Asian Journal of Information Technology 10 (4): 142-148, 2011
ISSN: 1682-3915

An Effective Classification Technique for Microarray Gene Expression by Blending of LPP and SVM

[1]J. Jacinth Salome and [2]R.M. Suresh
[1]Department of Computer Science, Arignar Anna Government Arts College,
Walajapet, Vellore District, Tamil Nadu, India
[2]R.M.D Engineering College, Tamil Nadu, India

Abstract: Now a days, data mining has been exploited to retrieve the valuable information in a wide spread fields especially in DNA microarray technology. The DNA mciroarray technology produces a huge amount of gene data i.e., expression levels of thousands of genes for a very few samples. From the microarray gene data, the process of extracting the required knowledge remains an open challenge. In order to retrieve the required information, gene classification is vital however, the task is complex because of the data characteristics, high dimensionality and smaller sample size. In this study, the propose an effective gene classification technique based on LPP and SVM. In the proposed gene classification technique firstly, the high dimensionality of the microarray gene data is reduced using LPP. The LPP is chosen for the dimensionality reduction because of its ability of preserving locality of neighborhood relationship. Secondly, the SVM is trained by the dimensionality reduced gene data for effective classification. SVM has the ability to learn with very few samples and so it is selected for the proposed technique. Hence, the classification technique developed with the blending of LPP and SVM results in effectual and powerful classification of gene expression data. Moreover, a comparative study is made with the ANN-based and PCA-based gene classification techniques.

Key words: Gene classification, microarray gene expression, dimensionality reduction, Locality Preserving Projection (LPP), Support Vector Machine (SVM), India

INTRODUCTION

Extremely large volume of data and small amount of knowledge extraction competence enhances the attention in the field of data mining (Dehuri and Cho, 2008). Data mining or the proficient innovation of valuable, non-obvious information from a massive collection of data (Bigus, 1996) has an objective to discover knowledge out of data and present it in a form that is easily understandable to humans (Labib and Malek, 2005). Data mining plays a vital role in twenty first century-the information age (Ramamohanarao *et al*., 2005). Mostly the data mining tasks consist of classification, regression, clustering, rule generation, discovering association rules, summarization, dependency modeling and sequence analysis (Mitra *et al*., 2002). They contribute in various fields of research such as information sharing and collaboration, security association mining, classification and clustering, intelligence text mining, spatial and temporal crime pattern mining and criminal/terrorist network analysis and more (Chen, 2008). DNA microarray technology is also the field that exploits the data mining techniques.

Presently, the enhanced DNA microarray technology has resulted in expression levels of thousands of genes being recorded over just a few tens of different samples (Shang and Shen, 2005). While the DNAmicro array technology considerably expedite the procedure of discovering the utility of genes, the amount of data generated by this technology also pretenses a challenge for the biologists to carry out the analysis (Kim *et al*., 2006). Also, the molecular biologists face the challenges in determining the required knowledge from this kind of enormous amount of data (Slavkov *et al*., 2005). In this kind of knowledge seeking applications, information retrieval is one of the primary and most important technologies (Lee, 2007) to extract the entailed knowledge from the huge amount of data. Normally, information retrieval is a selection process in which the required information is extracted from a database (Wolfram, 2000). In microarray data analysis, the process of information retrieval system includes diagnosis of disease, categorizing disease and getting information which is useful to give possible treatments (Slavkov *et al*., 2005). This makes the gene classification as one of the main

Corresponding Author: J. Jacinth Salome, Department of Computer Science, Arignar Anna Government Arts College, Walajapet, Vellore District, Tamil Nadu, India

tasks in microarray gene expression analysis (Leung and Hung, 2009) because it is a basis for prediction of the functions of unknown genes (Hori *et al*., 2001).

In general, classification is the process of recognizing a set of models that demonstrate and differentiate data classes or concepts for the intention of being able to use the model in order to forecast the class of objects which has unknown class label (Zhong *et al*., 2006). It is one of the major data mining functions (Waiyamai *et al*., 2004) and a dynamic research region in the perspective of data stream (Ling *et al*., 2009). Classification is a major apprehension in most of the engineering and scientific regulation that includes biology, psychology, marketing, computer vision, artificial intelligence and medicines (Pradhan *et al*., 2009). Also, the classification of gene expressions has become the subject of numerous researches in order to find out the functionality of known or unknown genes (Shang and Shen, 2005).

