Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Clustering Microarray Data using Fuzzy Clustering with Viewpoints

Katerina N. Karayianni[1], George M. Spyrou[2] and Konstantina S. Nikita[1]

[1]School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
kkarayian@biosim.ntua.gr, knikita@ece.ntua.gr

[2]Biomedical Research Foundation
Academy of Athens
Athens, Greece
gspyrou@bioacademy.gr

***Abstract*—This paper studies the application of fuzzy clustering with viewpoints in order to cluster cell samples according to their gene expression profile. This method combines fuzzy clustering with external domain knowledge represented by the so-called viewpoints. The viewpoints that we employ are obtained from previously available expression data. The method was compared to the clustering algorithms of k-means, fuzzy c-means, affinity propagation, as well as a method of clustering microarray data that is based on prior biological knowledge, and has shown comparable/improved results over them.**

Keywords—Clustering; microarray data; prior knowledge; viewpoints

I. INTRODUCTION

Clustering gene expression data can be used to find genes with similar patterns of expression, in order to identify through this process the most representative genes to be further studied [1]. In addition, clustering can be applied to group cell samples of different conditions according to their expression profile. This explorative clustering process could be used to identify different subtypes of conditions, for example different cancer subtypes [2]. It could also be used to aid the labeling of unknown samples.

Performing clustering in microarrays is a challenging process due to the nature of the datasets used. One reason is the inherent noise in the data [3]. Especially for the case of clustering samples, data consist of rather a few numbers of samples, with each sample being described by a high-dimensional vector of gene measurements. This setup requires carefully designed methods in order to effectively perform the clustering process.

A concept that has been explored in data clustering is the incorporation of prior domain knowledge in the clustering process, resulting in methods that are semi-supervised in nature [4]. This approach has also been employed in the clustering of microarray data, for example in [5] and [6], in which previously obtained biologically-related knowledge was taken into account in order to improve the clustering process. In [6] the fuzzy c-means clustering algorithm is used in combination with Gene Ontology annotations, which enables to obtain clusters of functionally related genes, according to similarity in patterns of expressions, which is associated with common functional behavior. The approach of clustering samples from microarray experiments using a priori information has been followed in [7]. This method uses certain pre-defined classes of genes to guide the clustering process, which present significant relationship with the sample classes and that can be obtained for example from Gene Ontologies. Another method to cluster samples according to their microarray expression has been presented in [8]. The algorithm is based on finding groups of genes that are co-regulated in a way that is associated with the sample classes. The supervised clustering algorithm uses a measure of similarity among the gene attributes which is based on mutual information.

In this work, we employ a fuzzy clustering approach to cluster microarray data that uses prior domain knowledge. The method used is fuzzy clustering with viewpoints, developed by Pedrycz et al [9]. In this method, the researcher can impact directly on the cluster centers, using previously obtained knowledge. We employed this method in order to perform supervised clustering of cancer samples from various tissues, according to their microarray expression profiles. The purpose is to explore the different subtypes of conditions that could be assigned to these samples. Overall, the goal is to enable the construction of prediction models to aid in the identification of unlabeled samples, especially in multiclass problems. By making use of prior knowledge in the form of viewpoints, we can incorporate in the predictive processes the characteristics that we expect for the results to have.

To our best knowledge, this is the first application of fuzzy clustering with viewpoints to cluster microarray data samples.

II. DATASETS AND METHODS

A. Datasets

In order to test the performance of the selected algorithms we used publicly available microarray data. More specifically, we used a dataset used in the Van' t Veer et al. paper in [10], consisting of breast cells of two classes, one class from subjects without recurrence of cancer within 5 years of diagnosis and the other with recurrence. We also used three datasets from [2], which have been available by the authors in order to be used as benchmark data in similar studies. The first dataset (Armstrong dataset) includes samples from blood cells of three recorded

states (lymphoblastic leukemia with MLL translocations, conventional acute lymphoblastic leukemia and acute myelogenous leukemia). The Nutt dataset includes samples from cancer brain cells of four recorded conditions (glioblastomas, anaplastic oligodendrogliomas, both being classified as classic or non-classic). Lastly, the Pomeroy dataset includes brain cell samples of five recorded conditions (controversial medulloblastomas, malignant gliomas, atypical teratoid/rhabdoid tumors, normal tissues, primitive neuroectodermal tumors).

