Hybrid Algorithm for Clustering of Microarray Data

Emir Buza[1], Zikrija Avdagic[1], Samir Omanovic[1] and Aida Hajdarpasic[2]
[1]Faculty of Electrical Engineering, Department for Computer Science and Informatics
[2]Faculty of Medicine, Center for Genetics
University of Sarajevo
{emir.buza, zikrija.avdagic, samir.omanovic}@etf.unsa.ba, aida.hajdarpasic@gmail.com

***Abstract*—Clustering is a crucial step in the analysis of gene expression data. Its goal is to identify the natural clusters and provide a reliable estimate of the number of distinct clusters in a given data set. In this paper we propose new hybrid algorithm for clustering of microarray data based on spectral clustering and k-means. Our algorithm consist of four steps, including preprocessing or filtering step, and finding optimal number of clusters by using two different clustering methods based on hierarchical and partition-based approaches. Then, we cluster data based on similarity/dissimilarity metrics with spectral clustering. In the final step, we select centroid genes based on k-means results. The proposed method was tested on six data sets from GEMS microarray database. When compared with existing single or combination of clustering methods, our results indicate about 10% improvement in selection of representative genes.**

Keywords Bioinformatics, Clustering Analysis, Gene expression, Data mining, Spectral clustering, Microarray data

I. Introduction

Microarrays have been used in various biomedical application such as gene discovery [1], [2], disease diagnosis [3], pharmacogenomics (drug discovery) [4], and toxicology [5]. DNA microarray technology is currently the most popular technology widely used for gene expression profiles. It is common to process thousands of genes from a large set of information in a microarray dataset in one experiment [6], [7]. Microarray data is usually represented as a matrix ($N \times M$), where each row represents a gene and each column an experiment. There are two approaches for detecting gene groups: *biclustering* and *clustering* [8]. Row-column clustering is performed simultaneously in biclustering approach, while in clustering approach either row (gene) or column (experiment) clustering is performed. In this paper we focus on clustering approach.

It is very difficult to make the right classification of data by using only pure statistical methods, especially in cases where the number of dimensions (number of samples) reaches up to several tens of thousands. The first problem is *prepocessing* or *filtering* less important information (information that do not contribute to good classification of genes/samples) from those that are crucial for further analysis. The goal is to transform data set into a improved, appropriate and reliable form for clustering. Thus, the quality of the clustering is increased. The second problem is *selection of representative genes* which have highly relevant information about disease from a large number of genes in a data set.

The general problem for multidimensional data sets whose deep structure is unknown is that the number and distribution of clusters is unknown in advance. Thus, the well-known clustering problem is how can one decide how many representative weight vectors should be used? When a number and distribution of the input clusters is known in advance, the selection of a representative can be found in a few simple computations.

In general, there are three main methods for microarray classification: *supervised*, *unsupervised* and *statistical* methods. The first category includes methods [9], [10], [11] for classification which require training and testing steps. However, in practice, microarray data sets usually contain a large number of genes and relatively small number of samples, so it is very difficult to make a good selection microarray data on subsets for training and testing. On the other hand, these methods usually suffer from a problem of overfitting, thus the characteristics of training set are memorized instead of capturing the desired pattern. Thus, additional knowledge about structure of data set and its classes is usually needed.

The second category includes clustering methods [12], [8], [13], [14] which do not require special knowledge for classification of data into a predefined number of groups. However, these methods usually suffer from a problem of initial number of clusters and starting centroid points, especially partitioning based clustering methods. Statistical methods [22], [24] are very popular and usually used as the first step of preprocessing [25] for all above methods. These kind of methods require a lot of time for execution, especially if microarray has a lot of dimensions. The second problem is that they need the help of other methods for completing final classification task. Thus, we need a more effective method and software tools for analyzing gene expression profiles of the microarray experimental samples [26].

In this paper we propose new hybrid algorithm for clustering of microarray data based on spectral clustering and k-means. Our hybrid algorithm is composed of four steps. In the first step, we pre-process data into a reduced number of data dimensions (genes) by using only statistical methods such as average value and standard deviation. In the second step, we determine optimal number of clusters by using two different clustering methods based on hierarchical and partition-based approaches. We use two different instances with two distance metrics. In the third step, we cluster data by similarity/dissimilarity metrics with spectral clustering, based on previously determined number of clusters. In the final step, we select centroid genes based on k-means results from the second step and results from spectral clustering.