In the process of information retrieval in DNA microarray technology, gene classification is quite tough task because of the characteristics of the data which contain high dimensionality and small sample size (Leung and Hung, 2009). A combination of the tactics is repeatedly used in practice forclassification with gene expression data. Such classification measures normally contain the following steps: gene selection/dimension reduction in which a small amount of gene components are constructed from a huge number of genes and classification in which the samples are categorized into groups by applying standard statistical models on the gene components (Dai *et al*., 2006). Microarray experiments normally produce a large amount of datasets with expression values for thousands of genes but still not more than a few dozens of samples thus very exact arrangement of tissue samples in such high dimensional problems is a tricky task (Zhang *et al*., 2007). Moreover, there is a high redundancy in microarray data and numerous genes contain inappropriate information for precise classification of diseases or phenotypes (Osareh and Shadgar, 2009). So, an effective classification technique is necessary to get back the gene information from the microarray experimental data.

For the purpose of retrieving information from a microarray gene expression, researchers propose an effective gene classification technique based on LPP and SVM. As a first process in the proposed gene classification, the high dimensionality of the microarray gene data is reduced using LPP. The LPP is chosen for the dimensionality reduction because of its ability of preserving locality of neighborhood relationship. Next, the SVM is trained by the dimensionality reduced gene data for effective classification. SVM has the ability to learn with very few samples and so it is selected for the proposed technique. Hence, the classification technique developed with the blending of LPP and SVM results in effectual and powerful classification of gene expression data. Moreover, a comparative study is made with the ANN-based and PCA-based gene classification techniques.

LITERATURE REVIEW

A stomach cancer detection system which was based on Artificial Neural Network (ANN) and the Discrete Cosine Transform (DCT) was developed by Sarhan (2009). The classification features were extracted from stomach microarrays using the DCT by the proposed system. The features were extracted from the DCT coefficients and then applied to an ANN for classification (tumor or none-tumor). The microarray images employed in his study were acquired from the Stanford Medical Database (SMD). Simulation results illustrated that the proposed system produced a very high success rate.

Hang and Wu (2009) have discussed about an approach for cancer diagnosis using gene expression data. Their method symbolized each testing sample as a linear combination of all the training samples. The coefficient vector was acquired by l_1-regularized least square. Classification was accomplished by defining discriminating functions from the coefficient vector for each individual category. l_1 norm minimization led to sparse solution and they named the new approach as sparse representation. Numerical experiments proved that the sparse representation approach matched the best performance accomplished by Support Vector Machines (SVM).

Sheng *et al*. (2009) have enhanced Block Diagonal Linear Discriminant Analysis (BDLDA) (Pique-Regi and Ortega, 2006) and employed it to gene expression data. They enhanced feature selection in BDLDA by making use of an estimated error rate to choose the best model among all the candidate models. The estimated error rate was formulated from LDA and could be derived for each candidate block diagonal covariance structure. Their algorithm was optimized by repeatingthe model construction procedure after the removal of earlier selected features which led to improved classification robustness. Their algorithm was tested by using 10 fold cross validation. Iwen *et al*. (2008) have proposed a method that considered a larger subset of CAR-related (conjunctive association rules) and Boolean Association Rules (BARs). To address the computational complexities included with pre-classification CAR mining, those rules were compactly captured in a Boolean Structure Table

(BST) which was then employed to generate a BST classifier called BSTC. In comparison to the present leading CAR classifier, RCBT on numerous benchmark microarray datasets have demonstrated that the BSTC is competitive with RCBT's accuracy while reducing the exponential costs acquired by CAR mining. For this reason, BSTC extended the generalized CAR-based methods to larger datasets. Besides, contrasting from other association rule-based classifiers, BSTC easily generalized to multi-class gene expression datasets. BSTC's worst case per-query classification time was worse then CAR-based methods after all exponential time CAR mining was concluded (O (|S|2 • |G|) versus O (|S| •|G|)).