The information for the datasets used is summarized in Table I. The chip technology used to obtain the data was a customized Agilent oligochip for the Van't Veer dataset and single-channel Affymetrix chips for the others. In the table there is a description of the microarray chip used (Chip), the type of tissue that the samples came from (Tissue), the total number of samples used (#s), the number of the described classes of samples in each dataset (#cl), the number of samples from each type of class (Dist. Classes) and the recorded number of features in every sample (#f).

TABLE I.

Dataset	***Chip***	***Tissue***	***#s***	***#cl***	***Dist. Classes***	***#f***
Van't Veer	Agilent oligo	Breast	97	2	51, 46	4348
Armstrong	Affy	Blood	72	3	4,20,28	2194
Nutt	Affy	Brain	50	4	14,7,14,15	1377
Pomeroy	Affy	Brain	42	5	10,10,10,4,8	1379

B. Fuzzy clustering with viewpoints algorithm

In the method of fuzzy clustering with viewpoints, the fuzzy clustering process is affected by the previously available knowledge (domain knowledge), which is expressed in the manner of the so-called viewpoints. The method is a variation of the fuzzy c-means (FCM) clustering algorithm, which allows the clustered data to belong to more than one cluster, according to the different values of the resulting partition matrix. The fuzzy partitioning is done through an iterative process that optimizes a particular objective function and in each step the values of the partition matrix and the resulting prototypes are being updated. The optimization stops when certain criteria for termination are met. In fuzzy clustering with viewpoints, the method incorporates the viewpoints introduced by the users. These viewpoints are considered to be representatives of the data and are used as members of the prototypes family during the clustering process. The resulting objective function of the clustering process incorporates the selected measure of distance of the data from the viewpoints. A detailed description of the algorithm can be found in [9]. In our implementation we used viewpoints of numerical nature and we implemented the equivalent formulas.

The formula for the objective function used in the fuzzy clustering with viewpoints algorithm is the following:

$$Q = \sum_{k=1}^{N}\sum_{i=1}^{c}\sum_{\substack{j=1 \\ i,j:b_{ij}=0}}^{n} u_{ik}^{f}(x_{kj}-v_{ij})^2 + \sum_{k=1}^{N}\sum_{i=1}^{c}\sum_{\substack{j=1 \\ i,j:b_{ij}=1}}^{n} u_{ik}^{f}(x_{kj}-f_{ij})^2 \quad (1)$$

where N is the number of samples to be clustered, c is the preselected number of clusters, n is the dimension of the description vector for each sample, u_{ik} is the value of the partition matrix that describes the possibility of sample k to participate to the cluster i, x_{kj} is the *jth* feature of sample k and v_{ij} is the *jth* feature of the prototype for cluster i. The b_{ij} variable is equal to 1 when the *jth* feature of the prototype for cluster i is determined by a viewpoint, otherwise its value is 0. The f_{ij} value describes the viewpoint regarding the *jth* feature of the prototype of cluster i. The second part of the sum is considered only in the case that we have a viewpoint for the prototype of cluster i in the *jth* dimension. Else, the first part of the sum is used.

C. Clustering methods used for comparison

In order to evaluate the performance of fuzzy clustering with viewpoints in the particular clustering task, we compared it with three other well-established clustering algorithms, the classic methods of k-means [11] and the fuzzy c-means clustering [12], and affinity propagation [13]. Affinity propagation considers all samples as possible exemplars in a simultaneous way and exchanges real-valued messages among the samples, in order to progressively improve the selection of exemplars and the relevant clustering. In all methods, the number of clusters to be obtained had been predefined, resulting into a supervised type of clustering, which is more appropriate for comparison with the fuzzy clustering with viewpoints algorithm, which uses some form of a priori knowledge. Among the methods that we chose, the fuzzy c-means is a soft-clustering method, while the others produce crisp clustering results. In order to be able to compare the results of soft and hard clustering methods, we obtained from the fuzzy methods the equivalent hard clusters from the resulting partition matrices. We did so by assigning a sample to the cluster for which it has the highest value in the partition matrix.