The paper is organized as follows: background information on clustering is briefly reviewed in section II. In section III, we propose new hybrid algorithm for clustering of microarray

ISBN: 978-1-61804-233-0

data. Implementation and results are presented in section IV, and in section V we conclude the paper with some remarks.

II. Background

Data clustering is a method of grouping objects into meaningful categories. Clustering is used for discovery of similarity degree among forms, data exploration to discover underlying structures, compression for organizing data and other application for any scientific field that collects data. Supervised clustering methods use labels for easier cluster identification. However, labeling a large set of sample patterns can be costly, thus *unsupervised* methods are used. Some advantages of unsupervised methods are detection of the gradual change of pattern over time, identification of features useful for categorization, gaining insight into the nature or structure of the data during the early stages of an investigation.

In the context of expression profile for microarray data clustering has four main steps: 1) feature selection or extraction, 2) design of clustering algorithm, 3) validation of clusters and 4) interpretation of results. In recent years, there are many published studies in this field [8], [13], [14], [15], [16], [17]. In summary [6], these approaches can be grouped in five major groups based on: a) *similarity and distance measures*, b) *hierarchical clustering*, c) *partition clustering*, d) *model-based* and e) *feature-based methods*. The goal of these approaches is to group objects into disjoint clusters in such a way that objects with high similarity to each other belong to one cluster.

Methods based on similarity and distance measures vary in the way the distance between data is measured. The examples of distances often used are: Euclidean distance, correlation distance based on Pearson correlation coefficient, Mahalanobis distance, cosine angle and squared Pearson correlation. Hierarchical clustering is divided into agglomerative and divisive. Agglomerative algorithms begin with individual gene expression pattern cluster and successively merge them into smaller groups (bottom-up approach), while divisive approach start with whole set and divide it into smaller groups (top-down approach).

Partition clustering starts with pre-defined number of clusters with initial cluster centroids, allocate each data point to the cluster which has the nearest centroid, and compute the new centroids of the clusters. Then, alternate between second and third step until no data points change clusters [18]. Two classic representative algorithms are k-center and k-means. Model-based approaches are based on the fundamental relation between expression profile and a function of model parameters such as mixture-model and hidden Markov model. Feature-based methods focus on overall shapes of characteristic features, such as local shape-based similarity measure and scale-space signals. An exhaustive reviews of exiting approaches for the analysis of gene expression data can be found in [8], [14].

III. Hybrid algorithm

Our hybrid algorithm is composed of four steps as illustrated in Fig. 1: A. Data Set Prepocessing, B. Selection of the number of clusters, C. Spectral clustering and D. Selection of representative genes. In this section, we describe each step in detail.

A. Data Set Prepocessing

The first and preliminary step for classification and clustering methods is data cleaning and reduction of data dimensionality. Experimental studies have shown that microarray data sets contain bigger number of genes than the number of samples. Most of these genes are without appropriate expression values, thus they do not carry any important information for experiments.

In general, methods for dimension reduction can be categorized using different criteria like: normalization, discretization to binary values, simple statistical techniques (average value, standard variation, standard deviation, or combination of these), selection of the best original features and selection of a representative gene.

For our case we used statistical methods such as average value and standard deviation for cleaning (reduction) of mircoarray data as described in Eq. 1, Eq. 2 and Algorithm 1.

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N}(x_{ij}), \forall j = 1, 2, ..., M \tag{1}$$

$$s_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2}, \forall j = 1, 2, ..., M \tag{2}$$

Algorithm 1 Cleaning (reduction) microarray Data set

Require:
$data$ - microarray dataset;
$\bar{x}_j$ - average value by j column;
s_j - standard deviation by columns;

$k \leftarrow 0$ ▷ Set k to 0
for $j \leq$ size of number columns of $data$; $j \leftarrow j + 1$ **do**
  if $(s_j > \delta \cdot \bar{x}_j)$ **then**
    $k \leftarrow k + 1$ ▷ increment k when if statement is true
    for $i \leq$ size of number rows of $data$; $i \leftarrow i + 1$ **do**
      $g_{ik} \leftarrow x_{ij}$
    end for
  end if
end for
return (G) ▷ Return reduced Data set.

