I.J. Information Engineering and Electronic Business, **2013, 5, 42-48**
Published Online November 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2013.05.06

Microarray Gene-expression Data Classification using Less Gene Expressions by Combining Feature Selection Methods and Classifiers

Aarti Bhalla
School of Computer and Systems Sciences, Jawaharlal Nehru University
New Delhi, India
E-mail: aarti.bhalla.5@gmail.com

R. K. Agrawal
School of Computer and Systems Sciences, Jawaharlal Nehru University
New Delhi, India
E-mail: rka@mail.jnu.ac.in

***Abstract*−**Microarray Data, often characterised by high-dimensions and small samples, is used for cancer classification problems that classify the given (tissue) samples as deceased or healthy on the basis of analysis of gene expression profile. The goal of feature selection is to search the most relevant features from thousands of related features of a particular problem domain. The focus of this study is a method that relaxes the maximum accuracy criterion for feature selection and selects the combination of feature selection method and classifier that using small subset of features obtains accuracy not statistically indicatively different than the maximum accuracy. By selecting the classifier employing small number of features along with a good accuracy, the risk of over fitting (bias) is reduced. This has been corroborated empirically using some common attribute selection methods (ReliefF, SVM-RFE, FCBF, and Gain Ratio) and classifiers (3 Nearest Neighbour, Naive Bayes and SVM) applied to 6 different microarray cancer data sets. We use hypothesis testing to compare several configurations and select particular configurations that perform well with small genes on these data sets.

***Index Terms*−**Microarrays, Feature Selection, Hypothesis testing, Classification with less genes.

I. INTRODUCTION

A Microarray is a tool used for gene expression analysis. It comprises of a tiny membrane or glass slide having samples of many regularly arranged genes. Microarray analysis can detect thousands of genes in a small sample along with the expression of those genes. Microarray Datasets are often characterised by high-dimensions and small samples. “These datasets are used for cancer classification problems in human biology”[1]. The problem is to classify the (tissue) samples as deceased or healthy on the basis of analysis of gene expression profile of thousands of genes simultaneously.

In some problems, the task is to classify a given sample among more than two classes such as multiple type of lung cancer. These classification problems are challenging to automate given the high number of genes and contrastingly low number of samples. Such problems suffer from the ‘curse of dimensionality’ [2]. According to Bellman, more dimensionality results in more possibilities and therefore complete enumeration approach either becomes very difficult or impossible. Due to this phenomenon, high dimensionality not only increases the learning cost of the classifier, but also deteriorates its learning performance. Typically such datasets contain several features which are either irrelevant or redundant for a given problem and degrade the performance of the classifier. Besides, high features require a huge size of computer memory. Hence dimension reduction is required to overcome these difficulties.

Mainly there exist two types of dimension reduction techniques, namely, Feature Extraction and Feature Selection. In Feature Extraction high dimension data is mapped to low dimension subspaces either with linear or non linear mapping under some constraints [3]. But feature Extraction techniques are not suitable for microarray data since it would manipulate the features into a new set of features with loss of interpretability. The original features (genes) will lose their physical meaning and hence cause hindrance in the understanding of the biochemical processes responsible for cancer or other diseases. Feature Selection techniques [4], on the other hand, select the most informative features from the original dataset. Thus there is no loss of their physical interpretation. The basic theme of feature selection is to search the most relevant features from thousands of related ones of a particular area. It also helps in enhancing the execution speed and forecasting accuracy of the classifier algorithm. Feature selection (FS) has received great attention in recent years in many domains including classification of Microarray datasets.

Feature Selection techniques are categorized into two types of methods: filter method [3] and wrapper method [5]. The filters compute the information from statistical properties; thus, their results rely upon the information captured from the features. The filters execute quickly, but their results are sub optimal. The wrapper methods use a learning algorithm for feature selection; thus, the result of classification is often biased. The wrappers provide high classification accuracy, but execute with reduced speed. Due to small number of samples in a microarray data, the maximum accuracy criterion is inclined to generate classifiers that are prone to overfit the data set. “Overfitting is a particularly undesirable effect [6], in which, the results obtained from one data set may not generalize to a different data set so the new cases are more likely to be misclassified”. Adjusting the classifier behaviour to the given data increases its risk of overfitting. According to the well known heuristic Occam’s razor, the simplest hypothesis fitting the data is preferred [6]. Hence, among the classifiers with greater accuracy the one employing the smallest number of gene expressions has less risk of overfitting[1]. Besides, having fewer attributes facilitates biomedical interpretation [7].