Lastly, we performed a comparison of the fuzzy clustering with viewpoints method with the clustering method CAPIU described in [7], which uses prior biological knowledge in the form of pre-defined classes of genes, in order to cluster the microarray samples. We applied fuzzy clustering with viewpoints to the same real datasets used in [7], namely the Chiaretti [14] and the Spira [15] microarray datasets. The performance of CAPIU was evaluated by calculating the adjusted Rand index, thus we used the same measure in order to compare the clustering performance.

D. Clustering performance measures

The evaluation of the clustering methods is done using three external measures of performance evaluation, which compare how close is the resulting clustering to the already known classes of the clustered samples. In addition, we used one internal measure of performance to evaluate the results based on the clustered data. Lastly, we calculated the prediction error that measures the percentage of the samples that are

mislabeled by the clustering process, by assuming that the members of a cluster are similar in terms of their labeling and that one label can be assigned to a cluster. The label that we assigned in each cluster is the label type that is in majority in the samples of that cluster. In the case of equal numbers of samples with different labels, we used a second best choice measure, which was to use the label of the sample that is closer to the cluster center.

Regarding the external measures of performance, we chose to use the adjusted Rand, the Jaccard and the Fowlkes-Mallows indices. These indices are used as measures to evaluate how well a method can recover the structure of the data, as this is obtained from the already known labeling of the samples. The adjusted Rand index is a measure of how well the actual samples' structure is recovered. The index takes values from -1 to 1, with the values close to 1 indicating the best recovery of the true partitions of the samples [2]. In addition, this is an index which is unbiased in terms of the algorithm and the number of the clusters used. The Jaccard index [16] is another statistical index that we used in order to judge the similarity among the resulting hard clustering of the clustering methods used and the benchmark labeling of the samples. The values of the index are in the range of [0,1] with higher values indicating higher level of similarity among the two partitions. In the Fowlkes-Mallows index, higher values show a greater similarity of the clustering with the known labeling of the samples. As for the internal measure of performance selected, this was the Dunn index [17]. Higher values of the Dunn index indicate better clustering in the sense that the clusters are well-separated and relatively compact (higher inter-cluster distance and smaller intra-cluster distance).

The implementation of the fuzzy clustering with viewpoints algorithm has been done in R. For the k-means, fuzzy c-means, affinity propagation methods, as well as for the selected measures of performance, we used the available methods offered in the R packages repository.

III. Results and Discussion

In this section we summarize the results obtained by applying the selected clustering methods to the chosen datasets. In the clustering performed, we predefined the number of the resulting clusters to be equal to the recorded number of classes in each dataset.

The viewpoints used in the fuzzy clustering with viewpoints method have been constructed by computing the average expression value for every feature (probe/gene) among the samples that have the particular label. In order to create the viewpoints we used a part of the samples from each dataset, with representatives from all the classes of samples and used the rest for the actual clustering process. In order to have comparable results, we used the same set of samples that was clustered with the fuzzy clustering with viewpoints method, to be clustered by the other three selected methods. In real case applications, it would be beneficial to use a sufficient number of samples to obtain good average values for the viewpoints. Microarray experiment repositories could be used in order to obtain the necessary samples for the viewpoints.

The clustering results for the four different datasets using the selected clustering methods and measures of performance are summarized in Tables II-V. It has to be noted that the k-means and fuzzy c-means algorithms produce non-deterministic results. For those methods the algorithms were executed multiple times in order to obtain consensus average results for the indices. The affinity propagation algorithm produces deterministic results, thus the results came from a single run. The fuzzy clustering with viewpoints method was performed once for each dataset, since the use of viewpoints that describe all the dimensions of the data space, as in our case, gives a deterministic clustering outcome, for given viewpoints.

TABLE II.

