MACHINE LEARNING BASED APPROACHES FOR CANCER CLASSIFICATION USING GENE EXPRESSION DATA

Amit Bhola[1] and Arvind Kumar Tiwari[2]

[1]Department of CSE, Kashi Institute of Technology, Varanasi, U.P., India
[2]Department of Computer Science and Engineering, IIT (B.H.U.), Varanasi, India

ABSTRACT

The classification of different types of tumor is of great importance in cancer diagnosis and drug discovery. Earlier studies on cancer classification have limited diagnostic ability. The recent development of DNA microarray technology has made monitoring of thousands of gene expression simultaneously. By using this abundance of gene expression data researchers are exploring the possibilities of cancer classification. There are number of methods proposed with good results, but lot of issues still need to be addressed. This paper present an overview of various cancer classification methods and evaluate these proposed methods based on their classification accuracy, computational time and ability to reveal gene information. We have also evaluated and introduced various proposed gene selection method. In this paper, several issues related to cancer classification have also been discussed.

KEYWORDS

Microarray data, Feature Selection, Cancer Classification, Gene Expression data.

1. INTRODUCTION

Cancer research is one of the major research area in medical field. In providing better treatment to patient, it is important to precisely predict different type of tumor. Earlier cancer prediction has always been clinical based and morphological [1]. Systematic approaches based on global gene expression have been proposed, in order to understand the problem of cancer classification. The recent development of microarray technology has motivated the simultaneous monitoring of genes and cancer classification using gene expression data [2, 3, 4, 5, 6]. In its early stage of development, result obtained so far is promising.

It is possible to monitor the expression pattern for large amount of genes simultaneously, hence large amount of gene data has been produced by the development of DNA microarray technology. This technology allows us to analyze the gene data quickly and precisely at one time. The gene expression data is different from any other data as: (1) Gene expression data is very high dimensional, and it usually contains thousands of genes. (2) The publicly available data size is very small or very large (that contains noisy data). (3) Most genes are irrelevant to cancer distinction. The existing classification methods is unable to handle this kind of data effectively. In order to obtain promising results, many researchers proposed to do gene selection before cancer classification. It helps to improve the running time and reduce data size by performing gene selection prior to classification. Gene selection also improves the classification accuracy by removing a large number of irrelevant genes [7]. There are several issues besides gene selection, which are related to cancer. These issues include biological relevance vs statistical relevance of cancer classifiers, the gene contamination problem and asymmetrical classification errors.

DOI : 10.5121/mlaij.2015.2401

2. DNA MICROARRAY AND GENE EXPRESSION

Microarray technology is one of the important recent breakthroughs in experimental molecular biology. This novel technology for thousands of genes concurrently allows the supervising of expression levels in cells and has been increasingly used in cancer research [8,9] to understand more of the molecular variations among tumors so that a more reliable classification becomes possible.

There are two main types of microarray systems [10]: the cDNA microarrays developed in the Brown and Botstein Laboratory at Stanford [11] and the high-density oligonucleotide chips from the Affymetrix Company [12]. The cDNA microarrays are also known as spotted arrays [13], where the probes are mechanically deposited onto modified glass microscope slides using a robotic arrayer. Oligonucleotide chips are synthesized in silico (e.g., via photolithographic synthesis as in Affymetrix GeneChip arrays). For a more detailed introduction and comparison of the biology and technology of the two systems, refer [14]. DNA microarrays provides gene expression data which can be characterized by many measured variables (genes) on only a few observations (experiments), although both the number of experiments and genes per experiment are growing rapidly [15]. The number of genes on a single array is usually in the thousands while the number of experiments is only a few tens or hundreds. There are two different ways to view data: (1) data points as genes, and (2) data points as samples (e.g. patients). In the way (1), the data is presented by expression levels across different samples, thus there will be a large number of features and a small number of samples. In the way (2), the data is represented by expression levels of different genes, thus the case will be a large number of samples with a few attributes. In this thesis, all the discussions and studies on gene expression profiles are based on the first manner of data presentation.

Microarray experiments raise many statistical questions in many diversified research fields, such as image analysis, experimental design, cluster and discriminant analysis, and multiple hypothesis testing [10].