In Algoritam 1, $1 \leq \delta \leq 2$. For our case, δ is set to 1.75.

B. Selection of the number of clusters

For determining the optimal number of clusters, we used four clustering algorithms: two k-means and two Hierarchical Agglomerative Clustering (HAC), as illustrated in Fig. 1.B. In other words, we use two clustering methods with different similarity (disimilarity) measures. In our case, for the distance measure we selected Euclidean distance and the correlation distance based on Pearson correlation coefficient. After that the process of clustering of reduced microarray data set is repeated C_{max} times for each of four selected clustering algorithms. Upper boundary C_{max} in the worst case could be specified as

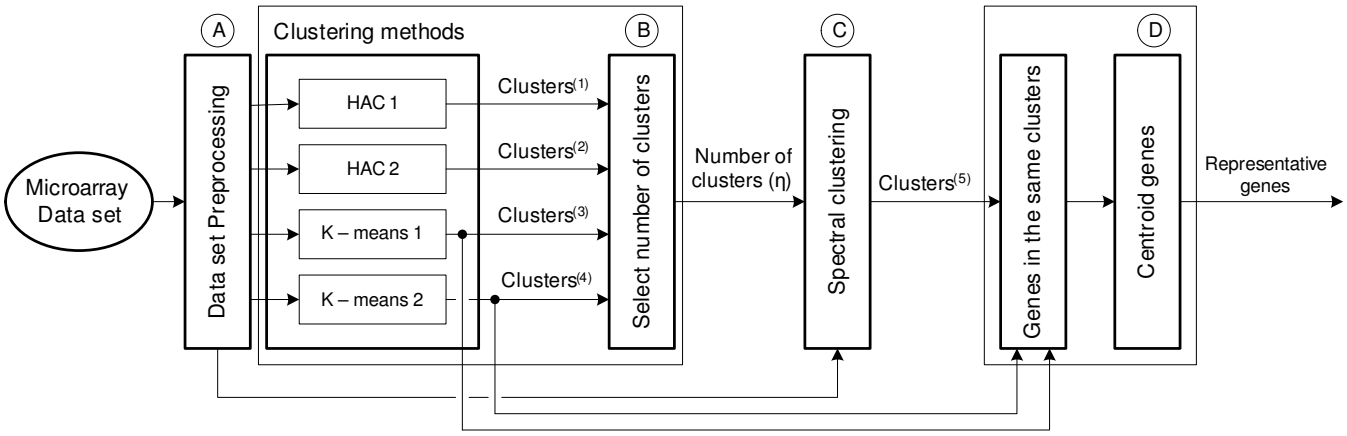


Fig. 1. Basic steps of Hybrid Algorithm for Clustering of Microarray data and selection of representative genes.

$N/2$, where N is number of experiments in the data set. In better case C_{max} should be selected so that it is sufficiently large than the optimal number of clusters (this upper boundary could be specified based on our knowledge of the data set).

During the process of execution of methods clustering, for each clustering algorithm l $(l = 1, 2, ..., 4)$ we calculated the sum of the variances of points within predefined number of clusters $k = 1, 2, ..., C_{max}$ based on Eq. 3.

$$Q_{lk} = \sum_{k=1}^{C_{max}} ((\sum_{x_i \in C_k} 1) - 1) \times \frac{1}{\sum_{x_i \in C_k} 1} \times \sum_{j=1, x_i \in C_k}^{M} (x_{ij} - \bar{x}_j)^2) \quad (3)$$

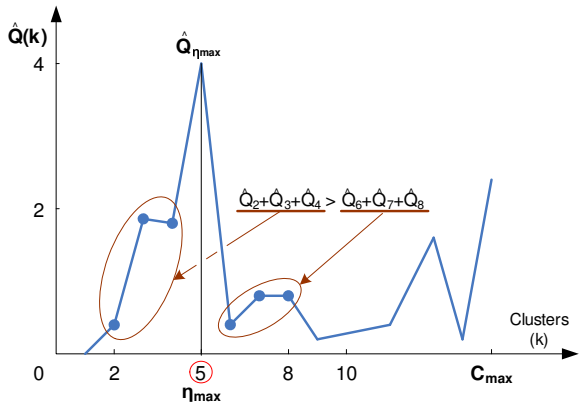
Suppose Q_{1k}, Q_{2k}, Q_{3k} and Q_{4k} are sums of variances for each of four clustering methods executed based on Eq. 3, where is $k = 1, 2, ..., C_{max}$, respectively. Finally, the optimal number of clusters η is calculated based on Eq. 4, Eq. 5, and Eq. 6.