Clustering Method	**Van't Veer Dataset**				
	Adjusted Rand Index	***Jaccard Index***	***Fowlkes-Mallows Index***	***Dunn index***	***Prediction Error***
K-means	0	**0.51**	**0.71**	**0.94**	0.37
Fuzzy c-means	0.09	0.37	0.54	**0.94**	0.32
Affinity propagation	-0.06	0.35	0.52	**0.94**	0.47
Fuzzy clustering with viewpoints	**0.46**	0.4	0.57	0.69	**0.16**

TABLE III.

Clustering Method	Armstrong Dataset				
	Adjusted Rand Index	*Jaccard Index*	*Fowlkes-Mallows Index*	*Dunn index*	*Prediction Error*
K-means	0.53	0.54	0.70	0.98	0.20
Fuzzy c-means	0.46	0.48	0.66	0.98	0.26
Affinity propagation	0.57	0.56	0.71	**1.01**	0.18
Fuzzy clustering with viewpoints	**0.82**	**0.78**	**0.88**	0.93	**0.07**

TABLE IV.

Clustering Method	Nutt Dataset				
	Adjusted Rand Index	*Jaccard Index*	*Fowlkes-Mallows Index*	*Dunn index*	*Prediction Error*
K-means	0.32	0.34	0.51	0.55	0.40
Fuzzy c-means	0.3	0.32	0.49	0.55	0.43
Affinity propagation	**0.36**	**0.4**	**0.63**	0.55	0.43
Fuzzy clustering with viewpoints	0.35	0.35	0.52	**0.69**	**0.33**

TABLE V.

Clustering Method	Pomeroy Dataset				
	Adjusted Rand Index	*Jaccard Index*	*Fowlkes-Mallows Index*	*Dunn index*	*Prediction Error*
K-means	0.41	0.36	0.55	0.79	0.33
Fuzzy c-means	0.35	0.33	0.53	0.79	0.43
Affinity propagation	0.29	0.31	0.5	**0.82**	0.35
Fuzzy clustering with viewpoints	**0.46**	**0.4**	**0.57**	0.69	**0.26**

The results show a clear advantage of the fuzzy clustering with viewpoints method in terms of the prediction error in all four datasets. In relation to the adjusted Rand index, fuzzy clustering with viewpoints performed better than the other algorithms in three datasets and was slightly worse than the better affinity propagation algorithm only in the third dataset. The results of the adjusted Rand index are of additional importance, since they are corrected for chance. The results for the Jaccard and Fowlkes-Mallows indices show somewhat different patterns that the adjusted Rand index. We observe that the fuzzy with viewpoints algorithm performs better in two datasets and is the second best in the rest. In order to evaluate this result for the Fowkles-Mallows index, we must take into account that it is by definition proportional to the number of true positives identified by the clustering method.

Regarding the internal validation measure of how dense and well-separated are the resulting clusters, the results appear to be dataset dependent.

The following table, Table VI, shows the results obtained when using a variation of the initial viewpoints for each dataset. More specifically, we calculated for each feature the standard deviation of the expression values. Then, instead of using as a viewpoint the average expression values, we randomly chose for each feature one point from a uniform distribution, with values in the range of $[m - s/2, m + s/2]$, where m is the average and s is the standard deviation of the expression value of that feature. We also repeated this process for wider ranges, with the performance of the clustering to be decreasing, as expected. Within the selected range, better results of the fuzzy clustering with viewpoints method are still maintained in relation to the other methods. By this, we demonstrate that the selection of the viewpoints does have an impact to the performance of the clustering, yet there is some level of tolerance regarding to the suitability of the selected viewpoints.

TABLE VI.

Fuzzy clustering with variation in the viewpoints					
Dataset	***Adjusted Rand Index***	***Jaccard Index***	***Fowlkes-Mallows Index***	***Dunn index***	***Prediction Error***
Van't Veer	0.15	0.4	0.57	0.96	0.29
Armstrong	0.75	0.72	0.84	0.93	0.09
Nutt	0.28	0.3	0.46	0.76	0.4
Pomeroy	0.47	0.41	0.58	0.7	0.28

Lastly, we performed a comparison of the fuzzy clustering with viewpoints method with the CAPIU algorithm. We employed our method in the Chiaretti and Spira datasets as in [7] and used the same measure of performance, the adjusted (corrected) Rand index, for one hundred repetitions of the algorithm. In each run we constructed different viewpoints in a random manner, in order to minimize the bias from the selection of particular viewpoints. In order to construct the viewpoints we did stratified sampling in the samples using a certain number of samples from each class and kept the remaining samples to perform the actual clustering.