2.1 The Cancer Classification Problem

Classification problem has been broadly studied by researchers in the area of databases, statistics, and machine learning. In the past, many classification algorithms have been proposed, like the linear discrimination analysis, decision tree methods, the bayesian network, etc. The gene expression changes are related to different types of cancers as the researchers have been working on cancer classification using gene expression, for the last few years.

Even though various conventional methods for classification of cancer in clinical practice can be commonly ambiguous or imperfect. Molecular level diagnostics with microarray gene expression profiles is capable of suggesting the methodology of objective, efficient and accurate cancer classification. Hong Hee Won et al., [16] has proposed concept of the collection of network classifiers learned from the negatively correlated characteristics to accurately classify the cancer disease and systematically estimate the performances of the proposed technique on the datasets. Investigational observation reveals the assemble classifier with negatively correlated characteristics that provides the most excellent recognition rate on the datasets.

Hu et al., [17] proposed and analyzed the classification performances of cancer cell by using unsupervised and supervised learning methods. A single hidden layer FFNN with back propagation training is developed and implemented for supervised learning. Non fuzzy, fuzzy and c-means clustering approaches, are used for unsupervised learning.

3. CANCER CLASSIFICATION METHODS

A total of 7 classification algorithms have been used in this comparative study. The classifiers in Weka have been categorized into different groups such as Bayes, Functions, Lazy, Rules, Tree based classifiers etc. A good mix of algorithms have been chosen from these groups that include Naive Bayes (from Bayes), k-Nearest Neighbour, Support Vector Machine, Random Forest, Bagging and AdaBoost. These algorithms are described in following section.

3.1 Naive Bayesian

Naive Bayesian classifier is developed on Bayes Conditional Probability Rule used for performing classification tasks, assuming attributes as statistically autonomous. The word Naive means as strong. Each and every attribute of the data set are considered as independent and strong of each other [18]. Naïve Bayesian is one of the simple and effective classifiers. It uses very few number of parameters that results in low variance making it an effective tool for classification. In Naïve Bayes, the existence of a feature is unrelated to the existence of any other feature thus making it space efficient and fast. The advantage of Naïve Bayesian classifiers is that the training data can be small to predict the parameters for classification. Naïve bayes classifier works well in many real world complex situations that include spam detection, language detection, and sentiment analysis.

3.2 Support Vector Machine

SVMs [19] are supervised learning methods originally used for binary classification and regression. They are the successful application of the kernel idea to large margin classifiers and have proved to be powerful tools. Nowadays SVMs are used in various research and engineering areas ranging from breast cancer diagnosis, recommendation system, database marketing, or detection of protein homologies, to text categorization, or face recognition, etc. The contributions of this dissertation cover the general framework of SVMs. Hence, their applicative scope is potentially very vast.

3.3 K-Nearest Neighbour

K-nearest-neighbor classifier uses the same distance metric. K-NN is a lazy learning or an instance based learning or where the function is approximated locally and all computation is postponed until classification [20]. In this algorithm the final classification is decided by a majority vote of its neighbors. K-NN supports numeric class problems by calculating the average target value of the nearest problems.

This algorithm is one of the highly accurate machine learning algorithms that involves no learning cost and builds a new model for each test. The testing may become costly if the number of instances in the input data set increases.

3.4 Adaboost

Boosting is an approach to machine learning which is based on the idea of making a highly accurate prediction rule by joining many relatively weak and inaccurate rules. The first practical boosting algorithm was AdaBoost algorithm of Freund and Schapire [21], which remains one of the most widely studied and used, with applications in numerous fields. Over the years, a great variety of attempts have been made to “explain” AdaBoost as a learning algorithm, that is, to understand why it works, how it works, and when it works (or fails). It is by understanding the nature of learning at its foundation, both generally and with regard to particular algorithms and phenomena, that the field is able to move forward.