Ruiz *et al.* (2006) have proposed a new heuristic to choose relevant gene subsets so as to use them for the classification task. Their method was on the basis of statistical significance of appending a gene from a ranked-list to the final subset. The efficiency and efficacy of their technique was established through widespread comparisons with other representative heuristics. Their approach demonstrated an excellent performance at recognizing relevant genes and also with respect to the computational cost.

Au *et al.* (2005) have proposed an attribute clustering method which was able to group genes based on their interdependence in order to mine meaningful patterns from the gene expression data. Their method grouped interdependent attributes into clusters by optimizing acriterion function obtained from an information measure that exhibited the interdependence between attributes. Meaningful clusters of genes were determined by applying their algorithm to gene expression data. To analyze the performance of their approach, they applied it to two recognized gene expression data sets and compared their results with those acquired by other methods. Their experiments proved that their method was able to determine the meaningful clusters of genes.

Shang and Shen (2005) and Hori *et al.* (2001) have proposed an application of supervised machine learning approaches to the classification of the yeast *S. cerevisiae* gene expression data. For the first time, established feature selection techniques based on information gain ranking and principal component analysis were employed to that data set to hold learning and classification. Different classifiers were implemented to examine the effect of combining feature selection and classification methods. Learning classifiers that were implemented comprise K-Nearest Neighbours (KNN), Naive bayes and decision trees. The provided results of comparative studies showed that effective feature selection is necessary for the development of classifiers anticipated for use in high dimension domains. Specifically amongst a large corpus of systematic experiments executed, best classification performance was accomplished using a subset of features chosen by means of information gain ranking for KNN and Naive bayes classifiers. Naive bayes was also carried out precisely with a comparatively small set of linearly transformed principal features in categorizing this complex data set. Their research also demonstrated that the feature selection aids to increase computational efficiency at the same time as to improve classification accuracy.

THE PROPOSED GENE CLASSIFICATION TECHNIQUE BASED ON LPP AND SVM

It is well known that the gene classification using microarray gene expression data is quite difficult because of the characteristics of the data, high dimensionality and small sample size. Researchers propose a technique for efficient gene classification based on LPP and SVM; the technique is described here. The proposed technique is comprised of two stages, dimensionality reduction and SVM-based classification. Let, the microarray gene expression data be: X_{jk}; $0 \le j \le n_g$, $0 \le k \le n_s$ where, n_g represents the number of genes from which the data is taken and n_s represents the number of samples. The gene data is of higher dimension and so it is subjected to dimensionality reduction. In the dimensionality reduction, the high dimensional gene data X_{ij} is converted to low dimensional data. The resultant low dimensional data is classified using a well-trained SVM. The SVM is trained by the gene data of different classes. Once the SVM is well trained by the low dimensional gene data of various classes, it will be ready to classify any of the similar gene expression data. So, prior to classification, the SVM has to be trained with the aid of the gene data of different classes, Y_{ijk}; $0 \le i \le n_c$ where n_c is the total number of classes. The training process of SVM is described further.

Dimensionality reduction by LPP: Dimensionality reduction, one of the two stages of the proposed gene classification technique is performed using LPP (He and Niyogi, 2003). From the gene data of different classes Y_{ijk}, a concatenated matrix is obtained as given in the Eq. 1. In the concatenated matrix, the gene data of all the classes are combined and it is given as a single matrix. The matrix Y_{conc} is given as follows:

$$Y_{conc_{jl}} = \sum_{i=0}^{n_s-1} Y'_{ijl} \quad (1)$$

Where:

$$Y'_{ijl} = \begin{cases} Y_{ijl}; & \text{if } l \in (i, n_s(i+1)-1) \\ 0; & \text{Otherwise} \end{cases} \quad (2)$$

The concatenated matrix Y_{conc} of dimension $n_g \times n'_s$; $n'_s = n_s.n_c$, $n'_s << cn_g$ which is highly dimensional and so the dimensionality of the matrix is reduced using LPP. The LPP is a linear dimensionality reduction algorithm that shares most of the properties of data representation of nonlinear techniques namely, locally linear embedding or Laplacian Eigenmaps. The LPP procedure for dimensionality reduction constitutes of three steps, namely, generation of distance matrix, determining adjacency matrix and calculating dimensionality reduced matrix.

