in ascending order of their δ values. Hence, in the sorted list, the probability $Prob(x_i \in \mathcal{B})$ decreases as we move towards the tail of the list. A tuning parameter $\mathcal{P}$ is defined on the sorted list such that the first $\mathcal{P}\%$ points from the sorted list are chosen to be the SiMM points. The value of $\mathcal{P}\%$ should be chosen carefully so that the appropriate set of SiMM points is identified.

2.3 SiMM-TS clustering: the Second stage

In the second stage of the proposed technique, the SiMM points are excluded from the data set and the remaining points are reclustered into K clusters using a recently proposed MOGA-based method (Bandyopadhyay *et al.*, 2007). The MOGA-based algorithm uses a multiobjective GA, namely, NSGA-II (Deb *et al.*, 2002), for simultaneously optimizing two validity indices, namely XB index [Equation (1)] and J_m, where J_m is the minimizing criterion in the FCM algorithm (Bezdek, 1981) given by

$$J_m = \sum_{j=1}^{n} \sum_{k=1}^{K} u_{kj}^m D^2(z_k, x_j). \tag{3}$$

Note that the MOGA-based clustering algorithm simultaneously optimizes the compactness of the clusters (minimizing global cluster variance) and the separation of the clusters (maximizing the *sep* criterion of Equation (1). This method yields a set of Pareto-optimal solutions, from which one solution providing the best value of the Silhouette index (Rousseeuw, 1987), described in Section 4.1.2, is selected as the final output.

Finally, each SiMM point is assigned to one of the K clusters that evolve out of the MOGA clustering method, using the nearest centroid rule, i.e. it is assigned to the cluster whose center is at minimum distance from the point.

3 COMPARATIVE STUDY

In order to establish the effectiveness of the proposed SiMM-TS algorithm, its performance is compared with some other techniques that use other methods in the two stages. In the first stage, for automatically determining the number of clusters, an iterated version of the FCM is used.

FCM (Bezdek, 1981) is a widely used partitional clustering algorithm. The objective of FCM technique is to use the principles of fuzzy sets to evolve a partition matrix $U(X)$ while minimizing the measure given by Equation (3). It is known that FCM algorithm sometimes gets stuck at some suboptimal solution (Groll and Jakel, 2005). In the iterated FCM (IFCM), the FCM algorithm is run for different values of K starting from 2 to $\sqrt{n}$, n being the number of data points. For each K, it is executed 10 times from different initial configurations and the run giving the best J_m value is taken. Among these best solutions for different K values, the solution producing the minimum XB index [Equation (1)] value is chosen as the best partitioning. The corresponding K and the partitioning matrix are considered for further processing.

In the second stage, the number of clusters is already known (determined in the first stage). FCM or single objective GA (SGA) is used for clustering in the second stage in addition to the MOGA-based method described earlier.

The performance of the proposed SiMM-TS (using VGA-MOGA combination) is also compared with those of well-known gene expression data clustering methods, namely, Average Linkage, SOM and a recently developed technique called Chinese Restaurant clustering (CRC) (Qin, 2006), and also with VGA and IFCM applied independently. One artificial and three real life gene expression data sets are considered for experiments.

4 EXPERIMENTAL RESULTS

This section first provides a description of the performance metrics and the data sets used for experiments. Thereafter, experimental results are demonstrated both visually and numerically.

4.1 Performance metrics

For evaluating the performance of the clustering algorithms, adjusted Rand index (Yeung and Ruzzo, 2001) and Silhouette index (Rousseeuw, 1987) are used for artificial (where true clustering is known) and real life (where true clustering is unknown) gene expression data sets, respectively. Also, two cluster visualization tools namely Eisen plot and cluster profile plot have been utilized.

4.1.1 Adjusted Rand Index (Yeung and Ruzzo, 2001): Suppose T is the true clustering of a gene expression data set based on domain knowledge and C a clustering result given by some clustering algorithm. Let a, b, c and d respectively denote the number of gene pairs belonging to the same cluster in both T and C, the number of pairs belonging to the same cluster in T but to different clusters in C, the number of pairs belonging to different clusters in T but to the same cluster in C and the number of pairs belonging to different clusters in both T and C. The adjusted Rand index $ARI(T, C)$ is then defined as follows:

$$ARI(T,C) = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}. \tag{4}$$

The value of $ARI(T, C)$ lies between 0 and 1 and higher value indicates that C is more similar to T. Also, $ARI(T, T) = 1$.