In [7] a boxplot of the values of the adjusted Rand index from one hundred repetitions of CAPIU with randomized gene class mappings is plotted, which is given in Fig. 1. It depicts the smallest non-outlier observation, the lower quartile (Q1), median (Q2), upper quartile (Q3) and largest non-outlier observation for the values of the adjusted Rand index, thus giving a measure of the spread and central tendency of its values. Fig. 2. shows the boxplots we plotted for the values of the adjusted Rand index when using fuzzy clustering with viewpoints in the same datasets. It can be observed that the results are comparable to the ones by the CAPIU method. In the case of the Chiaretti dataset, fuzzy clustering with viewpoints performs better, with all observations being above the value of 0.3, while this is the case only for the upper quartile in the CAPIU method. In the Spira dataset, the values for Q1 and Q3 are higher in CAPIU that in fuzzy clustering with viewpoints, yet in the latter the upper quartile of the observations spans in a wider range above 0.3 than in the CAPIU method. Judging the performance of fuzzy clustering with viewpoints in the two datasets, it can be observed that the performance of the method can be affected by the characteristics of the dataset used.

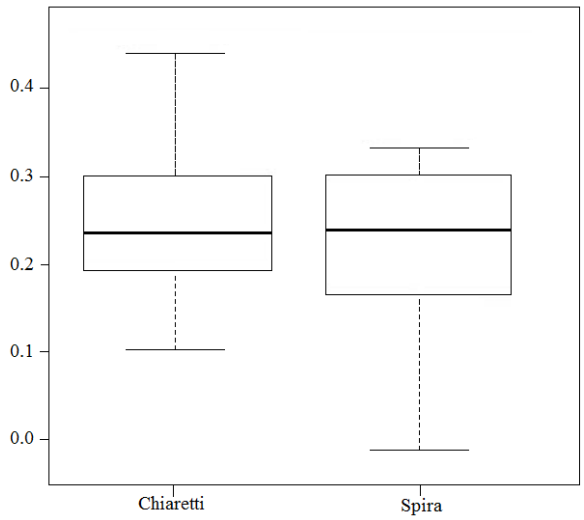

Figure 1. Boxplots of the adjusted Rand index with the CAPIU method, as taken from [7].

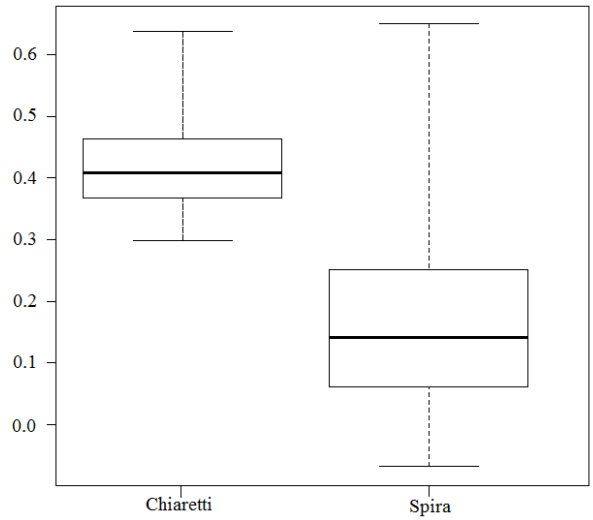

Figure 2. Boxplots of the adjusted Rand index with the fuzzy clustering with viewpoints method.

The results show a quite positive impact of using viewpoints to improve the clustering process. However, a better understanding of the differences amongst the selected measures of performance is necessary in order to explain the patterns that are observed in the results. In addition, comparing the performance of different clustering methods for the same dataset is a challenging process, in terms of the applicability of the selected measures of performance [18].