Generation of distance matrix: For the concatenated matrix Y_{conc}, the distance matrix of size $n_g \times n_g$ is determined as follows:

$$D_{xy} = \sqrt{\sum_{l=0}^{n_s} (Y_{conc_{xl}} - Y_{conc_{yl}})^2}; \; 0 \le x, \; y \le n_g \tag{3}$$

The determined distance matrix is based on the Euclidean distance calculated by considering each row of the Y_{conc} as a network node. The resultant D_{xy} is subjected to calculate adjacency matrix which can be determined based on the relationship of an element with every neighbor elements.

Determination of adjacency matrix: In the virtual network consisting of n_g nodes, the adjacency matrix is a $n_g \times n_g$ with binary entries representing if there is an edge between two nodes. Here, the adjacency matrix W is determined with the aid of the D_{xy} as follows:

$$W_{xy} = \begin{cases} 1; & \text{if } D_{xy} > 0 \\ 0; & \text{Otherwise} \end{cases} \tag{4}$$

From the Eq. 4, it can be shown that the adjacency matrix W_{xy} is constituted of binary values depending upon the distance calculated in D_{xy}.

Calculation of dimensionality reduced matrix: From the adjacency matrix W, a diagonal matrix A is determined as follows:

$$A_{xy} = \begin{cases} S_x; & \text{if } x = y \\ 0; & \text{Otherwise} \end{cases} \tag{5}$$

Where:

$$S_x = \sum_{y=0}^{n_g - 1} W_{xy} \tag{6}$$

Based on the A which is obtained from the Eq. 5, Z_1 and Z_2 are calculated as follows:

$$Z_1 = \frac{1}{2}\left(A_P + A_P^T\right) \tag{7}$$

$$Z_2 = \frac{1}{2}\left(L_P + L_P^T\right) \tag{8}$$

In Eq. 7 and 8, A_p and L_p can be determined by $A_p = Y_{conc}.Y.Y'_{conc}$ and $L_p = A\text{-}W$, respectively. The obtained Z_1 and Z_2 are subjected to a generalized eigenvector problem (He and Niyogi, 2003) as follows:

$$Z_2 E = \lambda Z_1 E \tag{9}$$

Once the eigenvectors are determined, the embedding is performed as:

$$\hat{Y} = E^T Y_{conc} \tag{10}$$

The $\hat{Y}$ obtained from the above equation is the dimensionality reduced gene data with size $n'_s \times n'_s$. The $\hat{Y}$ is utilized to train the SVM to classify the input microarray gene data.

Training process of SVM: SVMs pertain to the generalized linear classifier's family. SVMs are also regarded as a special case of Tikhonov regularization. A peculiar property is that they lessen the empirical classification error and increase the geometric margin at the same time. Therefore, they are also called as maximum margin classifiers. The SVM training intends to minimize an error function that is given as:

$$\arg\min P \sum_{j=0}^{n'_s - 1} \Omega_j + 0.5\alpha^T.\alpha \tag{11}$$

With the following constraints:

$$O_j(\alpha^T \phi(\hat{Y}_j) + g) \ge 1 - \Omega_j \tag{12}$$

And:

$$\Omega_j \ge 0 \tag{13}$$

In Eq. 11, P is the penalty constant, Ω is a parameter that handles the data and α is a matrix of coefficients. In the constraints given in Eq. 12 and 13, O_j is the class label of the jth dataset, b is a constant and ϕ is the kernel that transforms the input data to the feature space.

Hence by minimizing the error function, the SVM learns the training gene dataset $\hat{Y}$ well and so that it can classify the gene dataset that are similar to the training set.

Classification of gene data by SVM: From the training gene data, the SVM learns well about the class under which the given gene dataset is present. Once the SVM is trained well, it attains the ability to classify any gene dataset in the similar fashion. In the classification, firstly the gene dataset to be classified is subjected to dimensionality reduction i.e., the dimension of the gene dataset X_{ij} is reduced using LPP. This dimensionality reduction is performed in the similar fashion as performed for the gene dataset Y_{ijk}. Then, the dimension-reduced matrix is given to the trained SVM and so the class of the given microarray gene data is obtained in an effective manner.