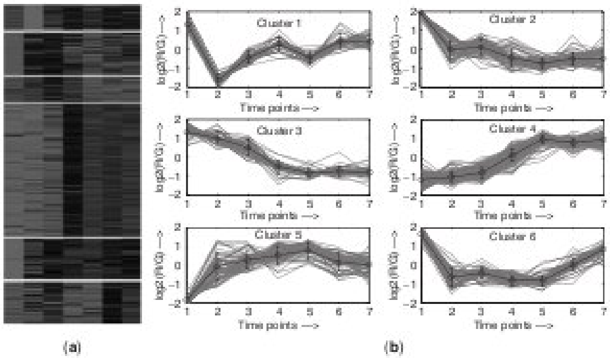

Fig. 1. Yeast Sporulation data clustered using the proposed SiMM-TS clustering method. (**a**) Eisen plot, (**b**) cluster profile plots.

4.1.2 Silhouette index (Rousseeuw, 1987): Silhouette index is a cluster validity index that is used to judge the quality of any clustering solution C. Suppose a represents the average distance of a point from the other points of the cluster to which the point is assigned, and b represents the minimum of the average distances of the point from the points of the other clusters. Now the silhouette width s of the point is defined as:

$$s = \frac{b-a}{\max\{a,b\}}. \tag{5}$$

Silhouette index $s(C)$ is the average Silhouette width of all the data points (genes) and it reflects the compactness and separation of clusters. The value of Silhouette index varies from -1 to 1 and higher value indicates better clustering result.

4.1.3 Eisen plot (Eisen *et al.*, 1998): In Eisen plot, (see Fig. 1a for example) the expression value of a gene at a specific time point is represented by coloring the corresponding cell of the data matrix with a color similar to the original color of its spot on the microarray. The shades of red color represent higher expression level, the shades of green color represent low expression level and the colors towards black represent absence of differential expression values. In our representation, the genes are ordered before plotting so that the genes that belong to the same cluster are placed one after another. The cluster boundaries are identified by white colored blank rows.

4.1.4 Cluster profile plot The cluster profile plot (see Fig. 1b for example) shows for each cluster the normalized gene expression values (light green) of the genes of that cluster with respect to the time points. Also, average expression values of the genes of the cluster over different time points are shown as a black line together with the SD within the cluster at each time point.

4.2 Input parameters

The GA-based clustering techniques (both for single objective and multiobjective) are executed for 100 generations with a fixed population size = 50. The crossover and mutation probabilities are chosen to be 0.8 and 0.01, respectively. The number of iterations for FCM algorithm is taken as 200 unless it converges before that. The fuzzy exponent m is chosen as in Kim *et al.*, 2006) and Dembele and Kastner, (2003), and the values of m for the data sets AD400_10_10, Sporulation, Serum and Rat CNS are obtained as 1.32, 1.34, 1.25 and 1.21,

Table 1. Comparison of SiMM-TS with other combination of algorithms in the two stages in terms of *ARI* and $s(C)$ for AD400_10_10 data

Algorithm	*K*	*ARI*	$s(C)$
SiMM-TS (VGA-MOGA)	10	0.6299	0.4523
IFCM	11	0.5054	0.2845
VGA	10	0.5274	0.3638
IFCM-FCM	11	0.5554	0.3692
IFCM-SGA	11	0.5835	0.3766
IFCM-MOGA	11	0.5854	0.3875
VGA-FCM	10	0.5894	0.4033
VGA-SGA	10	0.6045	0.4221
Average linkage	10	0.3877	0.2674
SOM	10	0.5147	0.3426
CRC	8	0.5803	0.3885

respectively. The parameter $\mathcal{P}$ for SiMM-TS clustering is chosen by experiment, i.e. the algorithm is run for different values of $\mathcal{P}$ ranging from 5% to 15%, and the best solution in terms of adjusted Rand index or Silhouette index is chosen. All the algorithms have been executed for 10 times and the average performance scores [the *ARI* values for artificial data where the true clustering is known and the $s(C)$ values for real life data where the true clustering is unknown] are reported.

