A Comparative Study of Clustering Algorithm For Lung Cancer Data

R. Krishnamoorthy [#1] and I. Laurence Aroquiaraj [*2]

Department of Computer Science, Periyar University
Salem - 636011, Tamil Nadu, India
krishna85pu@gmail.com
laurence.raj@gmail.com

Abstract— Literature survey is most important for understanding and gaining much more knowledge about specific area of a subject. In this survey on become an important topic in effective analysis of modify expression data due to its wide application in the biomedical industry. Gene clustering is the process of grouping related genes in the same cluster is at the foundation of different genomic studies that aim at analyzing the function of genes.K-means is a popular clustering algorithm that requires a huge initial set to start the clustering. K-means is an unsupervised clustering method which does not guarantee convergence. Numerous improvements to K-means have been done to make its performance better. Expectation Maximization is a statistical technique for maximum likelihood estimation using mixture models. It searches for a local maxima and generally converges very well. The proposed algorithm combines these three algorithms to generate optimum clusters which do not require a huge value of K and each cluster attains a more natural shape and guarantee convergence. Survival prediction is often accomplished by the TNM system that involves only three factors: tumor extent, lymph node involvement, and metastasis. Based on availability of large cancer datasets, it is possible to establish powerful prediction systems by using machine learning procedures and statistical methods. first briefly introduce the concepts of Clustering Techniques and discuss the basic elements of clustering on lung cancer data. The three benz mark algorithm of the most representative off-line clustering techniques,Expectation Maximization, fuzzy K-means clustering using Expectation Maximization, Fuzzy C- means clustering. The techniques are implemented and tested against a Lung Cancer Dataset. The performance of the three techniques are presented and compared.

Keywords— Expectation Maximization, Fuzzy K-Means, Fuzzy C-Means, Lung Cancer Overview and Images.

1. Introduction

At the conclusion of this lesson you should be able to:

- Explain what and why we need Data Mining
- Understand the difference between Data Mining and KDD in Databases

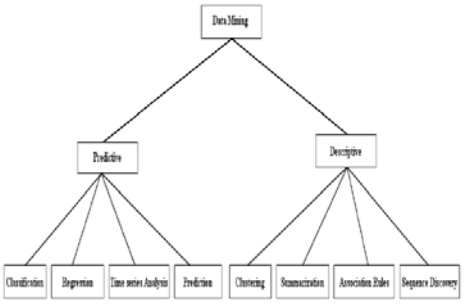
The past two decades has seen a dramatic increase in the amount of information or data being stored in electronic format. Data storage became as the availability of large amounts of computing power at low cost. There was also the introduction of new machine learning methods for knowledge representation based on logic programming etc, in addition to traditional statistical analysis of data. The new methods tend to be computationally intensive hence a demand for more processing power.

Database Management System (DBMS) gave access to the data stored but this was only a small part of what could be gained from the data. Traditional Online Transaction Processing (OLTP) are good at putting data into databases quickly, safely and efficiently but are not good at delivering meaningful analysis.

This is where Data Mining or Knowledge Discovery in Databases (KDD) has obvious benefits for the enterprises. It allows extracting diamonds of knowledge from historical data and predicting outcomes of future situations.

Data mining has been defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data".

Data Mining is briefly categorized into two models: Predictive and Descriptive. In turn, each model provides their nature specific tasks. These tasks are considered as the basic data mining tasks.



2. Research Background

Cancer is a class of diseases in which a cell, or a group of cells display uncontrolled growth (division beyond the normal limits), invasion (intrusion on and destruction of adjacent tissues), and sometimes metastasis (spread to other locations in the body via lymph or blood). These three malignant properties of cancers differentiate them from benign tumors, which are self-limited, and do not invade or metastasize.

3. Expectation Maximization Algorithm

The authors have designed an Intrusion response (IR) system cooperating with IDS using mobile agents distributed throughout the network, based on stigmergic properties. In [6], the authors introduced a self-organized ant colony based intrusion detection system (ANTIDS) to detect intrusions and compares its performance with linear genetic programming (LGP) [7], Support vector machines (SVM) [8] and Decision Trees (DT) [9]. Other works have made use of Multiple Adaptive Regression Splines (MARS) [10]. In [11], the authors have compared various data mining algorithms for detecting network intrusions.