In this endeavor, we mainly focus on a mechanism of feature selection which explores the possibility of loosening the maximum accuracy criterion by selecting a less accurate classifier, but not to the extent of compromising a good behaviour, in general. A technique called ‘hypothesis testing’ is used to indicate whether the numerical accuracies of two or more classifiers are significantly different. When difference is not significant, it implies that with the help of the given samples one technique cannot be considered as better than the other. For this, the classifier having accuracy statistically equivalent to the best accuracy but with lesser attributes is clearly an appropriate choice for classification of microarray datasets. The soundness of relaxing the criterion of maximum accuracy attained by a classifier is verified empirically on 6 different microarray data sets, joining several feature selection and classifier algorithms on the range from 1 to 200 features. We opted for commonly used feature selection methods and classifiers to do this study, applied hypothesis testing to select specific configurations (FS method and classifier) that behave well with few features.

In the remaining paper, various feature selection techniques in literature, are reviewed in section 2. The feature selection method, under consideration, is then delineated in section 3. In section 4 the datasets and experimental settings are discussed. Section 5 explains the experimental results. Lastly, a brief conclusion is drawn in section 6.

II. FEATURE SELECTION TECHNIQUES

Feature Selection techniques are basically categorized into two types of approaches: “filter method” and “wrapper method”. A Filter method assesses the relevance of a given feature subset using only characteristics of that subset, without the help of any learning method. In literature, various methods based on filters are proposed such as ranking, sequential forward search, backward elimination search [8], incremental approach [9] etc. These methods are simple and involve less computation efforts.

In contrast, a wrapper method assesses the adequacy of a feature subset on the basis of the performance of a given classifier learnt from the training data. Wrapper methods aim to identify a minimal subset of relevant and essential features r out of d features that minimize the classification error such that r < d. For a predefined r, a straightforward approach is to determine a subset of r features from C(d, r) combinations. The floating search [10] and branch and bound (BB) [11];[12] algorithms are two examples of methods used to perform this search. But, it is computationally not feasible for medium and high dimensional datasets.

In some literature [13], a third type of feature selection method, that is, embedded approach is also introduced. In embedded approach, feature Selection procedure is embedded within the learning algorithm itself, judiciously weeding out some subset of features. Examples of such an approach are the CART algorithm [14] and the sparse logistic regression (SLogReg) method [13].

2.1 Feature Selection based on Relevance of features

Recently, some effective filter methods suited for high-dimension data have been devised. These methods fall in mainly two categories [15]: Unsupervised and Supervised. Unsupervised techniques do not use the class names of the training data. Supervised techniques make use of the class names of the training data for measuring the relevance of each feature. The ranking method allocates a score to each feature based on the basis of some pre-defined criterion (denoting statistical properties of feature) and then these are sorted in decending order of their score. Score represents the relevance of each feature in determining the class, thus denoting the discriminatory power of the feature. Next, top m (predefined number) features out of the sorted set are selected denoting the reduced set of features to be considered for classification of samples.

2.1.1 Unsupervised methods

Let D = $\{(x_1 ,c_1),\ldots,(x_n,c_n)\}$ be a training set where xi ∈ R^d, for i = 1,…, n, denotes the feature vector of the ith sample without class label. All features vectors are collected in a n × d matrix X, where the ith row contains xi, the feature vector , while the jth column, denoted Xj ∈ R^n , contains the n samples of the jth feature. Finally, Xij is denoted as the ith sample value of the jth feature.

Unsupervised measures such as Term-Variance (TV) criterion [16] scores each feature on the basis of its variance. The measure based on the ratio of AM

(Arithmetic mean) and GM (Geometric mean) of each feature value and the other dispersion measure, named Mean Median (MM), are proposed by Ferreira and Figueiredo [15].

2.1.2 Supervised methods

Let $D = \{(x_1, c_1), \cdots, (x_n, c_n)\}$ be a training set where $x_i \in R^d$, for i = 1,…, n, denotes the feature vector of the ith sample and ci is its class name.

Fisher ratio (FiR) [17] for binary problems ($c_i \in \{0,1\}$), Mutual Information (MI) [18] are some existing supervised measures to capture the relationship between the gene and the corresponding class label c.