4.3 Gene expression data sets

One artificial data set AD400_10_10 and three real life data sets namely, Yeast Sporulation, Human Fibroblasts Serum and Rat CNS data sets are considered for experiment.

4.3.1 AD400_10_10 This data set consists of expression levels of 400 genes across 10 time points. It is generated as in Yeung *et al.* (2001). The data set has 10 clusters, each containing 40 genes, representing ten different expression patterns.

4.3.2 Yeast Sporulation (Chu *et al.*, 1998) This data set consists of 6118 genes measured across 7 time points (0, 0.5, 2, 5, 7, 9 and 11.5 h) during the sporulation process of budding yeast. The data are then log-transformed. The Sporulation data set is publicly available at the website http://cmgm.stanford.edu/pbrown/sporulation. Among the 6118 genes, the genes whose expression levels did not change significantly during the harvesting have been ignored from further analysis. This is determined with a threshold level of 1.6 for the root mean squares of the log2-transformed ratios. The resulting set consists of 474 genes.

4.3.3 Human Fibroblasts Serum (Iyer *et al.*, 1999) This data set contains the expression levels of 8613 human genes. The data set has 13 dimensions corresponding to 12 time points (0, 0.25, 0.5, 1, 2, 4, 6, 8, 12, 16, 20 and 24 h) and 1 unsynchronized sample. A subset of 517 genes whose expression levels changed substantially across the time points have been chosen (Eisen *et al.*, 1998). This data set can be downloaded from http://www.sciencemag.org/feature/data/984559.shl.

4.3.4 Rat CNS (Wen *et al.*, 1998) The Rat CNS data set has been obtained by reverse transcription-coupled PCR to examine the expression levels of a set of 112 genes during rat central nervous system development over 9 time points. This data set is available at http://faculty.washington.edu/kayee/cluster.

Each data set is normalized so that each row has mean 0 and variance 1 (Z normalization) (Kim *et al.*, 2006).

4.4 Results for AD400_10_10 data

To establish the fact that the proposed SiMM-TS clustering algorithm (VGA followed by MOGA) performs better in terms of partitioning as well as determining the correct number of clusters, it was compared to the other combinations of algorithms in the two stages as discussed in Section 3 for AD400_10_10. The average ARI and s(C) values are reported in Table 1 (refer to the first eight rows) over 10 runs of each algorithm considered here. The table also contains the results produced by average linkage, SOM and CRC algorithms. The number of clusters found in a majority of the runs is reported. It is evident from the table that the VGA (and hence the algorithms that use VGA in the first stage) has correctly evaluated the number of clusters to be 10, whereas, IFCM and CRC wrongly found 11 and 8 clusters in the data set, respectively. Average linkage and SOM have been executed with 10 clusters. It also appears from the table that irrespective of any algorithm in the two stages, application of second stage clustering improves the *ARI* and $s(C)$ values, and the SiMM-TS (VGA-MOGA) technique provides the best of those values compared to any other combinations. Similar results have been obtained for other data sets also.

These results indicate that for this data set the proposed SiMM-TS clustering method with VGA and MOGA in the first and second stages, respectively, performs better than any other combination of algorithms in the two stages, and is also better than the other algorithms considered for comparison. This is because VGA is capable of evolving the number of clusters automatically due to its variable string length and associated operators. Also use of genetic algorithm instead of IFCM, in the first stage, enables the algorithm to come out of the local optima. Finally, the use of MOGA in the second stage allows the method to suitably balance different characteristics of clustering unlike single objective techniques, which concentrate only on one characteristic by totally ignoring the others. Henceforth, SiMM-TS will indicate the VGA-MOGA combination only.

4.5 Results for Yeast Sporulation data

Table 2 shows the Silhouette index values for algorithms SiMM-TS, IFCM, VGA, average linkage, SOM and CRC. It is evident from the table that VGA has determined the number of clusters as 6, IFCM finds it to be 7 whereas CRC algorithm gives $K=8$. Note that among these three methods, VGA provides the best $s(C)$ value. Thus for average linkage and SOM, K is chosen to be 6. From the $s(C)$ values, it can be noticed that the proposed SiMM-TS clustering performs the best, IFCM and average linkage perform poorly, whereas SOM and CRC perform reasonably well.

