# Integrated Classifier: A Tool for Microarray Analysis

Shib Sankar Bhowmick[1,4(✉)], Indrajit Saha[2], Luis Rato[3], and Debotosh Bhattacharjee[4]

[1] Department of Electronics and Communication Engineering, Heritage Institute of Technology, Kolkata, India
shibsankar.ece@gmail.com
[2] Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, India
[3] Department of Informatics, University of Evora, Evora, Portugal
[4] Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

**Abstract.** Microarray technology has been developed and applied in different biological context, especially for the purpose of monitoring the expression levels of thousands of genes simultaneously. In this regard, analysis of such data requires sophisticated computational tools. Hence, we confined ourselves to propose a tool for the analysis of microarray data. For this purpose, a feature selection scheme is integrated with the classical supervised classifiers like Support Vector Machine, $K$-Nearest Neighbor, Decision Tree and Naive Bayes, separately to improve the classification performance, named as Integrated Classifiers. Here feature selection scheme generates bootstrap samples that are used to create diverse and informative features using Principal Component Analysis. Thereafter, such features are multiplied with the original data in order create training and testing data for the classifiers. Final classification results are obtained on test data by computing posterior probability. The performance of the proposed integrated classifiers with respect to their conventional classifiers is demonstrated on 12 microarray datasets. The results show that the integrated classifiers boost the performance up to 25.90% for a dataset, while the average performance gain is 9.74%, over the conventional classifiers. The superiority of the results has also been established through statistical significance test.

**Keywords:** Feature selection · Microarray · Principle component analysis · Supervised classifiers · Statistical significance test

## 1 Introduction

Microarray technology facilitates the researchers to simultaneously measure the expression levels of thousands of genes [1]. Generally, the technology works on

S.S. Bhowmick and I. Saha—Contributed equally.


J.K. Mandal et al. (Eds.): CICBA 2017, Part II, CCIS 776, pp. 30–43, 2017.
DOI: 10.1007/978-981-10-6430-2_3

glass slide, where the DNA molecules are fixed at specific location in an orderly manner [2]. Different technologies are used to fix these DNA molecules. Moreover, the fixed DNA molecules may correspond to the short stretch of an oligonucleotides, representing a gene. Microarray technology helps in understanding and analyzing large number of gene expressions in an efficient manner as well as it assists in exploring the genetic causes of anomalies occurring in a human body. All these analysis using microarray technology creating huge amount of data, analytical precision of which is influenced by a number of variables. Therefore, it is extremely important to reduce these huge data in to an informative one so that the best genes can be distinguished. Such set of genes is differentially expressed in normal and disease samples. To identify these differentially expressed genes, machine learning technology can be used.

Over the last decades, several methods for the integration of classifiers have been developed [3]. One such example of classifier integration is found in [4]. In this approach, evolutionary strategy is used with the integration of Multi-Layer Perception [5] to design a hybrid system for performing classification task. Recently, sequential integration of the classifiers is also proposed, where weights are assigned to the training samples. Based on the weights, samples are then propagated to the subsequent classifier as training data. Adaptive Boosting [6] is an example of such type of integrated classifier. In other approaches, different feature subsets are assigned to each single classifier and latter integration is performed on their results, e.g., mixture of experts [7] and ensemble averaging [8]. Moreover, classifiers are subjected to integrate by various forms of combination along with feature selection while implementing the intelligent decision making process. In this paper, we confined ourselves to this specific domain, referring to the classification problem where it is hard to find a single classifier that can be used for all pattern recognition tasks, since each has its own domain of competence. The above facts motivated us to propose a new technique for constructing Integrated Classifier (IC) that can use aggregated bootstrap samples after Principal Component Analysis (PCA). We expect the IC to exploit the strengths of the base classifier along with feature selection for microarray data to produce the high quality classification results which will overcome the performance of base classifier.

Unlike the other methods, in this study, PCA is used to compute additional features for training the classifier by increasing the diversity in the training set. To train the classifiers, the training dataset is split into different number of rotational non-overlapping subsets. Subsequently, PCA is used for each subset and all the principal components are retained to create diverse and informative features that preserve the variability information of the original training data. Thereafter, such informative features are multiplied with the original data to create the training and testing data for the classifiers. Finally, the posterior probability is computed to get the classification results while testing. In this study, we have used Support Vector Machine (SVM) [9], $K$-Nearest Neighbor ($K$-NN) [10], Decision Tree (DT) [11] and Naive Bayes (NB) [12] as an underlying base classifier to integrate with the above feature selection scheme and named

as individually, $i$SVM, $iK$-NN, $i$DT and $i$NB, all together Integrated classifiers (ICs). The performance of the proposed method is demonstrated in comparison with its Conventional Classifiers (CCs) on 12 microarray datasets [13–15] to see the effectiveness in the classification task. The superiority of proposed ICs are established quantitatively, and visually. Moreover, statistical significance test, called Friedman test [16], is conducted to judge the efficacy of the results produced by ICs.