Information Gain Feature Selection Filter: It determines the relevance of a feature by evaluating its information gain with respect to the class. Claude Shannon proposed this measure in information theory, to compute the value or "information content" of messages. It is denoted by:-

"InfoGain (Class, Attribute) = H(Class) - H(Class | Attribute)." (1)

The function H( ) gives the entropy of an attribute.

GainRatio Feature Selection Filter: This filter attempts to overcome this bias of the information gain filter which selects attributes having a large number of outcomes. *Gain ratio* "assesses the worth of a feature by measuring the gain ratio with respect to the class". The feature having the largest gain ratio is chosen as the splitting attribute.

"GainR (Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute)." (2)

ReliefF filter method "selects features that are deemed relevant to the class concept, even though many of them could be highly correlated to each other" [19]. This algorithm "estimates the relevance of features according to how well their values distinguish among the instances of the same and different classes that are near each other "[20]. These type of filters are suited to high-dimensional datasets, in terms of accuracy, time, and memory efficiency.

The feature selection methods discussed so far mainly focused on finding relevant features based on individual evaluation. However, subset of features so obtained may have redundancy thus causing degradation in the classifier's performance. It is pointed out [22] that feature relevance alone is inadequate for efficient feature selection in high-dimensional data. It is imperative to define feature redundancy and perform efficient analysis of feature relevance as well as feature redundancy for optimal feature selection in high dimensional microarray data.

2.2. Feature Selection based on Feature Relevance and Redundancy

Feature subset evaluation implicitly handles feature redundancy with feature relevance. Although these methods [5]; [22] produce better results than methods that do not handle feature redundancy, yet the high computational cost of the subset search makes them inefficient for high-dimensional data. Among existing heuristic search strategies for subset evaluation, even greedy sequential search which reduces the search space from O(2d) to O(d2) can be very inefficient for high-dimensional data. For example, methods of subset evaluation such as forward selection, backward elimination and combination of two determines a minimal feature subset that satisfies some goodness measure and can eliminate both irrelevant features and redundant ones. "Let G be the current set of features. A feature is said to be redundant and hence should be removed from G if it is weakly relevant [21] and has a Markov blanket Mi within G" [20]. In literature [15]; [20], several approximation methods are there, which circumvent subset search by decoupling relevance and redundancy analysis and allow an effective way of finding a feature subset that is nearly close to an optimal feature subset. Correlation Coefficient [22], Maximal Information Compression Index [23] and Absolute Cosine [15] are some of the existing unsupervised measures that have been used in relevance/redundancy analysis for feature selection.

In the category of supervised measures, the Fast Correlation based filter (FCBF) method [20], follows a relevance-redundancy approach in two phases: calculate the Symmetrical uncertainty measure of each attribute and sort these values in descending order; eliminate redundant attributes from the sorted list heuristically using the concept of predominant correlation. The symmetrical uncertainty (SU) for measuring the correlation between two features, is given by:

"SymmU (Class, Attribute) = 2 * (H(Class) - H(Class | Attribute)) / H(Class) + H(Attribute)." (3)

Peng et al.[7] proposed a technique called 'minimum Redundancy Maximum Relevancy' (mRMR) criterion in which redundancy is computed in terms of Mutual Information(MI) "between pairs of features, whereas relevance is measured by the MI between each feature and the class label". These methods are efficient than subset evaluation schemes since redundancy computation is limited to a few pairs of the most relevant features.

Support Vector Machine (SVM) projects the original data to a higher dimension space with the help of a nonlinear kernel function. It then locates a hyper plane having the maximal margin to segregate the data into two classes. Those data points lying closer to the optimal hyper plane act as its Support Vectors, which are used for further classification. SVM combined with Recursive Feature Elimination, [3] is an embedded method. This algorithm discards attributes having less weight (significance) and thus weakly affecting the classification on each cycle. Despite SVM-RFE being more computation intensive in comparison to filter methods, it is deemed to be a quite sound approach for classification of microarray data [24].

III. FEATURE SELECTION USING FEW GENES

Although, using fewer attributes than the configuration with maximum accuracy lessens the exactness of the classifiers, but minimum error rate criterion does not give the optimal feature subset for classification. The small sample size of typical microarray data causes a non-negligible variance in estimating the error rate. Statistically, this implies that there could be more than one configuration having error rates equivalent to the error rate of the best. The differences in the error rates could occur due to chance.

Also, the configuration with maximum accuracy may have a bigger risk of overfitting than that of a less exact classifier. Thus, selecting a configuration that uses fewer genes than the most precise, with its accuracy being significantly equivalent is quite sensible.