3.5 Random Forest

Random forest [22] is an ensemble classifier which consists of many decision tree and gives class as outputs i.e., the mode of the class's output by individual trees. Random Forests gives many classification trees without pruning. Each classification tree gives a certain number of votes for each class. Among all the trees, the algorithm chooses the classification with the most number of votes. Random forest runs efficiently on large datasets but is comparatively slower than other algorithms. It can effectively estimate missing values and hence is suitable for handling datasets with large number of missing values.

3.6 Bagging

Bagging also known as “bootstrap aggregation”, is one of the simplest methods of arching and the first effective method of ensemble learning [23]. It is meta-algorithm, which helps to improve accuracy and stability, was initially designed for classification and is usually applied to decision tree models, but it can be used with any type of model for regression or classification. This method uses multiple versions of training set by using bootstrap (sampling with replacement). Every data sets is used to train a different model. The outputs of the models are combined by voting (in case of classification) or averaging (in case of regression) to create a single output. It is only effective when using unstable (i.e. a small change in the training set can cause a significant change in the model) nonlinear models.

4. RELATED RESEARCH WORK

It is important to efficiently identify microarray gene expression data because the amount of microarray data is usually very large. The analysis of DNA microarray data is divided into four branches: classification, clustering, gene identification, and modelling and analysis of gene regulatory networks. Many data mining and machine learning methods have been applied to solve them. Fuhrman et al. 2000 [24] have applied Information theory to gene identification problem. Thieffry et al. 1998 [25] have proposed Boolean network, Friedman et al. 2000 [26] have proposed Bayesian network, and Arkin et al. 1997 [27] have applied reverse engineering method to gene regulatory network modeling problem.

In classifying gene expression data various machine learning techniques have been used earlier, which includes Dudoit [10] proposed Fisher linear discriminant analysis, Li [28] have proposed k nearest neighbour, Khan [29] have proposed decision tree, Xu [30] have proposed multi-layer perceptron, Furey [31] have proposed support vector machine, Brown [32] proposed boosting, and Golub [2] have used self-organizing map. Table 1 shows relevant works on cancer classification. Jayashree Dev et al. [33] have focused on three different classification techniques: FLANN , PSO-FLANN and BPN and found that the integrated approach of Functional Link Artificial Neural Network (FLANN) and Particle Swarm Optimization (PSO) could be predict the disease as compared to other method. This proposed method overcomes the nonlinearity of the classification problem. This proposed algorithm could be developed in order to classify different types of cancer genes from huge amount of DNA microarray gene expression data.

Alok Sharma [34] have proposed an algorithm for gene expression data analysis. The algorithm initially divides genes into subsets, into comparatively small size, then selects informative smaller subsets of genes from a subset and combines the chosen genes with another subset to update the gene subset. They repeated this process until all subsets were merged into one informative subset.

They showed promising classification accuracy for all the test datasets. They represents the potency of the proposed algorithm by analyzing three different gene expression datasets.

A. Castano F. [35] have proposed classification technique for microarray gene expression, produced of genomic material by the light reflection analysis. This proposed algorithm was in two-stages, in the first stage, salient expression genes is identified using two filter algorithms from thousands of genes. During the second stage, the proposed methodology was performed by the new input variables are selected from gene subsets. The methodology was composed of a union of (EGRBF) Evolutionary Generalized Radial Basis Function and (LR) Logistic Regression neural networks. The modeling of high-dimensional patterns has shown to be highly accurate in earlier research. Finally, the results obtained were differentiated with nonparametric statistical tests and confirm good unity between LR and EGRBF models.

Chhanda Ray [36] have proposed an algorithm to analyze DNA microarray gene expression patterns for huge amount of DNA microarray data. This development technique was identified based on the collecting various DNA microarray gene expression patterns of the same organism and by monitoring the expression of thousands of genes. In this paper, classification of cancer genes was also focused based on the distribution probability of codes.

Venkateshet [37] have discussed methods to study thousands of genes in a single sample microarray analysis or gene expression profiling. By providing large amount of data, micro array analysis was providing challenges in various fields which could be processed to obtain useful information. This paper focuses on the gene samples obtained from biopsy samples are collected from colon cancer patients. They introduced a learning vector quantization method that determines artifact states and separate infectious genes from regular genes. Finally, organism was identified based on the variations of DNA microarray gene expression patterns.