The authors have used Naïve Bayes algorithm in building a network intrusion detection model [12]. In [13], the authors proposed Bayesian Belief network (BBN) with genetic and simulated annealing local search in order to build an efficient network intrusion detection model. Modeling intrusion detection system using hybrid intelligent systems is proposed in [15].

In this, DT and SVM The authors propose hierarchical Gaussian Mixture Model (HGMM) a novel type of Gaussian Mixture which detects network based attacks as anomalies using statistical processing classification in [18]. In [19], the authors use automated feature weighting for network anomaly detection.

They conclude that their proposed method not only increases the detection rate but also reduces false alarm rate as well. We believe that an unsupervised clustering approach offers some advantages over supervised learning approaches. One of the main benefits is that new applications can be identified by examining the connections that are grouped to form a new cluster.

The supervised approach can not discover new applications and can only classify network traffic for which it has labeled training data. This prompted us to use unsupervised clustering approach in building an efficient classifier for detecting anomaly based network intrusions.

Step 1:

Initialization step:

Initialize the hypothesis $\theta^0 = (\mu^0_1, \mu^0_2, \ldots, \mu^0_K)$

$$\theta^0_k = \mu^0_k$$

Where,

Notation	Descriptions
K	The current number of ans
σ	The standard Deviation
θ^0	0^{th} Iteration
μ	Mean Value

Table 3.1

Step 2:
Expectation step:

Estimate the expected value of the hidden variables z_{ij} (mean and standard deviation) using the current hypothesis $\theta^t = (\mu^t_1, \mu^t_2, \ldots, \mu^t_K)$

$$E(z_{ik}) = \frac{\exp\left[-\frac{(x_i - \mu^t_k)^2}{2\sigma^2}\right]}{\sum_{j=1}^{K} \exp\left[-\frac{(x_i - \mu^t_j)^2}{2\sigma^2}\right]}$$

$$E(z_{ik}) = \frac{\exp\left[-(x_i - \mu_k^t)^2\right]}{\sum_{j=1}^{K} \exp\left[-(x_i - \mu_j^t)^2\right]}$$

$$E(z_{ik}) = \frac{\exp\left[(x_i - \mu_k^t)^2\right]}{\sum_{j=1}^{K} \exp\left[(x_i - \mu_j^t)^2\right]}$$

Where,

Notation	Descriptions
t	The number of Iterations
$E(z_{ik})$	The Expected value of the hidden variable
k	Dimension
σ	Standard Deviation

Table 3.2

Step 3:

Maximization step:

It provide a new centroid of the parameters

$$\mu_k^{t+1} = \frac{\sum_{i=1}^{n} E(z_{ik}) x_i}{\sum_{i=1}^{n} E(z_{ik})}$$

Step 4:

Convergence step:

If the above current centroid values and previous centroid values are equal to finish the iteration. Otherwise go to step 2

4. Fuzzy K-Means Expectation Maximization Algorithm

Assigning each of the feature vectors to the nearest random seed vector, is the next step, and it can be achieved by computing the distance between each feature vector and all other seed vectors. Then the feature vector will be assigned to the seed vector such that the distance between them is the shortest. Also, each time an assignment happens the number of feature vectors assigned to that seed vector will be incremented. All seed vectors that are the centers of empty clusters, or have fewer vectors that selected *p* vectors, are eliminated and K is reduced.

Step 1:

Compute Weighted Fuzzy Average

Each cluster is then given a new prototype with the current K, and that would be the weighted fuzzy average (WFA) of each class, by initially taking the sample mean μ^0 and variance σ^2 to start the process. Then center a Gaussian over the current approximate WFA $\mu^{(r)}$ and iterate as follows.

$$w_p^{(r)} = \frac{\exp\left[-(x_p - \mu^{(r)})/2\sigma^2\right]}{\sum_{(m=1,P)} \exp\left[-(x_m - \mu^{(r)})/2\sigma^2\right]}$$

$$w_p^{(r)} = \frac{\exp\left[-(x_p - \mu^{(r)})\right]}{\sum_{(m=1,P)} \exp\left[-(x_m - \mu^{(r)})\right]}$$

$$w_p^{(r)} = \frac{\exp\left[(x_p - \mu^{(r)})\right]}{\sum_{(m=1,P)} \exp\left[(x_m - \mu^{(r)})\right]}$$