Table 2. $s(C)$ values for real life gene expression data sets

Algorithm	Sporulation		Serum		CNS	
	K	$s(C)$	K	$s(C)$	K	$s(C)$
SiMM-TS	6	0.6247	6	0.4289	6	0.5239
IFCM	7	0.4755	8	0.2995	5	0.4050
VGA	6	0.5703	6	0.3443	6	0.4486
Average linkage	6	0.5007	6	0.3092	6	0.3684
SOM	6	0.5845	6	0.3235	6	0.4122
CRC	8	0.5622	10	0.3174	4	0.4423

For the purpose of illustration, the Eisen plot and the cluster profile plots for the clustering solution found in one of the runs of SiMM-TS have been shown in Figure 1. The six clusters are very much clear from the Eisen plot (Fig. 1a). It is evident from the figure that the expression profiles of the genes of a cluster is similar to each other and they produce similar color patterns. Cluster profile plots (Fig. 1b) also demonstrate how the cluster profiles for the different groups of genes differ from each other, while the profiles within a group are reasonably similar. Refer to Figures 1–7 of the Supplementary Material for the Eisen plots and cluster profile plots of all the methods for this data set.

4.6 Results for human fibroblasts serum data

The $s(C)$ values obtained by the different clustering algorithms on Human Fibroblasts Serum data are shown in Table 2. Again, the VGA algorithm determines the number of clusters as 6, whereas IFCM and CRC found 8 and 10 clusters, respectively, with the VGA providing the best value of $s(C)$ among these three. The average linkage and SOM algorithms are therefore executed with 6 clusters. For this data set also, the proposed SiMM-TS clustering method provides much improved value of $s(C)$ compared to all the other algorithms including VGA, while IFCM and average linkage perform poorly. Figure 2 shows the Eisen plot and the cluster profile plots of the clustering solution obtained in one of the runs of SiMM-TS algorithm on this data.

4.7 Results for rat CNS data

Table 2 reports the $s(C)$ values for the clustering results obtained by the different algorithms on Rat CNS data. VGA gives the number of clusters as 6, which is the same as found in Wen *et al.* (1998). The IFCM and CRC found 5 and 4 clusters in the data set, respectively, but with poorer $s(C)$ values as compared to VGA. Thus for average linkage and SOM, $K=6$ is assumed. For this data set also, the proposed SiMM-TS method outperforms all the other algorithms in terms of $s(C)$. Figure 3 illustrates the Eisen plot and the cluster profile plots of the clustering solution obtained in one of the runs of the proposed SiMM-TS method on Rat CNS data.

The results indicate significant improvement in clustering performance using the proposed SiMM-TS clustering approach compared to the other algorithms. In all the cases considered here, VGA is found to outperform IFCM in determining both the appropriate number of clusters and the partitioning in terms of the ARI and $s(C)$ index. To justify that the solutions found by SiMM-TS is superior than any other algorithm applied separately, and that the results are statistically significant, statistical significance tests have been conducted and the results are reported in the next section.

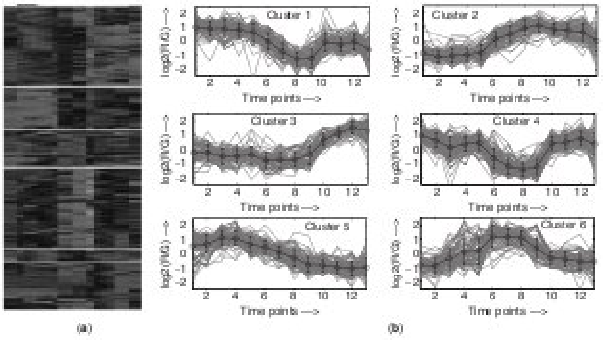

Fig. 2. Human Fibroblasts Serum data clustered using the proposed SiMM-TS clustering method. (**a**) Eisen plot, (**b**) cluster profile plots.