5. RECENT RESEARCH IN MICROARRAY TUMOR CLASSIFICATION

During the 20th century, biotechnological inventions have resulted in a broad range of approaches for explorations in the functional genomics field. Microarray technology is one of the recent advances which have offered by means of snapshots of which genes are expressed in cells of several tissues and diseases. The methods to obtain the consistent microarray data are continuously being improved and developed to meet the demands of biological researchers.

The recent Study Stable feature selection and classification algorithms [38] proposed by Sebastian and Krzysztof Fujarewicz. In this recent research, they presented that profiles of gene expression shows an alternative for clinical cancer classification. The dimension of obtained data sets is a major problem for classification by applying DNA microarrays. The researchers have proposed a multiclass gene selection method supported on Partial Least Squares (PLS) to select the genes for classification. The novel idea is to solve the multiclass selection problem with the Partial Least Squares method and decompose to a set sub problems of two class, one versus one (OvO) and one versus rest (OvR). This research focused on effective classification of informative genes. As an effect, a new approach to find a small subset of important genes is proposed. The obtainable method allows to find a more reliable classifier with fewer classifier error. An the same time this method produces more stable ordered feature lists in contrast with existing methods.

Table 1. Comparison of various earlier research

Author(s)	**Dataset**	**Methods**		**Accuracy (%)**
		Feature	**Classifier**	
Furey *et al.* (2000)	Leukemi	Signal to Noise ratio	SVM	94.1
	Colon			90.3
Li *et al.* (2000)	Leukemi	Logistic regression		94.1
Li *et al.* (2000)	Lympho	Genetic Algorithm	KNN	84.6
	Colon			94.1
Nguyen *et al.* (2002)	Leukemi	Principal component analysis	Logistic discriminant	94.2
	Lympho			98.1
	Colon			87.1
	Leukemi		Quadratic discriminant analysis	95.4
	Lympho			97.6
	Colon			87.1
	Leukemi	Partial least square	Logistic discriminant	95.9
	Lympho			96.9
	Colon			93.5
	Leukemi		Quadratic discriminant analysis	96.4
	Lympho			97.4
	Colon			91.9
Jayashree Dev et al. (2012)	Breast	Signature composition	BPN	56.12
			FLANN	63.34
			PSO-FLANN	92.36
A. Castano et al. (2011)	Breast	BARS	EGRBF LR	91.08
	CNS			
	Colon			
	Leukemi	FCBF		
	Lung			
	Gcm			
Student, S., & Fujarewicz, K. (2012)	Lung	Partial least square	SVM	95.5
	Leukemi		MSVM	97.5
	Blue cell		LDA	98.0
Alok Sharma and Kuldip K. Paliwal (2012)	Leukemi	Proposed algorithm	Bayesian Classification	96.3
	Lung			100.0
	Breast			100.0

Alok Sharma and Kuldip K. Paliwal proposed an innovative research on gene selection using Bayesian Classification [39]. In this study, they proposed a gene (or feature) selection algorithm by using Bayes classification approach. This algorithm can discover crucial gene subset for cancer classification problem. It begins at an empty feature subset and includes a feature that provides the maximum information to the current subset. The process of including features is terminated when no feature can add information to the current subset. The bayes classifier is used to judge the quality of features. It is considered to be the optimum classifier. The proposed algorithm is carried out on several publically available microarray datasets and better results have been obtained. The gene subset is acquired in the forward selection manner. It is observed that on three DNA microarray gene expression datasets, the proposed algorithm is exhibiting very promising classification performance when compared with several other feature selection techniques.

C. Chandrasekar, P.S. Meena proposed a new concept "Microarray Gene Expression for Cancer Classification by using Fast Extreme Learning Machine with ANP" [40]. DNA microarrays appears to be an efficient tool used in cancer diagnosis and molecular biology. In this recent work, Extreme Learning Machine (ELM) a new learning algorithm is used in order to achieve better consequences of the system accuracy. This algorithm overcomes the difficulties such as inappropriate learning rate, local minima and over fitting usually occurred by the iterative learning techniques and performs the training quickly. The performance of the ELM is improved by using Analytic Network Process (ANP). This technique is anticipated with the help of Lymphoma data set. The proposed technique gives better classification accuracies with lesser training time and implementation of complexity compared to the conventional techniques.