Where,

Notation	Descriptions
K	The current cluster
WFA	Waited Fuzzy Average
μ^0	Mean
σ^2	Variance
WFA $\mu^{(r)}$	Gaussian over the current approximate

Table 3.3

Step 2:

Compute Centroid Value

$$\mu^{(r+1)} = \sum_{(p=1,P)} w_p^{(r)} x_p, r = 0,1,2,.....$$

When eliminate the empty cluster, we get the selected previous vectors P, after that we go to EM algorithm

Step 3:

Initialization step:

Initialize the hypothesis $\theta^0 = (\mu^0_1, \mu^0_2, \ldots, \mu^0_K)$

$$\theta_k^0 = \mu_k^0$$

Where,

Notation	Descriptions
K	The current number of Gaussians
σ	The standard Deviation
θ^0	0^{th} Iteration
μ	Mean Value

Table 4.1

Step 4:

Expectation step:

Estimate the expected value of the hidden variables z_{ij} (mean and standard deviation) using the current hypothesis $\theta^t = (\mu^t_1, \mu^t_2, \ldots, \mu^t_K)$

$$E(z_{ik}) = \frac{\exp\left[-\frac{(x_i - \mu_k^t)^2}{2\sigma^2}\right]}{\sum_{j=1}^{K} \exp\left[-\frac{(x_i - \mu_j^t)^2}{2\sigma^2}\right]}$$

$$E(z_{ik}) = \frac{\exp\left[-(x_i - \mu_k^t)^2\right]}{\sum_{j=1}^{K} \exp\left[-(x_i - \mu_j^t)^2\right]}$$

$$E(z_{ik}) = \frac{\exp\left[(x_i - \mu_k^t)^2\right]}{\sum_{j=1}^{K} \exp\left[(x_i - \mu_j^t)^2\right]}$$

Where,

Notation	Descriptions
T	The number of Iterations
$E(z_{ik})$	The Expected value of the hidden variable
K	Dimension
σ	Standard Deviation

Table 4.2

Step 5:

Maximization step:

It provide a new centroid of the parameters

$$\mu_k^{t+1} = \frac{\sum_{i=1}^{n} E(z_{ik}) x_i}{\sum_{i=1}^{n} E(z_{ik})}$$

Step 6:

Convergence step:

If the above current centroid values and previous centroid values are equal to finish the iteration. Otherwise go to step 4

]5. Fuzzy C-Means Clustering.

Fuzzy C-Means (FCM) [31] is a method of Clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \|x_i - c_j\|^2$$

Once again, the expression $X = \{x_1, x_2,\ldots,x_n\}$ is a collection of data, where n is the number of data points and $C = \{c_1,c_2,\ldots,c_c\}$ is the set of corresponding cluster centers in the data set $\boldsymbol{X}$, where c is the number of clusters. u_{ij} is the membership degree of data x_i to the cluster centre c_j . Meanwhile, u_{ij} has to satisfy the following conditions:

$$u_{ij} \in [0,1], \quad \forall i = 1,\ldots,n, \forall j = 1,\ldots,c$$

$$\sum_{j=1}^{c} u_{ij} = 1, \quad \forall i = 1,\ldots,n$$

$U = (u_{ij})_{n*c}$ is a fuzzy partition matrix. $\|x_i - c_j\|$ Means the Euclidean distance between x_i and c_j. Parameter m is called the "Fuzziness Index", it is used to control the fuzziness of membership of each datum. The value of m should be within the range $m \in [1, \infty]$. There is no theoretical basis for the optimal selection of m, but value of m = 2.0 are usually chosen. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}, \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

Here in this step the iteration will be stopped when, $\max_{ij}\left\{\left\|u_{ij}^{(k+1)} - u_{ij}^{k}\right\|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

The algorithm is composed of the following steps:

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$

2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. Update$U^{(k)}$,$U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$$

4. If $\|U^{k+1} - U^{k}\| < \varepsilon$ then STOP; otherwise return to step 2.

6. Experimental Result Analysis.

A clustering method to form "natural" groups of the combinations. For reliable and easy interpretation in practice, we choose a hierarchical clustering approach. Figure 1 shows our final ensemble algorithm of clustering of cancer data (EACCD). Here the word ensemble refers to the sequence of the PAM procedures involved in the method. Early issues on ensemble clustering were discussed in [4] from the perspective of evidence accumulation. The work in [3] combined the *K*-means algorithm and linkage methods to form an ensemble method of discovering sample classes using gene expression profiles.