Hypothesis testing [25] is used to find those configurations the error rate of which, is not significantly different from that of the best configuration for a particular data set. In this study, to prevent over degradation of the classifier accuracy, we have followed the procedure given below, applicable to each data set:

1. Select the configuration having the best accuracy. Let’s call it *A*.
2. Compare this configuration with others having less number of attributes, by applying corrected resampled t-test with α = 0.05.
3. Discard configurations significantly different than A.
4. Sort the remaining configurations into a list in decreasing order of accuracy.
5. From this list, choose the top 10 configurations.
6. From these 10 configurations, choose the one with least number of attributes.

IV. DATASETS AND EXPERIMENTAL SETTING

4.1. Microarray cancer datasets

For conducting the experiments, 6 different microarray datasets were used, whose characteristics are listed in Table 1. These datasets are publically available for download at “Kent Ridge Biomedical Data Set Repository” [26], “Dataset Repository in ARFF (WEKA) of BioInformatics Research Group” [27], which stores the data with both experimental values and the gene names.

Most of the datasets deal with two class problems, except Lymphoma comprising of 9 distinct classes, MLL-Leukaemia with 3 different classes and GCM with 14 classes. The classification results for multiclass problems are expected to be different than binary class problems.This is because classification of a multiclass problem is more complex than classification of a binary class problem.

In Table 1 it can be seen that the number of samples are varying from 57 to 144 and the number of attributes oscillating from 4026 to 24481, which comprises a wholesome package of data sets to measure the model’s adequacy.

These datasets have been preprocessed for removal of missing values and other discrepancies using appropriate tool before actual processing.

TABLE 1: Data set description.

Data set	Features	Samples	Classes
Breast [27]	24481	78	2
Mll-leukemia [26]	12583	57	3
Lung-M [27]	7129	96	2
GCM [26]	16063	144	14
Lymphoma [27]	4026	96	9
Prostate [27]	12600	102	2

4.2 Experimental setting

Three classification methods are considered: “3 Nearest Neighbour (3-NN)”, “Naïve Bayes (NB)”, and “Support Vector Machines with linear kernel (SVM)”. “SVM is an effective classification method in this domain”[1]. Naïve Bayes method has a simple function. It requires attribute values to be independent given the class. When few features are selected, this condition can be easily accomplished. In this classifier, numeric attributes are modelled by a normal distribution.

The core equation for this classifier is known as ‘Bayes Rule’:

$$\text{“}P\,[Ci|D] = (P\,[D|Ci] * P[Ci]) \,/\, P[D]\text{”} \qquad (4)$$

where, Ci is class i and D is an attribute. We have also chosen 3-NN classifier which uses normalized Euclidean distance for finding the 3 training samples nearest to the given test sample, and forecasts the same class as these training instances. If multiple samples have the same distance to the test sample, the majority class is chosen.Three filter methods are selected: ReliefF, FCBF, Gain Ratio Attribute evaluation and an embedded one, “SVM-RFE”. 12 basic methods are produced by the combination of 4 feature selection methods and 3 classification methods whose acronyms are joined using '+'. Therefore, FCBF + 3NN means that 3NN is applied to data sets filtered with FCBF.

Feature selection methods were utilized for obtaining a specific number of features, covered in the range [1,200]: in steps of 1 in the sub range [1, 5], in steps of 5 in the sub range [5, 40], and in steps of 10 in [40,200]. In this manner, we obtained a total of 336 (3 * 4 * 28) different configurations.

The error rate for each data set has been determined with the "corrected resampled t-test" [25] for hypothesis testing. We applied 50 repetitions of Holdout method, where 90% of instances, selected randomly, are utilized for training and the rest 10% instances for testing. All experiments were performed using the data mining tool Weka [28], all parameters set with default values, but for SVM-RFE, the number of attributes eliminated during iteration are kept at 5%.

V. EXPERIMENTAL RESULTS

For each dataset, we first select the configuration (feature selection method + classifier) that provides

the maximum accuracy over the entire range of features. In case a tie occurs, the configuration with less features is selected. In Table 2, column 'Best', represents that configuration and number of attributes. Thereafter, the procedure listed in section 3 is applied.

The column "10 next best with less attributes", represents 3 out of the top 10 configurations with better accuracy: the configuration having the minimum number of attributes, the configuration having the best accuracy, and the configuration.

having the worst accuracy. In the table, '*' indicates that there are several configurations with the same accuracy.