## 2 Integrated Classifiers

In order to describe the Integrated Classifiers (ICs) some notations are used, such as a training set consisting of $N$ labelled instances $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{N}$ in which each instance $(x_i, y_i)$ is described by $m$ input attributes and an output attribute, i.e., $x \in \mathbb{R}^m$ and $y \in \mathbb{R}$, where $y$ takes a value from the label space $\{c_1, c_2, \ldots, c_z\}$. In a classification task, the goal is to use the information only from $\mathcal{L}$ to construct a classifier which performs well on unseen data. Let $X$ be an $N \times m$ matrix consisting of the values of $m$ input attributes for each training instance and $Y$ be an $N$ dimensional column vector containing the output attributes of each training instance in $\mathcal{L}$, which means that $\mathcal{L}$ can be expressed as concatenating $X$ and $Y$ horizontally, i.e., $\mathcal{L} = [XY]$. Let denote $\mathcal{S} = \{X_1, X_2, \ldots, X_m\}^T$, the attribute set comprised of $m$ input attributes. Note that the parameter $F$ which specifies the number of subsets for the given attribute set $\mathcal{S}$ that should be split off. In order to construct the training set for the classifier $IC$, the following steps are necessary:

**Step1:** Randomly split $\mathcal{S}$ into $F$ number of subsets. The lower and upper bounds of feature subsets are chosen as $F_{min} = 2$ and $F_{max} = \frac{m}{2}$, respectively such that $F_{min} \leqslant F \leqslant F_{max}$, i.e., the minimum number of subsets is 2 with atleast 2 features in each subset.

**Step2:** Repeat the following steps $F$ times for each subset, i.e., $f = 1, 2, \ldots, F$.

- **(a)** A new submatrix $X_f$ is constructed which corresponds to the data matrix $X$.
- **(b)** From this new submatrix, a bootstrap sample $X'_f$ is considered where the sample size is generally smaller than $X_f$.
- **(c)** $X'_f$ is then used for PCA and the coefficients of all computed principal components are stored in a new matrix $D_f$.

**Step3:** Arrange each $D_f$ into a block diagonal sparse matrix $R$ whose $f$th diagonal element is $D_f$, and then rearrange the columns of $R$ so that the order of them correspond to the original attributes in $\mathcal{S}$. During this rearrangement, columns with all zero values are removed from the sparse matrix. The rearranged rotation matrix is denoted by $R^a$ and the training set for classifier $IC$ is $[XR^a, Y]$.

The reason behind to do this rearrangement is that the feature set is split randomly and the order of the attribute or feature subsets is not the same