Sujata Dash, Bichitrananda Patra and B.K. Tripathy developed new research on Hybrid classification Data Mining technique used for improving the accuracy of microarray data set [41]. This recent work presented a comparison between dimension reduction technique, Hybrid feature selection scheme and a Partial Least Squares (PLS) method evaluates the comparative performance of four different supervised classification methods such as Multilayer Perceptron Network (MLP), Radial Basis Function Network (RBFN), Support Vector Machine with RBF kernel function (SVM-RBF) and Support Vector Machine by using Polynomial kernel function (Polynomial- SVM). The experimental outcome shows that the PLS regression method is an suitable feature selection method and a collective use of different feature selection and classification approaches makes it achievable to build high performance classification models for microarray data.

6. DATASET

There are various publicly available microarray datasets from cancer gene expression studies, including leukemia cancer, prostate tumor, colon cancer, lymphoma, breast cancer, NCI60, and lung cancer datasets. Among them five datasets (leukemia, prostate, lymphoma, breast, and lung) are used in this paper, mentioned in Table 2. All the cancer datasets have been collected from the repository of Artificial Intelligence Lab, Ljubljana [42].

1) ***Leukemia*** dataset consists of 72 samples: 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloid leukemia (AML). 63 bone marrow samples and 9 peripheral blood samples has been used to measure the source of gene expression. In these 72 samples gene expression levels were measured using high density oligonucleotide microarrays. Each sample contains 5147 gene expression levels.

2) ***Prostate*** dataset consists of 102 samples: 50 samples of normal tissue and 52 samples of prostate tumor. Prostate cancer is most common heterogeneous disease among humans, with respect to highly divergent clinical and histological outcomes. Each sample contains 12533 gene expression levels.

3) ***Breast cancer*** dataset consists of 24 samples: 14 samples of resistant to docetaxel treatment (resistant) and 10 samples of sensitive to docetaxel treatment (sensitive). The tumor response to neoadjuvant treatment was assessed after the samples, which (samples) were obtained before treatment. Each sample contains 12625 gene expression levels.

4) ***Lung cancer*** dataset consists of 34 samples: 17 samples of Squamous cell carcinoma (Squamous), 8 samples of Adenocarcinoma (Adenocarcinoma) and 9 samples of lung tissue (Normal). The tumor response to neoadjuvant treatment was assessed after the samples,

which (samples) were obtained before treatment. Each sample contains 12600 gene expression levels.

Table 2. Datasets

Dataset	**Sample**	**Genes (features)**	**Diagnostic classes**	**Predictive Accuracy (%)**
Leukemia	72	5147	2	98.57
Prostate	102	12533	2	98.8
Breast	24	12625	2	73.33
Lung	34	10541	3	94.07
DLBCL	77	7070	2	90.89

5) ***Lymphoma cancer*** dataset consists of 77 samples: 58 samples of Diffuse large B-cell lymphoma (DLBCL), 19 samples of Follicular lymphoma (FL). Two B-cell lineage malignancies follicular lymphomas (FL) and Diffuse large B-cell lymphomas (DLBCL) have very different clinical presentations, response to therapy and natural histories. However, Follicular lymphoma develop gradually over time in order to obtain the clinical and morphologic property of DLBCLs, some subsets of Diffuse large B-cell lymphoma have chromosomal translocations characteristic of FLs. Each sample contains 7070 gene expression.

7. EVALUATION

We have discussed various approaches for classification, in the preceding section. In this section we examine their performance, on experimental data.