The evaluate the performance of the various proposed algorithms Expectation Maximization clustering, Fuzzy K-Means clustering using Expectation Maximization and Fuzzy C-Means, these approaches were implemented in MATLAB. The lung cancer data was used for our experiments. This is the real time data set. The Xie - Beni index was used as validation measure for comparative analysis. The dataset is explained below.

Table of Lung Cancer Data

Patient	Survival time (X)	Stage (X_1)	Grade (X_2)	Histology (X_3)	Gender (X_4)
1	64	1	2	squamous	1
2	24	1	3	large	1
3	24	2	3	squamous	1
4	8	1	2	squamous	1
5	16	3	3	squamous	2
... ...	... ...	... ...	... ...	... ...	... ...
1402	4	4	4	squamous	2

6.1 Validity Measures

The validity indices [7] are used for measuring "goodness" of a clustering result comparing to other ones which were created by other clustering algorithms, or by the same algorithms but using different parameter values.

If the number of clusters is not known prior to commencing an algorithm, the clustering validity index may also be used to find the optimal number of clusters [27]. Xie-Beni Validity Index is discussed in [27, 5] which measures the compactness and separation of clusters.

6.2 Xie - Beni Validity Measure

In this study, the Xie – Beni index has been chosen as the cluster validity measure because it has been shown to detect the correct number of clusters in several experiments [5]. This validation is the combination of two functions. The first calculates the compactness of data in different clusters.

Table Notation used in Validity Index

Notations	**Validity Index**
k	Number of clusters
d	Number of dimensions
d(x, y)	Distance between two data elements
$\overline{X_j}$	Expected value in the j^{th} dimension
n_{ij}	Number of elements in i^{th} cluster j^{th} dimension
n_j	Number of elements in j^{th} dimension in the whole data set
z_i	Centre point of the i^{th} cluster
c_i	i^{th} cluster
$\|c_i\|$	Number of elements in the i^{th} cluster

Let S is the overall validity index be the compactness and s be the separation of the Brain Tumor dataset. The validity index expressed as:

$$S = \frac{\pi}{s}$$

Where.

$$\pi = \frac{\sum_{i=1}^{\kappa}\sum_{j=1}^{n}\mu_{ij}{}^{\varepsilon}\|X - Z_i\|^{\varepsilon}}{n}$$

And

$$S = (d_{min})^2$$

$$d_{min} = min_{ij}\|z_i - z_j\|$$

d_{min} Is the minimum distance between the cluster centers.Where n is the number of genes, k is the number of clusters, and z_i is the cluster centre of cluster Ci, μ_{ij} is taken as 0.5 for the elements. Smaller values p is indicate that the clusters are more compact and larger of s indicated the clusters are well separated. In this study, Xie- Beni validity index is used to validate the clusters obtained after applying the clustering algorithms.

7. Comparative Analysis Based On Clusters

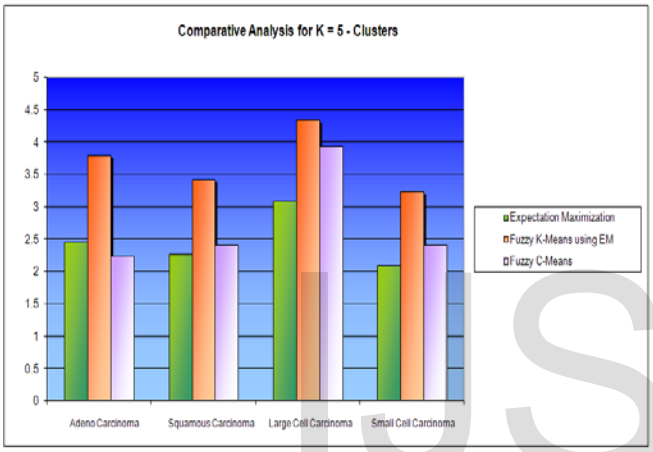
The comparative results based on Xie - Beni validity measure for all the depicted Gene expression data clustering algorithms for various Lung Cancer data sets taken for K= 5-clusters from the table.