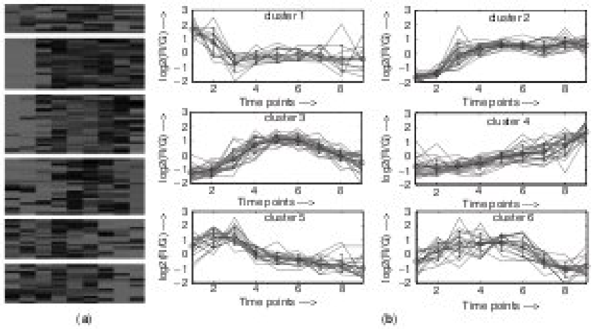

Fig. 3. Rat CNS data clustered using the proposed SiMM-TS clustering method. (**a**) Eisen plot, (**b**) cluster profile plots.

5 TEST FOR STATISTICAL SIGNIFICANCE

A non-parametric statistical significance test called Wilcoxon's rank sum test for independent samples (Hollander and Wolfe, 1999) has been conducted at the 5% significance level. Six groups, corresponding to six algorithms (1. SiMM-TS, 2. VGA, 3. IFCM, 4. average linkage, 5. SOM, 6. CRC), have been created for each data set. Each group consists of the performance scores [ARI for the artificial data and $s(C)$ for the real life data] produced by 10 consecutive runs of the corresponding algorithm. The median values of each group for all the data sets are shown in Table 3. Also the boxplots of the performance metric scores for the different groups are shown in Figures 19–22 of the Supplementary Material in the Supplementary Website.

It is evident from Table 3 that the median values for SiMM-TS are better than that for other algorithms. To establish that this goodness is statistically significant, Table 4 reports the P-values produced by Wilcoxon's rank sum test for comparison of two groups (group corresponding to SiMM-TS and a group corresponding to some other algorithm) at a time. As a null hypothesis, it is assumed that there are no significant difference between the median values of two groups. Whereas, the alternative hypothesis is that there is significant difference in the median values of the two groups. All the *P*-values reported in the table are less than 0.05 (5% significance level). For example, the rank sum test between algorithms SiMM-TS and IFCM for Sporulation data set provides a *P*-value of 2.2E-5, which is very small. This is strong evidence against the null hypothesis, indicating that the better median values of the performance metrics produced by SiMM-TS is statistically significant and has not occurred by chance. Similar results are obtained for all other data sets and for all other algorithms compared to SiMM-TS, establishing the significant superiority of the SiMM-TS algorithm.

6 BIOLOGICAL RELEVANCE

Biological interpretation of a cluster of genes can be best assessed by studying the functional annotation of the genes of that cluster. In this article, biological interpretations of the clusters are determined in terms of Gene Ontology (GO) (The Gene Ontology Consortium, 2000). For this purpose, a web-based Gene Ontology tool *FatiGO* (Al-Shahrour *et al*., 2004) (http://www.fatigo.org) has been utilized. FatiGO extracts the Gene Ontology terms for a query and a reference set of genes. In our experiment, a query is the set of genes of a cluster whose biological relevance is to be measured. The union of the genes from the other clusters is taken as the reference set. The GO level is fixed at 7.

Table 3. Median values of performance parameters [*ARI* for artificial and *s*(*C*) for real life data sets] over 10 consecutive runs of different algorithms

Algorithms	AD400_10_10	Sporulation	Serum	Rat CNS
SiMM-TS	0.6370	0.6353	0.4203	0.5147
IFCM	0.5011	0.4717	0.3002	0.4032
VGA	0.5204	0.5880	0.3498	0.4542
Average link	0.3877	0.5007	0.3092	0.3684
SOM	0.4891	0.5892	0.3345	0.4134
CRC	0.5162	0.5675	0.3227	0.4455

Table 4. *P*-values produced by Wilcoxon's rank sum test taking comparing SiMM-TS with other algorithms

Data sets	*P*-values				
	IFCM	VGA	Average link	SOM	CRC
AD400_10_10	2.1E−4	1.5E−3	5.2E−5	4.3E−5	1.1E−3
Sporulation	2.2E−5	2.1E−3	1.1E−5	1.2E−2	5.2E−3
Serum	1.5E−4	1.5E−4	5.4E−5	1.4E−4	1.4E−4
Rat CNS	1.1E−5	1.5E−4	5.4E−5	1.4E−4	1.7E−4

For illustration, similar clusters obtained using different algorithms for Yeast Sporulation data, have been analyzed. Cluster *selectivity* is assessed for validating the clusters. The cluster selectivity denotes the proportion of genes with a certain annotation (a biological process in our case) in the cluster relative to all genes in the data with this annotation. The selectivity is computed as the difference between percentage of genes with certain annotation in the query and reference set. A high selectivity thus indicates that these genes are well distinguished in terms of their expression profiles, among all the genes.

