Mathematical and Computational Applications, Vol. 16, No. 1, pp. 87-96, 2011.

PERFORMANCE OF AN ENSEMBLE CLUSTERING ALGORITHM ON BIOLOGICAL DATA SETS

Harun Pirim[1*], Dilip Gautam[2], Tanmay Bhowmik[2], Andy D. Perkins[2], Burak Ekşioglu[1], Ahmet Alkan[3]

[1] Industrial and Systems Engineering Department, Mississippi State University, 39762, USA

[2] Computer Science and Engineering Department, Mississippi State University, 39762, USA

[3] Sütçü İmam University, Electrical and Electronics Engineering, 46100, Turkey

harunpirim@gmail.com

Abstract– Ensemble clustering is a promising approach that combines the results of multiple clustering algorithms to obtain a consensus partition by merging different partitions based upon well-defined rules. In this study, we use an ensemble clustering approach for merging the results of five different clustering algorithms that are sometimes used in bioinformatics applications. The ensemble clustering result is tested on microarray data sets and compared with the results of the individual algorithms. An external cluster validation index, adjusted rand index (C-rand), and two internal cluster validation indices; silhouette, and modularity are used for comparison purposes.

Keywords- Ensemble Clustering, Rand Index, Silhouette Index, Modularity, Microarray Data Sets

1. INTRODUCTION

High throughput data technologies allow the production and analysis of biological data to address critical questions related to disease prediction and gene function, among others. Microarrays allow the measurement of expression levels of tens of thousands of genes simultaneously, in a single chip. Microarray measurements are eventually expressed as numbers indicating the relative expression values of each gene through an image processing process.

Gene co-expression networks constructed from the microarray data are very complex in terms of nodes and edges, since thousands of genes are represented by nodes and tens of thousands of relationships are represented by edges. A researcher is often interested in finding the effect of a treatment or time course in terms of changes in gene expression. This treatment or time course change leads researchers to focus on the genes that are significantly co-expressed under similar conditions. The classes to which genes belong are usually unknown, since most of the time there is little or no *a priori* information about the data, which requires analysis via an unsupervised learning technique. Clustering is an unsupervised learning technique that assigns objects into the same cluster based upon a cluster definition or criterion, which is the similarity between the objects being clustered.

Clustering has been studied for decades. However, there is no best clustering approach to be used for all applications. A particular clustering approach often has its own objective and assumptions about the data to cluster. Hence, combining multiple cluster-

ing approaches in an ensemble framework [1] may allow one to take advantage of the strengths of individual clustering approaches. In that sense, ensemble clustering is a promising approach to generating more accurate clusters than might be possible using an individual clustering approach.

In this study, we use an ensemble clustering approach as described in [2] for three different biological data sets. One of the data sets is the "Breast B" cancer diagnosis single channel microarray data set having 49 samples with 1213 attributes (corresponding to genes). The second is a protein data set that consists of 698 objects (corresponding to protein folds) with 125 attributes. The Breast B and protein data sets are detailed in and obtained from [3]. The third data set is a yeast cell cycle data set having 384 genes and 17 samples obtained from [4].

The base clustering algorithms used for the ensemble approach used here are hierarchical clustering (HC), K-means, dynamic tree cut (DTC), fuzzy C-means and a community structure finding algorithm (CSF). All of these algorithms except for fuzzy C-means were used and detailed in a previous study [5] to compare the performance of the individual algorithms with one another.

In order to evaluate the performance of the ensemble clustering approach, two internal and one external cluster validation indices are used. The internal validation indices are silhouette (S) [6] and modularity (Q) [7]. The external index is the adjusted rand index (C-rand) [8]. These indices are also described in section 3.

The remainder of the paper is organized as follows: section 2 gives background on ensemble clustering, section 3 concerns the application of the ensemble clustering on the three previously mentioned biological data sets and presents results, and section 4 concludes and indicates directions for possible future studies.

2. BACKGROUND

Combining the clustering results of many algorithms may result in high quality and robust clusters, since ensemble approaches such as bagging and boosting are used in classification problems and have proven to be effective [1]. The fact that the objects have various features (objects may be classified based on different features such as size, color, age, etc.) makes it difficult to find an optimal clustering of similar objects. In that sense, ensemble clustering is a promising heuristic.