IV. CONCLUSIONS

In the current work, we explored the use of the method fuzzy clustering with viewpoints in order to cluster microarray data from tissue samples. The purpose of the clustering is to effectively label the condition of unknown samples, as well as to explore the possible subtypes of conditions that these samples might have. In order to guide the clustering process, we propose the use of previously available microarray expression data and introduce them as viewpoints in the clustering process. The performance of the method was judged in four datasets in comparison to other three well-established clustering methods and a clustering method that uses prior biological knowledge. The results from the performance validation measures that we applied show that the particular approach has advantages over the first three clustering algorithms and is comparable to the prior knowledge clustering method. Future work can include the comparison of the fuzzy clustering with viewpoints method with additional clustering techniques that explore the concept of guiding the clustering process with previously obtained biological knowledge. Future work can also explore other possible types of viewpoints to be used in the enhanced clustering process, as well as their incorporation in a platform that could serve as a viewpoint repository.

REFERENCES

[1] J. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," Science, vol. 302, no. 5643, pp. 249-255, August 2003.

[2] M. Souto, I. G. Costa, D. Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study," BMC Bioinformatics, vol. 9, no. 497, 2008.

[3] L. Klebanov and A. Yakovlev, "How high is the level of technical noise in microarray data?," Biol Direct., vol. 2, no. 9, 2007.

[4] S. Basu, Semi-supervised clustering: probabilistic models, algorithms and experiments. University of Texas at Austin, 2005.

[5] D. Komura, H. Nakamura, S. Tsutsumi, H. Aburatani, and S. Ihara, "Incorporating prior knowledge into clustering of gene expression profiles," 15th International Conference on Genome Informatics, 2004.

[6] L. Tari, C. Baral, and S. Kim, ''Fuzzy c-means clustering with prior biological knowledge,'' *J. Biomed. Inf.*, vol. 42, pp. 74–81, 2009.

[7] H. Redestig, D. Repsilber, F. Sohler, and J. Selbig, ''Integrating functional knowledge during sample clustering for microarray data using unsupervised decision trees,'' Biom J, vol. 49, no. 2, pp. 214-229, 2007.

[8] P. Maji, "Mutual information based supervised attribute clustering for microarray sample classification," IEEE Trans. Knowl. Data Eng., vol. 24, no. 1, pp.127 -140, 2012.

[9] W. Pedrycz, V. Loia, and S. Senatore, "Fuzzy clustering with viewpoints," Fuzzy Systems, IEEE Transactions on, vol. 18, no. 2, pp. 274-284, April 2010.

[10] L. J. van 't Veer, H. Dai, M J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," Nature, vol. 415, pp. 530-536, 2002.

[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, pp. 281–297, 1967.

[12] J. Bezdek, Pattern recognition with fuzzy objective function algorithms. New York: Plenum Press, 1981.

[13] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," Science, vol. 315, pp. 972–976, February 2007.

[14] S. Chiaretti, X. Li, R. Gentleman, A. Vitale, M. Vignetti, F. Mandelli, J. Ritz, and R. Foa, "Gene expression prole of adult T-cell acute lymphocytic leukemia identies distinct subsets of patients with different response to therapy and survival," Blood, vol. 103, pp. 2771–2778, 2004.

[15] A. Spira, J. Beane, V. Shah, G. Liu, F. Schembri, X. Yang, J. Palma, and J. S. Brody, "Effects of cigarette smoke on the human airway epithelial cell transcriptome," in Proceeedings of the National Academy of Sciences of the United States of America, vol. 101, pp. 10143–10148, 2004.

[16] R. Real and J. M. Vargas, "The probabilistic basis of Jaccard's index of similarity," Systematic Biology, vol. 45, no. 3, pp. 380-385, 1996.

[17] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," Journal of Cybernetics, vol. 3, no. 3, pp. 32–57, 1973.

[18] V. Estivill-Castro, "Why so many clustering algorithms: a position paper," SIGKDD Explor. Newsl., vol. 4, no. 1, pp. 65-75, June 2002.