RESULTS AND DISCUSSION

The proposed technique for microarray gene classification has been implemented in the working platform of Matlab (Version 7.8). For evaluating the proposed technique, researchers have utilized the microarray gene samples of human acute leukemias. The SVM has been trained by two different classes of microarray gene data namely, Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). Thus obtained microarray gene expression data is of dimension, $n_c = 2$, $n_g = 7192$ and $n_s = 38$. The high dimensional gene expression data has been subjected to LPP-based dimensionality reduction and so a dimensionality reduced gene data with dimensions, 38×38 (i.e., $n_s = 38$) has been obtained. A sample of microarray gene dataset of two classes that has been used for training is shownn in the Table 1. While testing when a gene dataset is given the proposed technique has identified its belonging class. A sample of the gene dataset that has been subjected to classification and the class identified by the proposed technique is shown in the Table 2.

Some six samples for each cancer class are shown in the Table 1 and 2 for training and testing, respectively. In the testing, only the samples have been given and the proposed technique decides its belonging class. Thus classified samples are shown in the Table 2. The efficacy of the proposed technique has been determined by comparing it with some other classification techniques using Artificial Neural Network (ANN) and Principle Component Analysis (PCA). The comparison of the proposed technique with the ANN-based and PCA-based gene classification techniques with respect to the performance metrics, accuracy and error rate are shown in the Table 3. The performance metrics, accuracy and error rate can be calculated as follows:

$$\text{Accuracy} = \frac{\text{No. of samples classified exactly}}{\text{Totalno. of samples subjected to classification}} \quad (14)$$

$$\text{Error rate} = 1 - \frac{\text{No. of samples classified exactly}}{\text{Total no. of samples subjected to classification}} \quad (15)$$

From the Table 3, it can be shown that the proposed technique has provided more accuracy and less error rate rather than the ANN-based and PCA-based gene classification techniques. More accuracy and less error rate leads to effective classification of the given microarray gene data to the actual class of the gene.

Table 1: A sample of the microarray gene data corresponds to the cancer classes for training

	Sample											
	ALL (B-cell)						AML					
Gene	19769	23953	28373	9335	9692	14749	12	13	14	16	20	1
AFFX-BioB-5_at (endogenous control)	-214A	-135A	-106A	-72A	-413A	-67A	-20A	7A	-213A	-25A	-72A	-4A
AFFX-BioB-M_at (endogenous control)	-153A	-114A	-125A	-144A	-260A	-93A	-207A	-100A	-253A	-20A	-139A	-116A
AFFX-BioB-3_at (endogenous control)	-58A	265A	-76A	238A	7A	84A	-50A	-57A	136A	124A	-1A	-125A
AFFX-BioC-5_at (endogenous control)	88A	12A	168A	55A	-2A	25A	101A	132A	319A	325A	392A	241A
AFFX-BioC-3_at (endogenous control)	-295A	-419A	-230A	-399A	-541A	-179A	-370A	-377A	-209A	-396A	-323A	-191A
AFFX-BioDn-5_at (endogenous control)	-558A	-585A	-284A	-551A	-790A	-323A	-529A	-478A	-557A	-464A	-510A	-411A
AFFX-BioDn-3_at (endogenous control)	199A	158A	4A	131A	-275A	-135A	14A	-351A	40A	-221A	-350A	-31A
AFFX-CreX-5_at (endogenous control)	-176A	-253A	-122A	-179A	-463A	-127A	-365A	-290A	-243A	-390A	-202A	-240A
AFFX-CreX-3_at (endogenous control)	252A	49A	70A	126A	70A	-2A	153A	283A	119A	-1A	249A	149A
AFFX-BioB-5_st (endogenous control)	206A	31A	252A	-20A	-169A	-66A	29A	247A	-131A	358A	561A	24A

Table 2: A sample of the microarray gene data corresponds used to test the proposed technique