The ensemble clustering ensemble framework is usually constructed as in Figure 1. Ensemble clustering can be difficult in the sense that it is a challenging task to select base clustering algorithms, define a consensus function and merge individual partitions generated by clustering algorithms via the chosen consensus function [9]. Asur et al. [10] proposed an ensemble clustering framework to extract biologically relevant functional modules in protein-protein interaction (PPI) networks. Their method attempts to handle the noisy false positive interactions and specific topological interactions present in the network. They used graph clustering algorithms, repeated bisections, direct k-way partitioning, and multilevel k-way partitioning, to obtain the base clusters, and introduced two topology based distance matrices. One of the distance matrices is based on the clustering coefficient [11], and the other one is a distance matrix based on the betweenness [7] measure. The authors used a soft ensemble method such that proteins

were able to be assigned to more than one cluster, and they conducted an empirical evaluation of the different ensemble methods to show the superior performance of their ensemble framework [10].

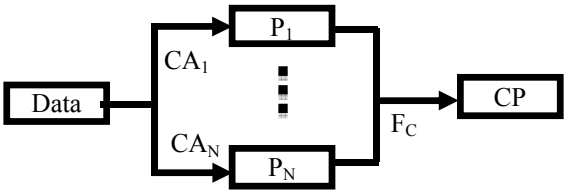

Figure 1. Ensemble clustering framework: CA refers to clustering algorithm, P refers to partition, F_C is the consensus function, and CP is the consensus partition.

There are a large number of fuzzy clustering algorithms with well-understood properties and benefits in various applications. Nevertheless, there has been very little analysis of using fuzzy clustering algorithms in regards to generating the base partitions in cluster ensembles. Wang [12] compared the use of hard and fuzzy C-means [13] algorithms in the well-known evidence-accumulation framework of cluster ensembles. In that study, it was observed that the fuzzy C-means approach requires much fewer base partitions for the cluster ensemble to converge, and is more tolerant of outliers in the data.

Avogadri and Valentili [14] proposed a fuzzy ensemble clustering approach to address the issue of unclear boundaries between the clusters from the biological and biomedical gene expression data analysis taking into account their inherent fuzziness. They had a goal of improving the accuracy and robustness of clustering results. After applying random projections to obtain lower dimensional gene expression data, they applied the fuzzy K-means algorithm on the low dimensional data to generate multiple fuzzy base clusters. Then, the fuzzy clusters were combined using a similarity matrix where the elements of the matrix were generated by the fuzzy t-norms algorithm, and finally, the fuzzy K-means algorithm was applied to the rows of the similarity matrix to obtain the consensus clustering. It was shown that the proposed ensemble approach is competitive with the other ensemble methods.

Microarray experiments often generate a great deal of data. If the data set is very large, it is possible to generate an ensemble of clustering solutions, or partition the data so that clustering may be performed on tractable-sized disjoint subsets [15]. The data can then be distributed at different sites, for which a distributed clustering solution with a final merging of partitions is a natural fit. Hore et al. [15] introduced two new approaches to combining partitions represented by sets of cluster centers. They stated that these approaches provided a final partition of data that was comparable to the best existing approaches and that the approaches could be 100,000 times faster while using much less memory. They compared the new algorithms against the best existing cluster ensemble approaches, clustering all of the data at once, and a clustering algorithm designed for very large data sets. Fuzzy and hard K-means based clustering algorithms were used for the comparison. It was shown that the centroid-based ensemble merging algorithms presented in the study generated partitions which were as good as the best

label vector method, or the method of clustering all the data at once. The proposed algorithms were also more efficient in terms of speed.

Asur et al. [9] applied an ensemble approach for clustering scale-free graphs. They used metrics based on the neighborhood metric (which uses the adjacency list of each node and considers the nodes as having several common neighbors), the clustering coefficient, and the shortest path betweenness of nodes in the network. The scale-free graph they used was from a budding yeast PPI network that contained 15147 interactions between 4741 proteins. It was reported that ensemble clustering can provide improvements in cluster quality for scale-free graphs based upon the preliminary results.

Galluccio et al. [16] proposed an ensemble clustering method called evidence accumulation clustering based on dual rooted prim tree cuts (EAC-DC). Their algorithm computes the co-association matrix based on a forward algorithm that repeatedly adds edges to Prim's minimum spanning tree (MST) to identify clusters until a satisfying criterion is met. A consensus cluster is then generated from the co-association matrix using spectral partitioning. Here, a MST is a fully connected sub-graph with no cycles and a dual-rooted tree is obtained by finding the union of two sub-trees. They applied their approach to the Iris data set [17], the Wisconsin breast cancer data set [18] (both obtained from [19]) and synthetic data sets, and presented a comparison of their results with other existing ensemble clustering methods.