$$\hat{Q}_k = \frac{1}{4} \sum_{l=1}^{4} (Q_{lk} - (\frac{1}{4} \sum_{j=1}^{4} Q_{jk}))^2, \forall k = 1, 2, ..., C_{max} \quad (4)$$

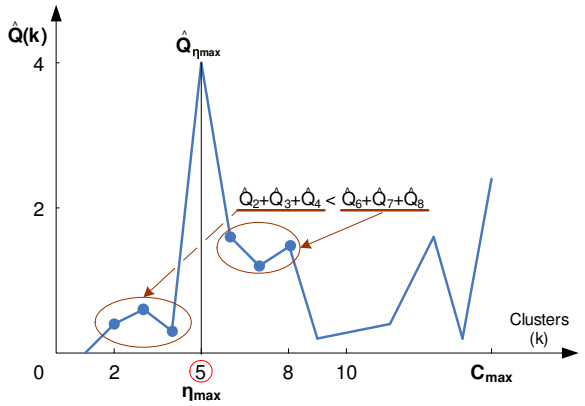
$$\eta_{max} = \arg \max_{2 \le k \le C_{max}} \{\hat{Q}_k\}, \quad (5)$$

$$\eta = \begin{cases} \eta_{max} - 1, & \text{if } \sum_{i=2}^{\eta_{max}-1} \hat{Q}_i > \sum_{\eta_{max}+1}^{2(\eta_{max}-1)} \hat{Q}_i \\ \eta_{max}, & \text{if } \sum_{i=2}^{\eta_{max}-1} \hat{Q}_i \le \sum_{\eta_{max}+1}^{2(\eta_{max}-1)} \hat{Q}_i \end{cases} \quad (6)$$

From Eq. 6, the maximal number of clusters C_{max} in the best case is selected as $C_{max} = 2 \times \eta_{max}$. A graphical representation of the process of finding the optimal number of clusters is shown in Fig. 2. Point η_{max} represents the value of k $(k = 1, 2, ..., C_{max})$, where the variance $\hat{Q}_k$ has the maximal value in point $\hat{Q}_{\eta_{max}}$.



(a) Left side is greater than the right side, i.e. the optimal number of clusters is calculated as $\eta = \eta_{max} - 1$.



(b) Right side is greater than the left side, i.e. the optimal number of clusters is calculated as $\eta = \eta_{max}$.

Fig. 2. Graphical representation of the process of finding the maximal and optimal number of clusters.

C. Spectral clustering

Recently, spectral clustering [19] became one of the most popular clustering algorithms. Traditional clustering algorithms, such as the k-means algorithm, use only simple metrics such as Euclidean distance to calculate distances between points in a data set and then make clusters by distance values (max or min). Unfortunately, they are not good enough for

clustering data into a predefined number of clusters.

In this paper we used the most commonly used normalized spectral clustering algorithm according to Ng et.al [20] and [21]. The basic technique of the spectral clustering is to perform dimensionality reduction before clustering in fewer dimensions. As input in the spectral clustering we used affinity matrix $A = a(_{ij}), i, j = 1, 2, .., n$ [19], [20], [23], which contains relative similarity for each pair of n points in the data set. This affinity matrix [23] is typically defined as $e^{\frac{-d^2}{\sigma^2}}$, in a way similar to the Gaussian kernel based on inter-point euclidean distance, where d is Euclidean distance between points and σ is a scale factor.