TABLE 2: Results for each Dataset

Dataset	Best Configuration			10 'next best' with less attributes			
	% Hit	Algorithm	Attrib.		% Hit	Algorithm	Attrib.
I. Breast	87.5	SVM-RFE+NB	5	Min. Attr	75	*	2
				Best %	75	SVM-RFE+SVM	4
				Worst %	67.15	ReliefF+NB	4
II. GCM	73.33	GainR+ 3NN	25	Min. Attr	60	GainR+ 3NN	15
				Best %	66.67	SVM-RFE+SVM	20
				Worst %	60	FCBF+ NB	15
III. Lung-M	100	ReliefF+3NN	2	Min. Attr	98.67	*	
				Best %	98.67	FCBF+NB	1
				Worst %	95.33	SVM-RFE+NB	1
IV. MLL	97.33	*	70	Min. Attr	92.89	ReliefF+3NN	5
				Best %	95.33	ReliefF+NB	40
				Worst %	92.89	GainR+3NN	5
V. Lymphoma	93.77	SVM-RFE+SVM	110	Min. Attr	92.37	SVM-RFE+SVM	35
				Best %	93.91	SVM-RFE+SVM	90
				Worst %	89.02	FCBF+SVM	100
VI. Prostate	95.52	FCBF+SVM	90	Min. Attr	94.79	ReliefF+3NN	10
				Best %	95.45	FCBF+SVM	35
				Worst %	94.24	*	

The ReliefF filter seems to work in an adequate manner for 4 out of the 6 data sets, but it performs poorly when dealing with multiclass datasets, i.e. Lymphoma and GCM. The FCBF filter gives satisfactory results for these datasets. In terms of average, the SVM-RFE model clearly obtains the best result.

By observing these results, one can find accurate classifiers using 5 or less features for two class problems, as indicated by 3 out of 4 binary problems under study. The binary problem which is left is correctly classified with 40 or less features. For Lung-M data set, a classifier is obtained that using just 1 gene expression has accuracy not significantly different from the best. This implies that for the available samples, using extra genes for classification may not necessitate a better performance of the classifier. If more attributes are added, then it can cause a fall in the classifier performance when processing test samples, thus showing signs of overfitting.

While applying feature selection, an important issue is to observe the reduction accomplished in the number of features. "For two-class problem, usually 50 to 60 informative genes are usually enough"[29]. For a multiclass classification problem more genes in the range 100 to 200 may be required. Moreover, a hold out method having 9:1 ratio of training and test set was performed 50 times for estimating the accuracy of each configuration, obtaining 336 different accuracy values(4 filters* 3classifiers *28 no. of attribute values). These values are analyzed and compared among themselves using corrected resampled T testing. MLL-

Leukaemia was the dataset with the smallest number of features for classification (0.07%), whereas Lymphoma was the dataset that utilizes the highest number (1.87%). Except Lymphoma, no data set requires more than 1.5% of the original feature set. The SVM-RFE model also acheives a significant diminution in the number of features required for the classification.

VI. CONCLUSION

Microarray datasets are classified using a combination of feature selection and classifier algorithm. Several studies emphasize on maximising the % accuracy of the induced classifier. However this criterion does not produce an optimal feature subset. The presence of extraneous attributes lead to the risk of overfitting of the classifier on a specific dataset, besides obstructing the biomedical interpretation of selected genes. In this study, a method is presented that relaxes the maximum accuracy criterion while choosing a configuration, selects that configuration employing lesser genes while having accuracy not statistically significantly different from the maximum accuracy attained. Such a configuration with good accuracy and few genes is determined with the help of Hypothesis testing. Empirical results obtained by conducting experiments on 6 different microarray data sets considering 3 classifier methods and 4 feature selection algorithms advocate the soundness of this criterion.

REFERENCES

[1] Alonso-González, C.J., et al., Microarray gene expression classification with few genes: criteria to combine attribute selection and classification methods. Expert Systems with Applications 39, 7270–7280, 2012.

[2] Bellman, R., Adaptive Control Processes. A Guided Tour. Princeton University Press 1961.

[3] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V., Gene selection for cancer classification using support vector machines. Machine Learning, 46(1–3), 389–422, 2002.

[4] Guyon, I., Elisseeff, A., An Introduction to Variable and feature Selection, Journal of Machine Learning Research, 3, 1157-1182, 2003.