Hu and Yoo [1] used a cluster ensemble in gene expression analysis. In their ensemble framework, the partitions generated by each individual clustering algorithm are converted into a distance matrix. The distance matrices are then combined to construct a weighted graph. A graph partitioning approach is then used to generate the final set of clusters. It was reported that the ensemble approach yields better results than the best individual approach on both synthetic and yeast gene expression data sets.

Fred and Jain [2] combined multiple partitions using evidence accumulation. Each partition generated by a clustering algorithm was used as a new piece of knowledge, to help uncover the relationships between objects. We used their approach for our ensemble. The core idea behind the ensemble approach here is constructing the co-association matrix by employing a voting mechanism for the partitions generated using individual clustering algorithms. A co-association matrix C is constructed based upon the formulation below, where n_{ij} is the number of times the object pair (i,j) is assigned to the same cluster among the N different partitions:

$$C(i,j) = n_{ij} / N$$

After constructing the co-association matrix, Fred and Jain [2] used single linkage hierarchical clustering to obtain the new cluster tree (dendrogram) and then used a cut-off value corresponding to the maximum life time (difference between merge points where branching starts) on the tree. They also employed the same ensemble framework using K-means partitions with different parameters. They tested their algorithms on ten different data sets, comparing the results with other ensemble clustering methods. They reported that their ensemble approach could identify the clusters with arbitrary shapes and sizes, and performed better than the other combination methods.

3. APPLICATION TO BIOLOGICAL DATA

As mentioned in the introduction, we use the ensemble approach described in [2]. However, we select a different set of base clustering algorithms that may produce better results, as stated by Fred and Jain, since the application of evidence accumulation techniques to more powerful clustering methods can lead to better partitions [2]. The ensemble clustering algorithm steps are as follows:

1. For each partition generated by the base clustering algorithms, construct a binary partition matrix P(i, j), where $1 \leq i \leq n$ and $1 \leq k \leq r$; n = number of attributes; r = number of unique clusters
2. Initialize the non-diagonal elements of co-association matrix C(n, n) to zero and the diagonal elements to 1.
3. For k = 1 to r do
4. For i = 1 to n - 1 do
5. If (P(i, k) = 1) then
6. For j = i + 1 to n do
7. If P(j, k) = 1) then
8. Update the co-association matrix as:
   C(i, j) = C(i, j) + (1/N) where N = number of clusterings from clustering algorithms.
9. Obtain the distance matrix as D(i, j) = 1 – abs(C(i, j))
10. Use Hierarchical Clustering with Complete Linkage (HC-CL) to generate the dendrogram
11. Cut the tree at suitable point by visually inspecting the tree (tree height giving the maximum cluster lifetime)

For example, for two different partitions of a data set with six objects in each: (1, 1, 1, 2, 2, 2) and (1, 1, 2, 2, 2, 2), the co-association and distance matrices are given below:

$$\begin{pmatrix} 1 & 1 & 0.5 & 0 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 1 & 1 & 1 \\ 0 & 0 & 0.5 & 1 & 1 & 1 \\ 0 & 0 & 0.5 & 1 & 1 & 1 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 & 0 & 0.5 & 1 & 1 & 1 \\ 0 & 0 & 0.5 & 1 & 1 & 1 \\ 0.5 & 0.5 & 0 & 0.5 & 0.5 & 0.5 \\ 1 & 1 & 0.5 & 0 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0 & 0 \end{pmatrix}$$

Figure 2. Co-association and distance matrices.

Using the distance matrix as an input for HC-CL, the clusters obtained for this example are shown in Figure 3.

We use HC, K-means, C-means, DTC, and CSF to generate five different partitions. The workflow to generate the partitions is shown in Figure 4.

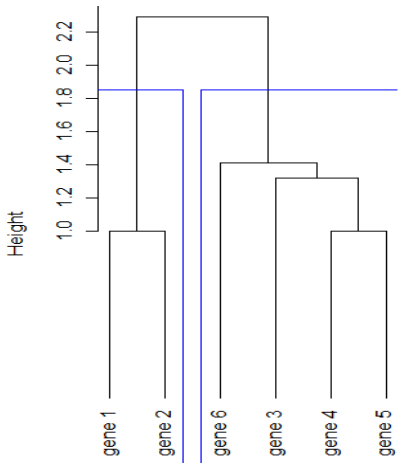

Figure 3. The ensemble clusters obtained using HC-CL from the example problem.