The basic steps of this algorithm are presented below:

- input: number k of clusters which need to construct
- for a given dataset of n points $X = x_1, ..., x_n \in R^l$ form the affinity matrix $A \in R^{n \times n}$ which is defined by $a_{ij} = e^{\frac{-d^2(x_i, x_j)}{\sigma^2}}$, $i, j = 1, 2, ..., n$, where $d(x_i, x_j)$ is some distance function (e.g. Euclidean) between points x_i and x_j
- compute degree matrix $D = diag(d_i)$ where $d_i = \sum_{j=1}^{n} a_{ij}$
- compute the normalized Laplacian matrix $L = D^{\frac{-1}{2}} \times L \times D^{\frac{1}{2}}$, where L is Laplacian matrix defined as $L = D - S$
- perform the eigen value decomposition by equation $L \times \upsilon = \lambda \times \upsilon$, where $\upsilon \in R^{n \times n}$ matrix of eigen vectors and $\lambda \in R^{n \times n}$ is matrix of eigen values
- form matrix $U \in R^{n \times k}$ from matrix υ, $u_{ij} = \upsilon_{im}$ where $i = 1, 2, ..., n; j = 1, 2, ..., k$ and m is k largest eigen vectors which are selected as the last k columns from matrix υ
- construct the normalized matrix Y from the obtained matrix U $y_{ij} = \frac{u_{ij}}{(\sum_{l=1}^{k} u_{il}^2)^{\frac{1}{2}}}$, where $i = 1, 2, ..., n; j, l = 1, 2, ..., k$
- clustering n points $y_i \in R^k, i = 1, 2, ..., n$ with K-means algorithm into $C_1, C_2, ..., C_k$ clusters

For our case, we found that the affinity matrix A (Eq. 7) gives the good results for microarray data set.

$$A = e^{-\sin(\frac{\arccos(R)}{2})^2} \tag{7}$$

where R is Pearson's correlation coefficient given by Eq. 8, x and y are points in d-dimensional space, $\bar{x}, \bar{y}$ are average values for x and y, and n is number of variables.

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{8}$$

D. Selection of representative genes

Selection of representative genes is executed by Algorithm 2. Selection is performed based on results of Spectral clustering and selective k-means results. Genes which exists in the same clusters (from results of Spectral and two k-means clustering) are selected as potential representative genes. In the second step, only genes with minimal Euclidean distance between potential selected genes ($g_k \in G$) and center of Spectral clusters are selected as the best representative genes ($\hat{X}$).

Algorithm 2 Selection of representative genes

Require:
X - microarray data set;
$C^{(e)}$ - k-means (Euclidean) clusters;
$C^{(c)}$ - k-means (Correlation) clusters;
$C^{(s)}$ - Spectral clustering clusters for number of clusters;

Relabel $C_k^1 \leftarrow C_k^{(e)}$ according to centers of clusters $C^{(s)}$ and $C^{(e)}$, where $C_k^{(e)} = \{C_1^{(e)}, C_2^{(e)}, ..., C_\eta^{(e)}\}$ for k = 1, 2, ... η
Relabel $C_k^2 \leftarrow C_k^{(c)}$ according to centers of clusters $C^{(s)}$ and $C^{(c)}$, where $C_k^{(c)} = \{C_1^{(c)}, C_2^{(c)}, ..., C_\eta^{(c)}\}$ for k = 1, 2, ... η
for i=1 to 2; $i \leftarrow i + 1$ **do**
  for k=1 to η; $k = k + 1$ **do**
    $G_k^i \leftarrow X_{C_k^i} \bigcap X_{C_k^{(s)}}$
  end for
end for
for k=1 to η; $k = k + 1$ **do**
  $G_k \leftarrow G_k^1 \bigcup G_k^2$
  $\hat{x}_k \leftarrow \min_{i, g_{ik} \in G_k} d^2(m_k^{(s)}, g_{ik})$, where $\{\hat{x}_k\} \in \hat{X}$
  ▷ d^2 is Euclidean square distance between $m_k^{(s)}$ (center of cluster $C_k^{(s)}$) and g_{ik}.
end for
return $(\hat{X})$ ▷ Return representative genes

IV. Implementation and Results

A. Implementation

Our method has been implemented in MATLAB version 7.11.0 (R2010b) with Statistics Toolbox. The clustering was performed on Intel Core2 1.80 GHz CPU with 4GB of RAM.

The effectiveness of our method has been verified on data sets selected from [27]. This approach allows convenient verification and comparison with similar algorithms or methods.

B. Results

In order to evaluate our hybrid algorithm we used several data sets. Data sets are used for evaluation our hybrid algorithm as described in Table I. The first column is data set identification, the second is number of clusters, the third is number of experiments, and the forth is number of genes in a data set.