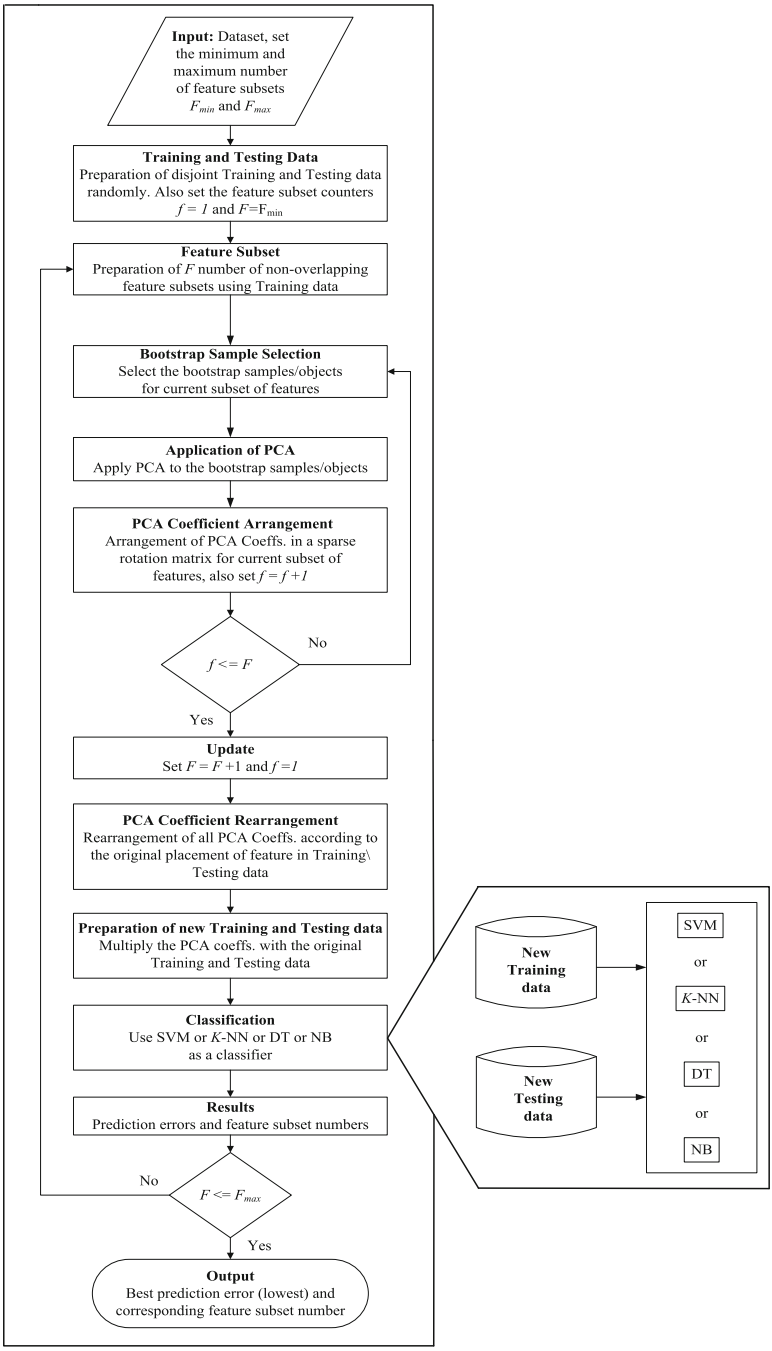


**Fig. 1.** Block diagram of Integrated Classifiers

as original feature set. Thus, to multiply the generated PCA coefficients from the subsets with its corresponding original attributes, we need to rearrange the columns of $R$. In the testing phase of the classification, if $\mathcal{T}$ is the test sample and $IC_i(\mathcal{T}R^a)$ be the posterior probability produced by the classifier $IC$ on the hypothesis that $\mathcal{T}$ belongs to class $c_i$. Then the confidence for a class is determined by the posterior probability. Formally, it can be defined as follows.

$$\psi_i(\mathcal{T}) = IC_i(\mathcal{T}R^a), \quad where \quad i = 1, 2, \ldots, z \tag{1}$$

Here $\mathcal{T}$ is assigned to the class with the largest confidence. Figure 1 shows the block diagram representation of ICs, where SVM, NB, $K$-NN and DT are used separately instate of IC. Due to the process of random feature subdivision, the classifier will get new sets of training and testing data in each iteration, which will help to diversify the classification results. The ICs are applied on microarray datasets to see how it performs on these large attribute datasets.

## 3 Results and Discussion

### 3.1 Microarray Data

In recent years, microarray data have been extensively studied for gene expression analysis in biological and biomedical research. The rapid development of DNA Microarray technology have enabled the simultaneous measurement of the expression levels of thousands of genes. The use of microarrays facilitate the researchers to classify differentially expressed genes between two or more groups of patients. Generally, the expression values of genes are measured at different time points. A microarray gene expression dataset consisting of $G$ genes taken at $T$ time points, can be thought as a $G \times T$ two-dimensional matrix $M = [g_{ij}]$, where each element of $g_{ij}$ represents the expression level of the $i$th gene that has been taken at $j$th time point. To classify the group of genes, here the problem has been modeled as a classification task. Hence, we have applied Integrated Classifiers for microarray data classification. The superiority of the ICs over CCs has been demonstrated on 12 benchmark microarray datasets [13–15]. Details of the considered benchmark microarray datasets [13–15] are given in Table 1, where the first column presents the information about the name of different datasets, the second and third columns give information about microarray types and tissue types. Rest of the columns provide knowledge about size of the dataset, number of classes, samples per class, class name and number of attributes used in each dataset, respectively.

### 3.2 Experimental Setup

In this experiment, the parameters of SVM such as $\gamma$ for kernel function and the soft margin $\mathcal{C}$ (cost parameter), are set to be 0.5 and 2.0, respectively. Note that, RBF (Radial Basis Function) kernel is used here for SVM. The $K$ value for the $K$-NN classifier is chosen as 13 for the satisfactory operation of the classifier and for the case of DT, C4.5 classifier is used.