	Samples											
	ALL						AML					
Gene	19769TA+ Norel	406TA+(ML) Norel	4466 Norel	1245TA- Norel	16125TA- Norel	23368TA- Norel	15 PK) Norel	19 (PK) Relap	10 (PK) Relap	9 (PK) Relap	SH 5	SH 13
AFFX-BioB-5_at (endogenous control)	-214A	-342A	-87A	22A	-243A	-130A	-21A	-202A	-112A	-118A	-90A	-137A
AFFX-BioB-M_at (endogenous control)	-153A	-200A	-248A	-153A	-218A	-177A	-13A	-274A	-185A	-142A	-87A	-51A
AFFX-BioB-3_at (endogenous control)	-58A	41A	262A	17A	-163A	-28A	8A	59A	24A	212A	102A	-82A
AFFX-BioC-5_at (endogenous control)	88A	328A	295A	276A	182A	266A	38A	309A	170A	314A	319A	178A
AFFX-BioC-3_at (endogenous control)	-295A	-224A	-226A	-211A	-289A	-170A	-128A	-456A	-197A	-401A	-283A	-135A
AFFX-BioDn-5_at (endogenous control)	-558A	-427A	-493A	-250A	-268A	-326A	-245A	-581A	-400A	-452A	-385A	-320A
AFFX-BioDn-3_at (endogenous control)	199A	-656A	367A	55A	-285A	-222A	409A	-159A	-215A	-336A	-726A	-13A
AFFX-CreX-5_at (endogenous control)	-176A	-292A	-452A	-141A	-172A	-93A	-102A	-343A	-227A	-310A	-271A	-11A
AFFX-CreX-3_at (endogenous control)	252A	137A	194A	0A	52A	10A	85A	236A	100A	177A	-12A	112A
AFFX-BioB-5_st (endogenous control)	206A	-144A	162A	500A	-134A	159A	281A	-7A	307A	-131A	-104A	-176A

Table 3: Performance comparison between the proposed gene classification technique and the ANN-based as well as PCA-based gene classification techniques

	Gene classification techniques		
Performance metrics	Proposed classification technique	ANN-based gene classification	PCA-based gene classification
Accuracy (%)	97.2973	91.8919	83.3333
Error rate (%)	2.7027	8.1081	16.6667

CONCLUSION

In this study, researchers have proposed an effective gene classification technique based on LPP and SVM. The LPP has been utilized for the dimensionality reduction and the SVM has been trained for effectual gene classification. The technique has been tested by classifying the microarray gene expression data of human acute leukemias. The proposed technique has classified the *AML* and *ALL* gene expressions well. In the testing of the proposed technique when any of the gene expression data has been given, the SVM has identified the class of the corresponding data. As the LPP and SVM have good positive features in their task of dimensionality reduction and classification respectively, the blending of them has led the proposed technique to effectual and powerful classification. The comparative results have shown that the proposed technique possesses better accuracy and lesser error rate than the ANN-based and PCA-based gene classification techniques. Hence, this means of gene classification have paved the way for effective information retrieval in the microarray gene expression data.

REFERENCES

Au, W.H., K.C.C. Chan, A.K.C. Wong and Y. Wang, 2005. Attribute clustering for grouping, selection and classification of gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 2: 83-101.

Bigus, J.P., 1996. Data Mining with Neural Networks. McGraw-Hill, New York.

Chen, H., 2008. Homeland security data mining using social network analysis. Proceedings of the 1st European Conference on Intelligence and Security Informatics, (ISI'08), Esbjerg, Denmark, pp: 4-4.

Dai, J.J., L. Lieu and D. Rocke, 2006. Dimension reduction for classification with gene expression microarray data. Stat. Appl. Genet. Mol. Biol., 5: 1-21.

Dehuri, S. and S.B. Cho, 2008. Multi-objective classification rule mining using gene expression programming. Proceedings of the 3rd International Conference on Convergence and Hybrid Information Technology, Nov. 11-13, Busan, pp: 754-760.

Hang, X. and F.X. Wu, 2009. Sparse representation for classification of tumors using gene expression data. J. Biomed. Biotechnol., 2009: 403689-403689.

He, X. and P. Niyogi, 2003. Locality Preserving Projections. In: Advances in Neural Information Processing Systems, Becker, S., S. Thrun and K. Ober-Mayer (Eds.). MIT Press, Cambridge, MA. USA.

Hori, G., M. Inoue, S.I. Nishimura and H.I. Nakahara, 2001. Blind gene classification-an application of a signal separation method. Proceedings of the 12th International Conference on Genome Informatics, (CI'10), Tokyo, pp: 255-256.