Some results are presented in Fig. 3, 4, 5, where (a) depicts the sum of the variances of points within predefined number of clusters $k = 1, 2, ..., C_{max}$ (Eq. 3), and (b) depicts the optimal number of clusters η (Eq. 4).

We compare our results manually with other existing clustering models including the spectral clustering, k-means and hierarchical agglomerative clustering with different selected measures. We emphasize the fact that our main goal it to

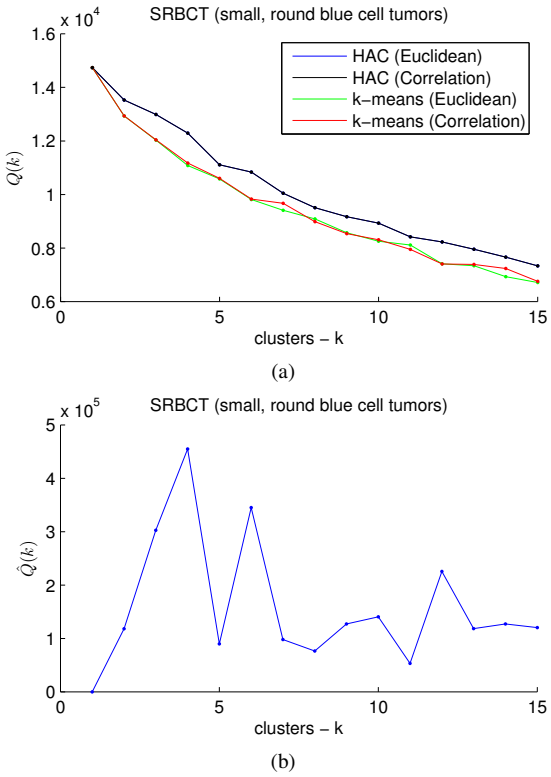


Fig. 3. Representation of Small, Round Blue Cell Tumors (SRBCT) Data set, (a) the sum of the variances of points within predefined number of clusters $k = 1, 2, ..., C_{max}$ (Eq. 3), (b) the optimal number of clusters $\eta = 4$ (Eq. 4).

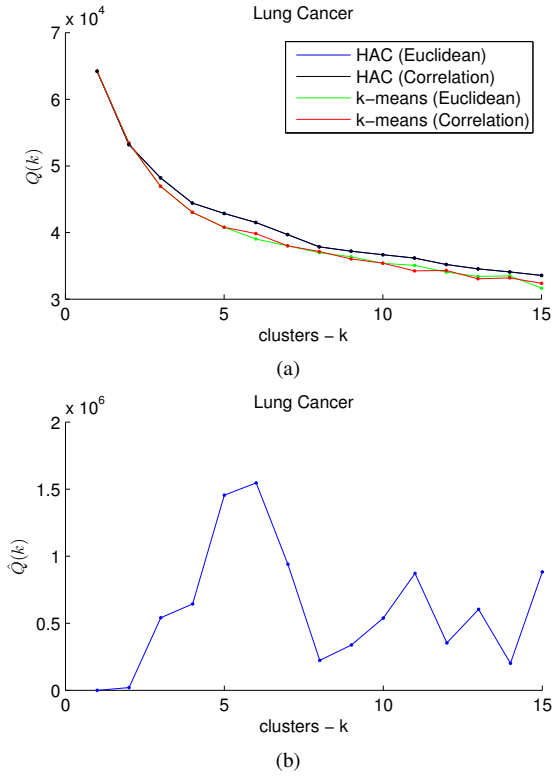


Fig. 4. Representation of Lung Cancer Data set, (a) the sum of the variances of points within predefined number of clusters $k = 1, 2, ..., C_{max}$ (Eq. 3), (b) the optimal number of clusters $\eta = 5$ (Eq. 4).