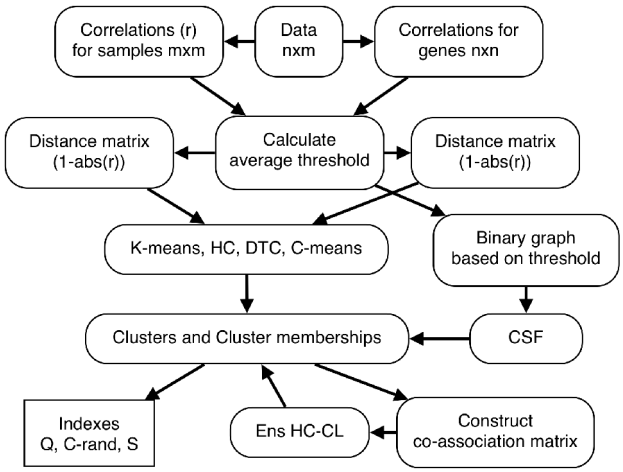

Figure 4. The workflow to generate the partitions, the cluster ensemble and the index values, adapted from [5].

The workflow starts with reading the gene expression data set that is represented as an *n x m* matrix. Correlation values between objects (genes) and attributes (samples) are calculated, and then the average of the absolute correlation values are used as a threshold to convert the complete graph to a binary graph to be used in the CSF algorithm. A distance matrix is constructed from the correlation matrix. Distance matrices for genes and samples are used in the K-means, HC, and the DTC algorithms to generate the partitions.

We use the CSF algorithm [20] for the ease of implementation from the igraph package [21] in R [22]. We use DTC as in [23]. We report the best result obtained using various parameters for DTC. We use K-means HC as implemented in the R base package, and the fuzzy C-means method as implemented in the R package e1071.

Internal and external cluster validation indices are used to evaluate the performance of the individual algorithms and the ensemble clustering. Silhouette, modularity and C-rand values are calculated using the clusterSim, igraph and flexclust packages in R, respectively. The silhouette index reflects the compactness and separation of clusters [24].

The silhouette index can take on values between -1 and 1, with higher values indicating better partitions. Modularity is a measure of data represented as a network, which is represented by the difference between the fraction of edges in the same clusters and the fraction of edges in the same clusters if they were connected randomly [7]. The modularity index can take on values between -1/2 and 1, with higher modularity values also indicating better partitions. C-rand is a measure of agreement between two partitions. The maximum value of the C-rand index is 1, meaning that two partitions (one of the partitions is the true "correct" partition) are exactly the same. It can also take negative values.

The performances of the individual algorithms and the ensemble approach based upon the three cluster validation indices are summarized in Table 1.

Table 1. Performance measures of individual and ensemble clustering algorithms for both genes and samples of three data sets.

Gene/Protein, Sample						
Dataset	**Method**	**Real**	**Alt**	**Silhouette**	**C-rand**	**Modularity**
Breast B	HC-CL	-, 4	3, 4	0.008, 0.364	-, 0.0189	0.014, 0.037
(1213 genes)	CSF	-, 4	38, 3	-0.072, 0.324	-, 0.206	0.032, **0.130**
(49 samples)	DTC	-, 4	12, 2	-0.037, **0.400**	-, -0.002	0.028, -0.032
(real cluster for samples)	K-means	-, 4	3, 4	**0.039**, 0.091	-, 0.230	**0.069**, 0.061
	C-means	-, 4	3, 4	0.019, 0.091	-, **0.230**	0.056, 0.061
	Ens. HC-CL	-, 4	4, 4	0.025, 0.046	-,0.125	0.027, 0.062
Yeast	HC-CL	5, -	5, 4	0.390, 0.290	0.452, -	**0.409**, 0.425
(384 genes, 17 samples)	CSF	5, -	4, 4	0.216, 0.279	0.316, -	0.346, 0.428
(real clusters if genes)	DTC	5, -	7, -	0.395, -	0.353, -	0.396, -
	K-means	5, -	5, 4	0.417, 0.176	0.500, -	0.403, 0.376
	C-means	5, -	5, 4	**0.497**, 0.290	**0.500**, -	0.405, 0.425
	Ens. HC-CL	5, -	5, -	0.151, **0.343**	0.351, -	0.215, **0.451**
Protein	HC-CL	4, -	4, -	0.344, 0.120	**0.199**, -	0.063, 0.304
(698 proteins, 125 samples)	CSF	4, -	5, 8	0.188, **0.410**	0.135, -	**0.097**, 0.164
(real cluster of proteins)	DTC	4, -	9, 5	-0.297, 0.088	0.081, -	0.018, 0.317
	K-means	4, -	4, 4	**0.379**, 0.113	0.127, -	0.056, 0.291
	C-means	4, -	4, 4	0.379, 0.075	0.127, -	0.062, **0.443**
	Ens. HC-CL	4, -	4, 3	0.078, 0.044	0.157, -	0.022, 0.180