Table 3. Accuracy (%) for Leukemia dataset

Classifier	**IG**	**RelifF**	**SVMRFE**	**PSO**	**FCFB**
NB	97.2	98.6	100.0	97.2	100.0
k-NN	97.2	95.8	90.3	91.7	98.6
RF	97.2	98.6	95.8	97.2	98.6
SVM	97.2	98.6	98.6	98.6	94.4
Bagging	91.7	93.1	93.1	91.7	93.1
AdaBoost	97.2	97.2	97.2	97.2	98.6

The results of accuracy for all the five selected datasets are as shown in Tables 3,4,5,6 and 7. Although the results are different between datasets, Recursive feature elimination is the best, and RelifF is the second among the five chosen feature selection technique. The different type of data causes the difference in performance of datasets. Based on the results, the optimal feature-classifier combination has been produced which gives the best performance on the classification.

Table 4. Accuracy (%) for Prostate dataset

Classifier	**IG**	**RelifF**	**SVMRFE**	**PSO**	**FCFB**
NB	92.2	93.1	92.2	62.7	93.1
k-NN	91.2	92.2	97.1	87.3	94.1
RF	86.3	92.2	95.1	90.2	95.1
SVM	93.1	94.1	95.1	93.1	96.1
Bagging	85.3	86.3	88.2	84.3	87.3
AdaBoost	89.2	92.2	92.2	93.1	98.0

Row is the list of feature selection methods: Information Gain (IG), RelifF, Support vector machine Recursive feature elimination (SVMRFE), Particle swarm optimization (PSO), Fast correlation based feature selection (FCBF).

Table 5. Accuracy (%) for Breast Cancer dataset

Classifier	**IG**	**RelifF**	**SVMRFE**	**PSO**	**FCFB**
NB	94.1	92.1	97.0	96.1	95.1
k-NN	96.1	93.1	97.5	89.7	94.6
RF	91.6	90.1	93.6	88.2	92.6
SVM	68.5	68.5	94.6	92.6	94.1
Bagging	94.1	92.6	92.1	88.7	93.6
AdaBoost	78.3	78.3	77.3	72.4	75.4

Column is the list of classifiers used: Naïve Bayes (NB), k-Nearest Neighbour, Random Forest, Support Vector Machine, Bagging and AdaBoost. The accuracy estimates for the different methods applied to the five data sets.

Table 6. Accuracy (%) for Lung cancer dataset

Classifier	**IG**	**RelifF**	**SVMRFE**	**PSO**	**FCBF**
NB	79.4	85.3	91.2	76.5	94.1
k-NN	79.4	82.4	97.1	70.6	100.0
RF	88.2	88.2	85.3	73.5	97.1
SVM	79.4	79.4	94.1	76.5	85.3
Bagging	94.1	85.3	79.4	55.9	88.2
AdaBoost	91.2	94.1	91.2	73.5	91.2

As we can see, the classification approach Support vector machine performs significantly better than the other approaches on the selected five cancer data set when applied with Recursive feature elimination. Naive Bayes performs better than SVM AdaBoost performs better than other methods on the leukemia and breast cancer data sets. We can also see that k-Nearest Neighbour performs better in case of Lung cancer but with other data SVM performs well with SVMRFE.

Table 7. Accuracy (%) for Lymphoma dataset

Classifier	IG	RelifF	SVMRFE	PSO	FCBF
NB	88.3	89.6	96.1	80.5	96.1
k-NN	97.4	96.1	100.0	85.7	94.8
RF	90.9	93.5	96.1	87.0	96.1
SVM	75.3	75.3	100.0	93.5	98.7
Bagging	87.0	90.9	90.9	87.0	87.0
AdaBoost	92.2	88.3	88.3	93.5	97.4

8. CONCLUSION

Classification problem has been largely studied by researchers in the field of machine learning, statistics, and databases. Various classification technique have been proposed in the past, like the linear discrimination analysis, the bayesian network, the decision tree methods, etc. For the last few year researchers have started exploring cancer classification using gene expression. Recent studies have shown that gene expression changes are related to different types of cancers. Many proposed cancer classification methods are from the machine learning area, statistical and computational, ranging from the old k-NN (nearest neighbour) analysis, to the new SVM (support vector machine). Recent approaches shows that no single classifier that is better than other. Few methods are not extensible to multi-class problems whereas works well on binary-class problems, while others are more flexible and general. Most of these proposed algorithms on gene classification only process to improve the accuracy of the classification and does not notice the running time whereas most gene classifiers proposed are quite computationally expensive. Cancer classification using gene expression data is exceptionally well from the previous classification data. Through this research work, we hope to better understand the problem of cancer classification which helps to develop more systematic and productive classification algorithms.