The performance of six algorithms, namely, SiMM-TS, VGA, IFCM, average linkage, SOM and CRC on Yeast Sporulation data are examined. From Figure 4, it is evident that cluster 1 obtained by SiMM-TS clustering corresponds to cluster 6 of IFCM, cluster 6 of VGA, cluster 4 of average linkage and cluster 3 of SOM and cluster 4 of CRC clustering solutions, because they provide similar expression patterns. A part of the FatiGO results (first six most selective annotations) of the clusters mentioned for the six algorithms applied on Sporulation data is available in the Supplementary Material. Here, biological processes have been used for annotation. The figure indicates that for most of the annotations, SiMM-TS algorithm provides better selectivity compared to the other algorithms. All the algorithms, except average linkage, shows maximum selectivity to the genes involved in rRNA processing. Average linkage shows maximum selectivity to M phase of Mitotic cell cycle. However, SiMM-TS shows the best maximum selectivity of 61.36% (62.5%−1.14%) compared to VGA (52.42%), IFCM (34.5%), average linkage (5.36%), SOM (34.5%) and CRC (56.55%). Similar results have been obtained for all other similar clusters produced by different algorithms. Readers can refer to Figures 8–18 and Table 1 of the Supplementary Material available in the afore-mentioned website for FatiGO results on other similar clusters of the Yeast Sporulation data.

For the purpose of illustration, Figure 5 shows the boxplots representing the maximum selectivity for the first six most selective annotations for the clusters produced by different algorithms for Yeast Sporulation data. It is evident from the figure that the range of selectivity values for SiMM-TS clustering is better compared to other algorithms. Overall, the results indicate the superiority of the proposed two-stage clustering method.

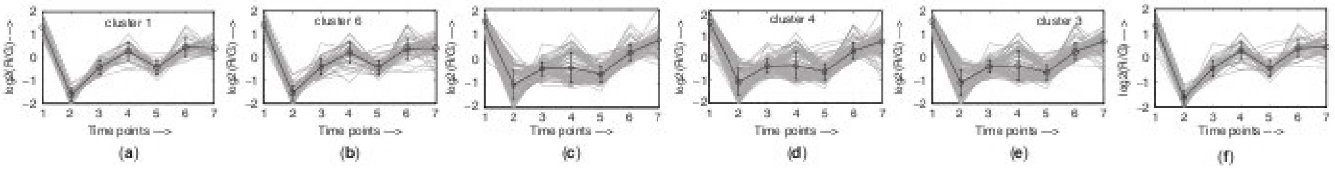

Fig. 4. Similar clusters found on Sporulation data by different algorithms: (**a**) Cluster 1 of SiMM-TS (39 genes), (**b**) Cluster 6 of VGA (44 genes), (**c**) Cluster 6 of IFCM (87 genes), (**d**) Cluster 4 of average linkage (92 genes), (**e**) Cluster 3 of SOM (89 genes) and (**f**) Cluster 4 of CRC (35 genes).

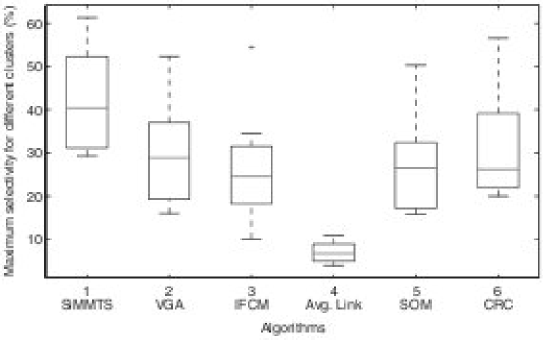

Fig. 5. Boxplots showing the range of best selectivity values for different algorithms on Yeast Sporulation data.