[5] Kohavi, R., John, G., H., Wrappers for feature subset selection. Artificial Intelligence, 97 (1-2), 273-324, 1997.

[6] Mitchell, T.M., Machine learning. McGraw-Hill International ed., 1997.

[7] Peng, H., Long, F., Ding, C., Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, 8, 1226-1238, 2005.

[8] Nakariyakul, S., Casasent, D., Adaptive branch and bound algorithm for selecting optimal features. Pattern Recognition Letters, 28, 12, 1415–1427, 2007.

[9] Kittler, J., Pierre, A., D., Pattern Recognition: A Statistical Approach, PHI, 1982.

[10] Ruiz, R., Aguilar, J., S., Riqueline, J., Best agglomerative ranked subset for feature selection.JMLR: Workshop and Conference Proc., 4, New Challenges for feature selection, 148–162, 2009.

[11] Pudil, P., Novovicova, J., Kittler, J., Floating search methods in feature selection. Pattern Recognition Letters, 15, 11, 1119–1125, 1994b.

[12] Chen, X., An improved branch and bound algorithm for feature selection. Pattern Recognition Letters, 24, 1925–1933. , 2003.

[13] Shevade, S., Keerthi, S., A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics, 19, 17, 2246–2253, 2003.

[14] Breiman, L., Friedman, J., Olshen, R., Stone, C., Classification and regression trees, Chap-man &Hall London, 1984.

[15] Ferreira, A., Figueiredo, M., Unsupervised feature selection for sparse data. 19th European Symposium on Artificial Neural Networks-ESANN, Bruges, Belgium, 339–344, 2011.

[16] Liu, L., Kang, J., Yu, J., Wang, Z., A comparative study on unsupervised feature selection methods for text clustering. IEEE International Conference on Natural Language Processing and Knowledge Engineering, 597–601, 2005.

[17] Bishop, C., M., Neural Networks for Pattern Recognition. Oxford University, Oxford, 1995.

[18] Zaffalon, M., Hutter, M., Robust feature selection by mutual information distributions. In proceedings of the 18th international conference on artificial Intelligence, 577–584, 2002.

[19] Kira, K., & Rendell, L.A., A practical approach to feature selection. D. Sleeman, P. Edwards (Eds.), Machine learning: proceedings of international conference (ICML-92), 249–256, 1992.

[20] Yu, L., Liu, H., Efficient feature selection via analysis of relevance and redundancy. J. Machine Learning Res. 5, 1205–1224, 2004.

[21] Koller, D., Sahami, M., Towards optimal feature selection. In Proceedings of the ThirteenthInternational Conference on Machine Learning, 284–292, 1996.

[22] Hall, M., Correlation-based feature selection for discrete and numeric class machine learning. In: ICML Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann, pp. 359–366, 2000.

[23] Mitra, P., Murthy, C., Pal, S., Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Machine Intell. 24, 301–312, 2002.

[24] Saeys, Y., Inza, I., & naga, P. L., A review of feature selection techniques in bioinformatics. Bioinformatics, 23, 2507–2517, 2005.

[25] Nadeau, C., Bengio, Y., Inference for the generalization error. Machine Learning, 52, 239–281, 2003.

[26] Li, J., & Liu, H., Kent Ridge Bio-medical Dataset, <http://datam.i2r.astar. edu.sg/datasets/krbd/>, 2011.

[27] Aguilar-Ruiz, J. S., Dataset Repository in ARFF (WEKA) of BioInformatics Research Group. Pablo de Olavide University and University of Seville. <http:// www.upo.es/eps/aguilar/datasets.html>, 2011.

[28] Witten, I., & Frank, E., Data mining: Practical machine learning tools and techniques (2nd ed.), Morgan Kaufman, 2005.

[29] Golub, T., Stomin, D., & Tamayo, P., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286, 531–537, 1999.

Authors' Profiles

Aarti Bhalla has done B.Tech in Computer Science and Engineering from Guru Gobind Singh Indraprastha University, Delhi. She is currently doing M.Tech from School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi under the supervision of Dr. R.K. Agrawal (see below). Her current area of research is Data Mining and pattern recognition.

Ramesh Kumar Agrawal received M. Tech. degree in computer application from Indian Institute of Technology, Delhi. He has done his Ph.D. in computational physics from Delhi University. Presently, he is working as a professor in School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. His current areas of research are classification, feature extraction and selection for pattern recognition problems in domains of image processing, security and bioinformatics.