REFERENCES

[1] Azuaje, F. "Interpretation of genome expression patterns: computational challenges and opportunities." IEEE engineering in medicine and biology magazine: the quarterly magazine of the Engineering in Medicine & Biology Society 19.6 (1999): 119-119.

[2] Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." science 286.5439 (1999): 531-537.

[3] D. Slonim, P. Tamayo, J. Mesirov, T. Golub, and E. Lander. Class prediction and discovery using gene expression data. In Proc. 4th Int. Conf. on Computational Molecular Biology (RECOMB), 2000, pages 263–272.

[4] Lakhani, Sunil R., and Alan Ashworth. "Microarray and histopathological analysis of tumours: the future and the past?." Nature Reviews Cancer 1.2 (2001): 151-157.

[5] Nguyen, Danh V., and David M. Rocke. "Classification of acute leukemia based on DNA microarray gene expressions using partial least squares." Methods of Microarray Data Analysis. Springer US, 2002. 109-124.

[6] Berns, Anton. "Cancer: Gene expression in diagnosis." Nature 403.6769 (2000): 491-492.

[7] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning 46.1-3 (2002): 389-422.

[8] Schneider, Michel, Michael Tognolli, and Amos Bairoch. "The Swiss-Prot protein knowledgebase and ExPASy: providing the plant community with high quality proteomic data and tools." Plant Physiology and Biochemistry 42.12 (2004): 1013-1021.

[9] Apweiler, Rolf, et al. "UniProt: the universal protein knowledgebase." Nucleic acids research 32.suppl 1 (2004): D115-D119.

[10] Dudoit, Sandrine, Jane Fridlyand, and Terence P. Speed. "Comparison of discrimination methods for the classification of tumors using gene expression data." Journal of the American statistical association 97.457 (2002): 77-87.

[11] DeRisi, Joseph L., Vishwanath R. Iyer, and Patrick O. Brown. "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science 278.5338 (1997): 680-686.

[12] Lockhart, David J., et al. "Expression monitoring by hybridization to high-density oligonucleotide arrays." Nature biotechnology 14.13 (1996): 1675-1680.

[13] Miller, Lance D., et al. "Optimal gene expression analysis by microarrays." Cancer cell 2.5 (2002): 353-361.

[14] Harrington, Christina A., Carsten Rosenow, and Jacques Retief. "Monitoring gene expression using DNA microarrays." Current opinion in Microbiology 3.3 (2000): 285-291.

[15] Nguyen, Danh V., and David M. Rocke. "Tumor classification by partial least squares using microarray gene expression data." Bioinformatics 18.1 (2002): 39-50.

[16] Hong-Hee Won; Sung-Bae Cho; “Paired neural network with negatively correlated features for cancer classification in DNA gene expression profiles”, Proceedings of the International Joint Conference on Neural Networks, 2003, Vol. 3, Pp. 1708 – 1713.

[17] Hu, Y.; Ashenayi, K.; Veltri, R.; O'Dowd, G.; Miller, G.; Hurst, R.; Bonner, R.; “A comparison of neural network and fuzzy c-means methods in bladder cancer cell classification”, IEEE International Conference on Neural Networks, IEEE World Congress on Computational Intelligence, 1994, Vol. 6, Pp. 3461 – 3466.

[18] Domingos, Pedro, and Michael Pazzani. "On the optimality of the simple Bayesian classifier under zero-one loss." Machine learning 29.2-3 (1997): 103-130.

[19] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning 46.1-3 (2002): 389-422.

[20] D. Coomans, and D. L. Massart. "Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules." Analytica Chimica Acta136 (1982): 15-27.

[21] Freund, Yoav, and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." Journal of computer and system sciences 55.1 (1997): 119-139.

[22] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[23] Quinlan, J. Ross. "Bagging, boosting, and C4. 5." AAAI/IAAI, Vol. 1. 1996.