Iwen, M.A., W. Lang and J.M. Patel, 2008. Scalable rule-based gene expression data classification. Proceedings of the IEEE 24th International Conference on Data Engineering (DE'08), USA., pp: 1062-1071.

Kim, S., Y. Tak and L. Tari, 2006. Mining Gene Expression Profiles with Biological Prior Knowledge. Proceedings of the Life Science Systems and Applications Workshop, July 2006, Bethesda, Maryland, pp: 1-2.

Labib, N.M. and M.N. Malek, 2005. Data mining for cancer management in egypt case study: Childhood acute lymphoblastic leukemia. World Acad. Sci. Eng. Technol., 8: 309-314.

Lee, J.W., 2007. A model for information retrieval agent system based on keywords distribution. Proceedings of the International Conference on Multimedia and Ubiquitous Engineering, April 26-28, Seoul, pp: 413-418.

Leung, Y.Y. and Y.S. Hung, 2009. An integrated approach to feature selection and classification for microarray data with outlier detection. Proceedings of 8th Annual International Conference on Computational Systems Bioinformatics, Aug. 10-12, Stanford, CA. USA., pp: 1-4.

Ling, C., Z. Ling-Jun and T. Li, 2009. Stream data classification using improved fisher discriminate analysis. J. Comput., 4: 208-214.

Mitra, S., S.K. Pal and P. Mitra, 2002. Data mining in soft computing framework: A survey. IEEE. Trans. Neural Networks, 13: 3-14.

Osareh, A. and B. Shadgar, 2009. Classification and diagnostic prediction of cancers using gene microarray data analysis. J. Applied Sci., 9: 459-468.

Pique-Regi, R. and A. Ortega, 2006. Block diagonal linear discriminant analysis with sequential embedded feature selection. Proceedings of the IEEE International Conference May 14-19, Toulouse, pp: 5-5.

Pradhan, G., V. Korimilli, S.C. Satapathy, S. Pattnaik and B. Mitra, 2009. Design of simple ANN (SANN) model for data classification and its performance comparison with FLANN (Functional Link ANN). Int. J. Comput. Sci. Network Security, 9: 105-115.

Ramamohanarao, K., J. Bailey and H. Fan, 2005. Efficient mining of contrast patterns and their applications to classification. Proceedings of the 3rd International Conference on Intelligent Sensing and Information Processing, Dec. 14-17, Bangalore, pp: 39-47.

Ruiz, R., J.C. Riquelme and J.S. Aguilar-Ruiz, 2006. Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recognition, 39: 2383-2392.

Sarhan, A.M., 2009. Cancer classification based on microarray gene expression data using DCT and ANN. J. Theoretical Applied Inform. Technol., 6: 208-216.

Shang, C. and Q. Shen, 2005. Aiding classification of gene expression data with feature selection: A comparative study. Int. J. Comput. Intell. Res., 1: 68-76.

Sheng, L., R. Pique-Regi, S. Asgharzadeh and A. Ortega, 2009. Microarray classification using block diagonal linear discriminant analysis with embedded feature selection. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, April 19-24, Taipei, Taiwan, pp: 1757-1760.

Slavkov, I., S. Dzeroski, J. Struyf and S. Loskovska, 2005. Constrained clustering of gene expression profiles. Proceedings of the Conference on Data Mining and Data Warehouses at the 7th International Multi-Conference on Information Society, Oct. 10-17, Slovenia, pp: 212-215.

Waiyamai, K., C. Songsiri and T. Rakthanmanon, 2004. Object-oriented database mining: Use of object oriented concepts for improving data classification technique. Lecture Notes Comput. Sci., 3036: 303-309.

Wolfram, D., 2000. Applications of informetrics to information retrieval research. Informing Sci., 3: 1-6.

Zhang, L.J., Z.J. Li and X. Hu, 2007. A hybrid gene selection method for cancer classification. Proceedings of VLDB Workshop on Data Mining in Bioinformatics (DMB'07), Vienna, Austria, pp: 1-8.

Zhong, J., Y. Fu and J.L. Zhou, 2006. A classification approach based on evolutionary neural networks. Int. J. Comput. Intell. Res., 2: 72-75.