Clustering Algorithm	**DATASETS (XIE-BENI INDEX)**			
	Adeno Carcinoma	Squamous Carcinoma	Large Cell Carcinoma	Small Cell Carcinoma

Expectation Maximization	2.4645	2.2755	3.0882	2.1032
Fuzzy K-Means using EM	3.8058	3.4215	4.3458	3.2410
Fuzzy C-Means	2.2468	2.4158	3.9421	2.4125

$\epsilon = 0.04$

The above table shows the comparative analysis for 5 clusters. Here the threshold value is 0.04 for the Expectation Maximization, Fuzzy K-Means using EM and Fuzzy C-Means clustering algorithms. In the above comparison table the Fuzzy K-Means using EM Clustering algorithm produces the better result compared to other algorithms



From the survey of above chart represent the comparative analysis of various approaches for K= 5 - cluster. It can be observed from the chart that the Fuzzy K-Means using EM Clustering algorithm construct the enhanced outcome compared to other algorithms

8. CONCLUSION

The paper reviews the problems with simple K-means, and suggests improvements to the method. The Expectation Maximization is used to fit data better when the distribution or model of the data is known. When these two are combined we will get a clustering method that not only fits the number of clusters but also tries to make them compact and more meaningful. The Statistical techniques have been combined before with fuzzy logic theory and have shown to yield good results. We are using the new approach, the Gaussian for the clusters fits the data better using EM along with Fuzzy K-means.

Since the new method finds results for small value of K selected initially we can argue that we reduce the number of iterations overall. Since we only use the centers chosen initially and get centroids for the data. The EM can still perform to make the K-means work correctly, since EM will iterate to find best centers for the given data.

the comparative analysis of three clustering algorithms viz., Expectation Maximization, Fuzzy K-Means using EM and Fuzzy C-Means clustering algorithms have been analyzed using the Xie-Beni index for measuring the validation to compact and well separated the clusters. This measure is used to find the minimum value in the clustering. The four different datasets i.e., Adeno, Squamous, Large and Small cell carcinoma treatment outcome have been used for comparative analysis.

9.REFERENCES

[1] M.S Aldenderfer, R.K. Blashfield, Cluster Analysis, Sage Publications, Beverly Hills, USA, 1984.

[2] M.J.A. Berry, G. Linoff, Data Mining Techniques for Marketing, Sales and Customer Support. John Wiley & Sons, NY, USA, 1997.

[3]. J.B. MacQueen, Some methods for classification and analysis of multivariate observations, Proc. 5th Berkeley Symp. Probability Statistics, University of California Press, Berkeley, 1967, pp. 281-297.

[4]. P.S. Bradley, Usama M. Fayyad , Refining Initial Points for K-Means Clustering, Proc. 15th International Conf. on Machine Learning, 1998.

[5]. E. Forgy, Cluster analysis of multivariate data: efficiency versus Interpretability of classifications, Biometrics, 21, 1965, 768-776.

[6]. S.Z. Selim, M. A. Ismail, k-means type algorithms: a generalized convergence Theorem and characterization of local optimality, IEEE Trans. Pattern Analysis Machine Intelligence, 6, 1984, 81-87.

[7]. C. G. Looney, Interactive clustering and merging with a new fuzzy expected value, Pattern Recognition Lett., vol. 35, 2002, 187-197.

[8]. N. Watanabe, T. Imaizumi, Fuzzy k-means clustering with crisp regions, The 10th IEEE International Conference on Fuzzy Systems, Melbourne, 2001, p p.199-202.

[9].Thestatistichomepage,http://www.statsoftinc.com/textbook/stathome.html, January 17, 2005.

[10].The Expectation Maximization Algorithm, http://www.cs.unr.edu/~bebis/MathMethods/EM/lecture.pdf, September 25, 2004.

[11]. Hyungsuck Cho, Opto-Mechatronic Systems Handbook: Techniques and Applications, CRC Press, KAIST, Taejeon, South Korea, 2002

[12]. Programs and Data for Downloading, http://ultima.cs.unr.edu/programs.htm, October 4, 2004

[13]. Z. Ghahramani, M. Jordan. Supervised learning from incomplete data via an EM Algorithm. Advances in Neural Information Processing Systems, vol. 6, 1994, 120-127.