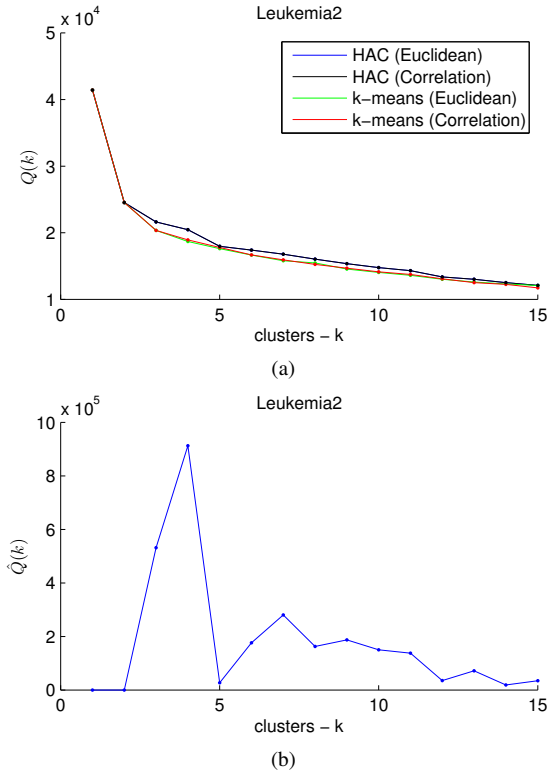


Fig. 5. Representation of Leukemia2 Data set, (a) the sum of the variances of points within predefined number of clusters $k = 1, 2, ..., C_{max}$ (Eq. 3), (b) the optimal number of clusters $\eta = 3$ (Eq. 4).

TABLE I. DATA SETS USED FOR EXPERIMENTATION

Data set name	Clusters	Experiments	Genes
Prostate Tumor	2	102	10509
SRBCT	4	83	2308
Lung Cancer	5	203	12600
DLBCL	2	77	5469
Brain Tumor2	4	50	10367
Leukemia2	3	72	11225

show the effectiveness of our hybrid algorithm in selection of representative genes [28], rather than application of single clustering method or their combination when using large microarray data sets [27].

The estimation accuracy when compared with existing single or combination of clustering methods is about 10% improved. We think that this information can be used for better prediction of unknown microarray genes.

V. CONCLUSION

In this paper we propose new hybrid algorithm for clustering of microarray data based on spectral clustering and k-means. Spectral clustering is a simple method for finding structure in data, which uses spectral properties of an associated pairwise similarity matrix.

The proposed method was tested on six Data sets from GEMS microarray data sets [27]. On a given set of data, our method improved selection of representative genes for about 10% when compared with existing single or combination of clustering methods. Thus, it is suitable for classification of unknown microarray genes.

ISBN: 978-1-61804-233-0

REFERENCES

[1] Nan, Xiaofei and Wang, Nan and Gong, Ping and Zhang, Chaoyang and Chen, Yixin and Wilkins, Dawn, *Biomarker Discovery Using 1-norm Regularization for Multiclass Earthworm Microarray Gene Expression Data*, Neurocomputing, Vol. 92, pp. 36–43, 2012

[2] Chu, Wei and Ghahramani, Zoubin and Falciani, Francesco and Wild, David L., *Biomarker Discovery in Microarray Gene Expression Data with Gaussian Processes*, Bioinformatics, Vol. 21(16), pp. 3385–3393, 2005

[3] Xin Zhao and Leo Wang-Kit Cheung, *Multiclass Kernel-Imbedded Gaussian Processes for Microarray Data Analysis*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, IEEE/ACM Transactions, Vol. 8(4), pp. 1545–5963, 2011

[4] Paul F. Predki, *Functional Protein Microarrays in Drug Discovery*, CRC Press, 2007

[5] Teresa Lettieri, *Recent Applications of DNA Microarray Technology to Toxicology and Ecotoxicology*, Environmental Health Perspectives, Vol. 114(1), pp. 4–9, 2006

[6] Amphun Chaiboonchoe and Sandhya Samarasinghe and Don Kulasiri, *Machine Learning for Childhood Acute Lymphoblastic Leukaemia Gene Expression Data Analysis: A Review*, Current Bioinformatics, Vol. 5(2), pp. 118–133, 2010

[7] Michael J. Korenberg, *Microarray data analysis: methods and applications*, Humana Press, 2007

[8] Jiang, Daxin and Tang, Chun and Zhang, Aidong, *Cluster analysis for gene expression data: A survey*, Knowledge and Data Engineering, IEEE Transactions, Vol. 16(11), pp. 1370–1386, 2004

[9] Perez M., Marwala T., *Microarray data feature selection using hybrid genetic algorithm simulated annealing*, Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention, Humana Press, pp. 1–5, 2012