7 DISCUSSION AND CONCLUSIONS

This article proposes a two-stage clustering algorithm (SiMM-TS) for clustering gene expression data using the idea of points having significant membership to multiple classes (SiMM points). A VGA-based clustering scheme and a recently proposed MOGA-based clustering technique are utilized in the process. The number of clusters in a gene expression data set is automatically evolved in the proposed SiMM-TS clustering technique. The performance of the proposed clustering method has been compared with that of the other combinations of algorithms in the two stages as well as with the average linkage, SOM, CRC, VGA and IFCM clustering algorithms to show its effectiveness on an artificial and three real life gene expression data sets.

In general it is found that the SiMM-TS clustering scheme outperforms all the other clustering methods significantly. Moreover, it is seen that VGA performs reasonably well in determining the appropriate value of the number of clusters of the gene expression data sets. However, it may be noted that due to the complex nature of the gene expression data sets, it is difficult to identify a single partitioning that can be claimed to be the best. The clustering solutions are evaluated both quantitatively (i.e. using adjusted Rand index and Silhouette index) and using some gene expression visualization tools. Statistical tests have also been conducted in order to establish the statistical significance of the results produced by the proposed technique. Finally, biological interpretation of the clustering solutions have been given.

ACKNOWLEDGEMENTS

The authors are thankful to the anonymous reviewers for their valuable comments.

REFERENCES

Al-Shahrour,F. *et al.* (2004) FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Bandyopadhyay,S. *et al.* (2007) Multiobjective genetic clustering for pixel classification in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.*, **45**, 1506–1511.

Bezdek,J.C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York.

Chu,S. *et al.* (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

Deb,K. *et al.* (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, **6**, 182–197.

Dembele,D. and Kastner,P. (2003) Fuzzy c-means method for clustering microarray data. *Bioinformatics*, **19**, 973–980.

Eisen,M.B. *et al.* (1998) Cluster analysis and display og genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Groll,L. and Jakel,J. (2005) A new convergence proof of fuzzy c-means. *IEEE Trans. Fuzzy Syst.*, **13**, 717–720.

Hollander,M. and Wolfe,D.A. (1999) *Nonparametric Statistical Methods*. Wiely, USA, 2nd edition.

Iyer,V.R. *et al.* (1999) The transcriptional program in the response of the human fibroblasts to serum. *Science*, **283**, 83–87.

Jain,A.K. and Dubes,R.C. (1988) *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ.

Kim,S.Y. *et al.* (2006) Effect of data normalization on fuzzy clustering of DNA microarray data. *BMC Bioinformatics*, **7**, doi: 10.1186/1471-2105-7-134.

Maulik,U. and Bandyopadhyay,S. (2000) Genetic algorithm based clustering technique. *Pattern Recognit.*, **33**, 1455–1465.

Maulik,U. and Bandyopadhyay,S. (2002) Performance evaluation of some clustering algorithms and validity indices. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**, 1650–1654.

Maulik,U. and Bandyopadhyay,S. (2003) Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification. *IEEE Trans. Geosci. Remote Sens.*, **41**, 1075–1081.

Pakhira,M.K. *et al.* (2005) A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. *Fuzzy Sets Syst.*, **155**, 191–214.

Qin,Z.S. (2006) Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics*, **22**, 1988–1997.

Rousseeuw,P. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Sharan,R. *et al.* (2003) CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics*, **19**, 1787–1799.

Tamayo,P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, **96**, 2907–2912.

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Tomida,S. *et al.* (2002) Analysis of expression profile using fuzzy adaptive resonance theory. *Bioinformatics*, **18**, 1073–1083.

Wen,X. *et al.* (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA*, **95**, 334–339.

Xie,X.L. and Beni,G. (1991) A validity measure for fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intelli.*, **13**, 841–847.

Yeung,K.Y. and Ruzzo,W.L. (2001) An empirical study on principal component analysis for clustering gene expression data. *Bioinformatics*, **17**, 763–774.

Yeung,K.Y. *et al.* (2001) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309–318.