[14]. S Gunn, Support Vector Machines for Classification and Regression, Technical Report. Image, Speech and Intelligent Systems Group University of Southampton, 1998.

[15]. A. P. Dempster, N.M. Laird, D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of Royal Statistics Society, vol. B- 39, 1977.

[16]. H. Surmann, A. Kanstein, K. Gosser, Self-Organizing and Genetic Algorithms for an Automatic Design of Fuzzy Control and Decision Systems, EUFIT 93 First European Congress on Fuzzy and Intelligent Technologies, Aachen, Vol. 2, 1993, pp. 1097-1104.

[17]. L.A. Zadeh, Fuzzy logic and approximate reasoning. Synthese, vol. 30, 1975, 407--428.

[18]. Vladimir Estivill-Castro, Jianhua Yang , A Fast and Robust General Purpose Clustering Algorithm, Pacific Rim International Conference on Artificial Intelligence, 2000, pp: 208-218.

[19]. P. Bradley, U. Fayyad, C. Reina, Scaling EM (Expectation Maximization) Clustering to Large Databases, Microsoft Research Report, MSR-TR-98-35,Aug 1998.

[20]. Wael Abd-Almageed, Christopher E. Smith, Mixture Models for Dynamix Statistica.Pressure Snakes, in Proc. IEEE International Conference on Pattern Recognition, August 2002, pp. 721-724.

[21]. A.V. Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, C. Mao, and K. Murphy, A coupled HMM for audio-visual speech recognition, in International Conference on Acoustics,Speech and Signal Processing (CASSP'02), Orlando, FL, USA, May 2002, 13-17.

[22]. A. Abu-Naser, N.P. Galatsanos, M.N. Wernick, D. Schonfeld, Object recognition based on impulse restoration with use of the expectation-maximization algorithm, Journal of the Optical Society of America A (Optics, Image Science and Vision), vol. 15, no. 9, September 1998, pp. 2327 - 40.

[23]. Wael Abd-Almageed, Aly El-Osery, Christopher E. Smith, Non-Parametric Expectation Maximization: A Learning Automata Approach, IEEE Conference on Systems, Man and Cybernetics, Washington DC, 2003.

[24]. Fisher,R., The use of multiple measurements in taxonomic problems. Ann. Eugenics, **7**, 1936, 179-188.

[25]. I. Gath, and A.B. Geva, Unsupervised Optimal Fuzzy Clustering IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 11, pp. 773-781, 1989. Conference on Machine Learning and Applications, 11-13 Dec. 2008, San Diego,California, USA.

[26]. Chen, D., Kai, X., Henson, D., Sheng, L., Schwartz, A. M., and Cheng, X. A clustering-based approach to predict outcome in cancer patients. To appear in Int. J.Data Mining and Bioinformatics.

[27]. Chen, D., Zhang, Z., Liu, Z. and Cheng, X. An Ensemble Method of Discovering Sample Classes Using Gene Expression Profiling. Data Mining in Biomedicine,Series: Springer Optimization and Its Applications , Vol. 7, Pardalos, Panos M.; Boginski, Vladimir L.; Vazacopoulos, Alkis (Eds.), Springer, 2007, pp. 39 - 46.

[28]. Fred, A. and Jain, A. K. Data clustering using evidence accumulation. Proc. of the 16th International Conference on Pattern Recognition, ICPR 2002, 276- 280.

[29].Greene, F. L., Compton, C. C., Fritz, A. G., Shah, J. P.,and Winchester, D. P. (Editors). AJCC Cancer Staging Atlas. Springer, 2006.

[30]. Kaplan, E. L., Meier, P. Nonparametric estimation from incomplete observations. J Am Stat Assoc, 1958, 53, 457-481.

[31]. Kaufman, L., and Rousseeuw, P.Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, 1990.

[32]. Klein, J.P. and Moeschberger, M. L. Survival Analysis: Techniques for Censored and Truncated Data, 2nd edition. 2003, New York: Springer Verlag.

[33]. Xing, K., Chen, D., Henson, D., and Sheng, L.A clustering-based approach to predict outcome in cancer patients.