The table exhibits silhouette, C-rand, and modularity values for the partitions identified in three different data sets, one for genes/proteins and one for samples. The number of clusters in a partition known *a priori* is given under the "Real" column, and the number of clusters selected by the user is given under "Alt" column of the table for each clustering method.

The "Breast B" data set includes 1213 genes and 49 samples. Real clusters (classes) are known for the samples, since the samples are the microarray experiments designed as control vs. treated or time course. The known clustering for Breast B was based upon the known estrogen receptor (ER) types: 25 ER positive samples and 24 ER negative samples. The final clustering into four classes consists of: 13 ER+ LN+, 12 ER+ LN-, 12 ER- LN-, 12 ER- LN- samples, where LN stands for lymphnode tumors [3]. The ensemble clustering using hierarchical clustering with complete linkage (Ens HC-CL),

choosing the number of desired clusters (Alt values in the table) as 4, resulted in the second highest silhouette value among all clustering results for the genes.

The yeast cell cycle data set is comprised of 384 genes measured over 17 time points (samples), with final values obtained by normalization and standardization. The real clusters correspond to 5 yeast cell cycles: early Gap 1 (G1) (beginning of Interphase), late G1, S (Synthesis), G2 (Prometaphase), M (Metaphase) [4]. The ensemble clustering approach resulted in the highest silhouette and modularity values (see the bolded values in the table) among all clustering algorithms for the set of samples.

The protein data set contained 698 proteins from 125 samples. The real clusters correspond to the 4 classes of protein-folds: α, β, α/β and α+β protein classes. The ensemble clustering approach found the second best C-rand value for the proteins shown in the table.

The ensemble clustering approach did not produce any negative values for any of the cluster validation indices. Hence, Ens HC-CL improves the negative valued partitions.

K-means and C-means resulted in high silhouette and C-rand values. The reason behind this may be that the data follow a specific distribution or have a specific simple shape, e.g., sphere, that K-means imposes [2].

HC and CSF gave high modularity values, which is reasonable since CSF is generally used to maximize the modularity [7] and there are also community structure finding algorithms, see [7], which are hierarchical like HC.

We wanted to see if the ensemble clusters obtained using the ensemble method generally agreed with the biological literature using the "Search for relationships between many genes, proteins, or keywords" option provided by Chilibot [25], [26], which is a freely available online tool that searches abstracts from the PubMed literature database [27] for specific relationships between proteins, genes, and keywords [5]. In this preliminary analysis, we examined only one cluster obtained from the Breast B data set.

We searched for the keyword "Cancer" along with the gene names contained in cluster 1 as we did in [5]. As output, Chilibot provided the number of PubMed literature database abstracts containing the keyword "Cancer" associated with a particular gene, along with the number of gene interactions, i.e., the number of abstracts where two genes appear together, for each gene. We refer to this later value as the "number of hits." Chilibot cannot handle more than fifty items per search [25]. Since the cluster had more than fifty genes, we split the search into smaller searches of appropriate size.

We calculated the average number of links from a gene to all other genes, along with the average percentage of genes to which each gene is linked in the cluster. The value of average hits per gene and average hit percentage in cluster is 12.5 and 9.05 respectively. These numbers appear to be rather high considering the analysis that we performed in [5], where we carried out a similar analysis for clusters of genes of Breast B data set created by CSF and DTC algorithms. Since we have not performed Chilibot analysis for all the clusters in this paper, we cannot make a fair comparison of our available results with the results in [5]. However, at least with respect to the result corresponding to cluster 1, it seems that there are a considerable number of associations between the genes in this cluster based upon the PubMed database.

4. CONCLUSION

In this study, we investigated an application of the ensemble clustering approach described in [2] using five different clustering algorithms that have not been reported in an ensemble framework before. We also evaluated the relative performance of the individual algorithms and the ensemble approach on three different biological data sets using two internal (silhouette and modularity) and one external (C-rand) validation index.

Computational experiments show that the ensemble clustering approach used tended to improve the quality of clusters for two of the data sets, based upon the ensemble clustering producing the best and second best values for at least one cluster validation index.