[24] Fuhrman, Stefanie, et al. "The application of Shannon entropy in the identification of putative drug targets." Biosystems 55.1 (2000): 5-14.

[25] Thieffry, D. and Thomas, R. (1998): Qualitative analysis of gene networks. Pacific Symposium on Biocomputing, 3:66-76.

[26] Friedman, Nir, et al. "Using Bayesian networks to analyze expression data." Journal of computational biology 7.3-4 (2000): 601-620.

[27] Arkin, Adam, Peidong Shen, and John Ross. "A test case of correlation metric construction of a reaction pathway from measurements." Science 277.5330 (1997): 1275-1279.

[28] Li, Leping, et al. "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method." Bioinformatics 17.12 (2001): 1131-1142.

[29] Khan, Javed, et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." Nature medicine 7.6 (2001): 673-679.

[30] Xu, Yan, et al. "Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett’s esophagus and esophageal cancer." Cancer Research 62.12 (2002): 3493-3497.

[31] Furey, Terrence S., et al. "Support vector machine classification and validation of cancer tissue samples using microarray expression data." Bioinformatics 16.10 (2000): 906-914.

[32] Brown, Michael PS, et al. "Knowledge-based analysis of microarray gene expression data by using support vector machines." Proceedings of the National Academy of Sciences 97.1 (2000): 262-267.

[33] Dev, Jayashree, et al. "A Classification Technique for Microarray Gene Expression Data using PSO-FLANN." International Journal on Computer Science and Engineering 4.9 (2012): 1534.

[34] Sharma, Alok, Seiya Imoto, and Satoru Miyano. "A top-r feature selection algorithm for microarray gene expression data." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 9.3 (2012): 754-764.

[35] Castaño, Adiel, et al. "Neuro-logistic models based on evolutionary generalized radial basis function for the microarray gene expression classification problem." Neural processing letters 34.2 (2011): 117-131.

[36] Ray, Chhanda. "Cancer Identification and Gene Classification using DNA Micro array Gene Expression Patterns." International Journal of Computer Science Issues 8.2 (2011): 155-160.

[37] Venkatesh and Thangaraj, "Investigation of Micro Array Gene Expression Using Linear Vector Quantization for Cancer", International Journal on Computer Science and Engineering, 2010, Vol. 02, No. 06, pp. 2114-2116.

[38] Student, Sebastian, and Krzysztof Fujarewicz. "Stable feature selection and classification algorithms for multiclass microarray data." Biology direct 7.1 (2012): 33.

[39] Sharma, Alokanand, and Kuldip K. Paliwal. "A gene selection algorithm using Bayesian classification approach." American Journal of Applied Sciences 9.1 (2012): 127-131.

[40] C. Chandrasekar, P.S. Meena /International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2, Mar-Apr 2012, pp.229- “Microarray Gene Expression for Cancer Classification by using Fast Extreme Learning Machine with ANP”.

[41] I.J. Information Engineering and Electronic Business, 2012, 2, 43-50 Published Online April 2012 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijieeb.2012.02.07 by Sujata Dash, Bichitrananda Patra and B.K. Tripathy “A Hybrid Data Mining Technique for Improving the Classification Accuracy of Microarray Data Set”.

[42] http://www.biolab.si/supp/bi-cancer/projections/

Authors

Mr. Amit Bhola is currently pursuing M.Tech. (Master of Technology) in CSE (Computer Science and Engineering) from Kashi Institute of Technology, Varanasi, India affiliated to Uttar Pradesh Technical University (UPTU), Lucknow. His research area includes machine learning computational approaches in the field of bioinformatics. He has published a paper in IJCA (International Journal of Computer Applications) and also has a conference paper in IEEE.

Dr. Arvind Kumar Tiwari received the BE degree in CSE from CCS university, Meerut, M.Tech. in CSE from UPTU, Lucknow and Ph.D. in CSE from IIT (BHU), Varanasi, India. He has worked as Professor and Vice Principal in GGS College of Modern Technology, Kharar, Punjab, India. His research interests include computational biology and pattern recognition.