[10] Greer, Braden, and Javed Khan, *Online analysis of microarray data using artificial neural networks*, Microarray Data Analysis, Humana Press, pp. 61–73, 2007

[11] Wutao Chen, Huijuan Lu, Mingyi Wang, Cheng Fang, *Gene Expression Data Classification Using Artificial Neural Network Ensembles Based on Samples Filtering*, Artificial Intelligence and Computational Intelligence, 2009. AICI '09. International Conference, Vol. 1, pp. 626–628, 2009

[12] Riccardo De Bin and Davide Risso, *A novel approach to the clustering of microarray data via nonparametric density estimation*, BMC Bioinformatics, pp. 1–8, 2011

[13] JXu, Rui and Wunsch, Donald, *Survey of clustering algorithms*, Neural Networks, IEEE Transactions, Vol. 16(3), pp. 645–678, 2005

[14] Madeira, Sara C and Oliveira, Arlindo L, *Biclustering algorithms for biological data analysis: a survey*, Computational Biology and Bioinformatics, IEEE/ACM Transactions, Vol. 1(1), pp. 24–45, 2004

[15] Shamir, Ron and Sharan, Roded, *Algorithmic approaches to clustering gene expression data*, Current Topics in Computational Biology, 2001

[16] Tibshirani, Robert and Hastie, Trevor and Eisen, Mike and Ross, Doug and Botstein, David and Brown, Pat, *Clustering methods for the analysis of DNA microarray data*, Dept. Statist., Stanford Univ., Stanford, CA, Tech. Rep, 1999

[17] Yin, Longde and Huang, Chun-Hsi *Clustering of Gene Expression Data: Performance and Similarity Analysis*, Computer and Computational Sciences, 2006. IMSCCS '06. First International Multi-Symposiums , Vol. 1, pp. 20–24, 2006

[18] Roger K. Blashfield and Mark S. Aldenderfer *Computer Programs for Performing Iterative Partitioning Cluster Analysis*, APPLIED PSYCHOLOGICAL MEASUREMENT, Vol. 2, No. 4, pp. 533-541, 1978

[19] U. Luxburg, *A Tutorial on Spectral Clustering*, Statistics and Computing 17(4), pp. 395 - 416, 2007

[20] A. Y. Ng., M. I. Jordan, Y. Weiss, *On spectral clustering: analysis and an algorithm*, Advances in Neural Information Processing Systems 14, pp. 849-856, 2001

[21] Implementation of four key algorithms of Spectral Graph Clustering using eigen vectors, *http://www.mathworks.com/matlabcentral/fileexchange/ 26354-spectral-clustering-algorithms/content/Spectral%20Clustering/ Jordan_Weiss.m*

[22] Matlab Statistics Toolbox, *http://www.mathworks.com/help/stats/index.html*

[23] D. Xu; P. Zhao; W. Gui; Ch. Yang; Y. Xie, *Research on spectral clustering algorithms based on building different affinity matrix*, Control and Decision Conference, pp. 3160-3165, 2013

[24] Hautaniemi S., Lehmussola A., Yli-Harja O., *DNA microarray data preprocessing*, First International Symposium on Control, Communications and Signal Processing, pp. 751–754, 2004

[25] Stiglic G., Kocbek S., Kokol P. *Unsupervised variance based preprocessing of microarray data*, Computer-Based Medical Systems, CBMS 2009. 22nd IEEE International Symposium, pp. 1–4, 2009

[26] Emanuel Weitschek, Giovanni Felici and Paola Bertolazzi, *MALA: A Microarray Clustering and Classification Software*, Database and Expert Systems Applications (DEXA), 23rd International Workshop, pp. 201–205, 2012

[27] GEMS (Gene Expression Model Selector) data sets, *http://www.gems-system.org/*

[28] Hanczar, Blaise and Courtine, Mélanie and Benis, Arriel and Hennegar, Corneliu and Clément, Karine and Zucker, Jean-Daniel., *Improving Classification of Microarray Data Using Prototype-based Feature Selection*, SIGKDD Explor. Newsl., Vol. 5(2), pp. 23–30, 2003

[29] Hong Chang; Dit-Yan Yeung, *Robust path-based spectral clustering with application to image segmentation*,Tenth IEEE International Conference on Computer Vision, vol. 1, pp.278-285, 2005