Using a different clustering algorithm than hierarchical clustering for the ensemble approach may improve the results generated by the ensemble as in [16]. Further experiments and investigation of a different combination of clustering algorithms and algorithm parameter settings and/or parameter settings for the ensemble approach are intended for future studies. Supplementary data including figures of the clusters from three data sets are available upon request.

5. REFERENCES

1. Hu, X. and Yoo, I., Cluster ensemble and its applications in gene expression analysis. In *2nd Asia-Pacific Bioinformatics Conference*. Dunedin, New Zealand. 29. Australian Computer Society, 2004.
2. Fred, A. L. N.; Jain, A. K. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **27** (6): 835–849, 2005.
3. Nascimento, M. C. V.; Toledo, F. M. B.; Carvalho, A. C. P. L. F., Investigation of a new GRASP-based clustering algorithm applied to biological data. *Computers and Operations Research*. **37**: 1381–1387, 2010.
4. Yeung, K. Y. ; Fraley, C.; Murua, A.; Raftery, A. E. and Ruzzo, W. L., Model-based clustering and data trasformations for gene expression data. *Bioinformatics*. **17**: 977–987, 2001.
5. Pirim, H.; Gautam, D.; Bhowmik, T.; Perkins, A. D. and Eksioglu, B., Performance evaluation of a community structure finding algorithm using modularity and C-rand measures. In *IEEE International Joint Conference on Neural Networks*, Barcelona, Spain, 18 – 23 July, 2010.
6. Rousseeuw, P. J., Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. **20**: 53–65, 1987.
7. Newman, M. E. J. and Girvan, M. Finding and evaluating community structure in networks, *Physical Review E*. **69**, 2004.
8. Hubert, L. and Arabie, P., Comparing partitions. *Journal of Classification*. **2**: 193–218, 1985.
9. Asur, S.; Parthasarathy, S. and Ucar, D., An ensemble approach for clustering scale-free graphs. In *LinkKDD*. Philadelphia, PA, 20 August. ACM, 2006.
10. Asur, S.; Ucar, D.; and Parthasarathy, S., An ensemble framework for clustering

protein-protein interaction networks. *Bioinformatics* **23**: i29-i40, 2007.
11. Watts, D. and Strogatz, S., Collective dynamics of small world networks. *Nature*. **393**(6684): 440–442, 1998.
12. Wang, T., Comparing Hard and Fuzzy C-Means for Evidence-Accumulation Clustering, In *FUZZ-IEEE*. Korea, 20-24 August, 2009.
13. Bezdek, J. C., *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum, 1981.
14. Avogadri, R.; and Valentini, G., Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. *Artificial Intelligence in Medicine* **45**: 173-183, 2008.
15. Hore, P.; Hall, L. O. and Goldgof. D. B., A scalable framework for cluster ensembles. *Pattern Recognition*. **42**: 676 – 688, 2009.
16. Galluccio, L.; Michel, O. J.J.; Comon, P.; Hero, Alfred O.; Kliger, M., Combining multiple partitions created with a graph-based construction for data clustering. *IEEE International Workshop on Machine Learning for Signal Processing, Grenoble: France*., 2009.
17. Fisher, R. A., The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**: 179-188, 1936.
18. Mangasarian, O. L. and Wolberg W. H., Cancer diagnosis via linear programming. *SIAM News*. **23**(5): 1–18, 1990.
19. Asuncion, A. and Newman, D. J., UCI machine learning repository. http://archive.ics.uci.edu/ml/, 2007.
20. Newman, M. E. J., Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*. **74**, 2006.
21. Csardi, G. and Nepusz, T., The igraph software package for complex network research. *InterJournal, Complex Systems*. 1695, 2006.
22. R Development Core Team. R: A language and environment for statistical computing. Version 2.9.2. from the World Wide Web: http://www.R- project.org.
23. Langfelder, P.; Zhang, B. and Horvath, S., Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics Applications Note*. **24** (5), 2008.
24. Bandyopadhyay, S.; Mukhopadhyay, A.; and Maulik, U., An improved algorithm for clustering gene expression data. *Bioinformatics*, **23**(21): 2859–2865, 2007.
25. Chan, H. and Sharp, M. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics.* **5**(1): 147, 2004.
26. “Chilibot” http://www.chilibot.net/ (current 15 May, 2010).
27. “PubMed,” http://www.ncbi.nlm.nih.gov/pubmed (current 22 April, 2010).