**Table 1.** Summery of the microarray datasets

| Dataset | Array type | Tissue | Size | Number of classes | Samples per class | Classes | Total number of input attributes |
|---|---|---|---|---|---|---|---|
| Armstrong-2002-v2 | Affymetrix | Blood | 72 | 3 | 24, 20, 28 | ALL, MLL, AML | 2194 |
| Bhattacharjee-2001 | Affymetrix | Lung | 203 | 5 | 139, 17, 6, 21, 20 | AD, NL, SCLC, SQ, COID | 1543 |
| Chowdary-2006 | Affymetrix | Breast, Colon | 104 | 2 | 62, 42 | B, C | 182 |
| Laiho-2007 | Affymetrix | Colon | 37 | 2 | 8, 29 | Serrated CRC, Conventional CRC | 2202 |
| Liang-2005 | Double Channel | Brain | 37 | 3 | 28, 6, 3 | GBM, ODG, Normal | 1411 |
| Nutt-2003-v1 | Affymetrix | Brain | 50 | 4 | 14, 7, 14, 15 | CG, CO, NG, NO | 1377 |
| Pomeroy-2002-v2 | Affymetrix | Brain | 42 | 5 | 10, 10, 10, 4, 8 | MD, Mglio, Rhab, Ncer, PNET | 7129 |
| Ramaswamy-2001 | Affymetrix | Multi-tissue | 190 | 14 | 11, 10, 11, 11, 22, 10, 11, 10, 30, 11, 11, 11, 11, 20 | BR, PR, LU, CR, LY, ML, BL, UT, LE, RE, PA, OV, ME, CNS, | 1369 |
| Risinger-2003 | Double Channel | Endometrium | 42 | 4 | 13, 3, 19, 7 | PS, CC, E, N | 1771 |
| Su-2001 | Affymetrix | Multi-tissue | 174 | 10 | 26, 8, 26, 23, 12, 11, 7, 27, 6, 28 | PR, BL, BR, CO, GA, KI, LI, OV, PA, LU | 1571 |
| West-2001 | Affymetrix | Breast | 49 | 2 | 25, 24 | ER+, ER− | 1198 |
| Yeoh-2002-v2 | Affymetrix | Bone Marrow | 248 | 2 | 43, 205 | T-ALL, B-ALL | 2526 |

### 3.3 Results

The performance of ICs is compared with the CCs like SVM, $K$-NN, C4.5 or DT and NB. As there is no separate training and testing data for the aforementioned datasets, hence, each of these datasets is randomly divided into 70% training and 30% testing datasets to compute the prediction error of each classifier. Tables 2 and 3 report the average results of prediction error produced by different integrated and conventional classifiers for microarray datasets, respectively. Figures 2(a–h) show the results for eight such best performing microarray datasets. In general, the results in Tables 2, 3 and Fig. 2 show that the average prediction error values corresponding to the ICs are better than the CCs. On the other hand, Tables 4 and 5 report the average values of Kappa-Index (KI) [17], Minkowski Score (MS) [18] and Adjusted Rand Index (ARI) [19] of different ICs and CCs for microarray datasets over 20 runs. The KI, MS and ARI values are also found better for ICs. Moreover, it is observed that the results of $i$SVM and SVM are superior in their corresponding groups, whereas the $i$SVM performs better than the SVM. Figures 3(a–h) show the boxplots indicating the

**Table 2.** Average values of prediction error (in %) and its standard deviation ($\sigma$) of different integrated classifiers for microarray datasets

| Dataset | Mean & $\sigma$ of integrated classifier | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $i$SVM | | $iK$ -NN | | $i$DT | | $i$NB | |
| Armstrong-2002-v2 | 00.53 | ±0.16 | 01.91 | ±0.75 | 22.02 | ±3.95 | 06.17 | ±1.81 |
| Bhattacharjee-2001 | 01.78 | ±1.18 | 03.45 | ±2.00 | 11.40 | ±1.21 | 12.35 | ±1.03 |
| Chowdary-2006 | 01.03 | ±1.42 | 13.24 | ±1.84 | 02.79 | ±1.08 | 04.56 | ±2.06 |
| Laiho-2007 | 01.67 | ±1.87 | 02.71 | ±2.54 | 06.25 | ±2.49 | 22.29 | ±3.39 |
| Liang-2005 | 02.80 | ±1.90 | 09.40 | ±6.01 | 18.80 | ±2.93 | 25.80 | ±4.16 |
| Nutt-2003-v1 | 07.42 | ±3.74 | 14.55 | ±7.83 | 24.70 | ±4.11 | 24.70 | ±5.35 |
| Pomeroy-2002-v2 | 04.63 | ±2.90 | 08.33 | ±2.57 | 28.52 | ±3.96 | 25.00 | ±4.79 |
| Ramaswamy-2001 | 30.00 | ±2.57 | 30.08 | ±3.35 | 29.88 | ±2.84 | 26.17 | ±2.77 |
| Risinger-2003 | 08.15 | ±5.29 | 15.19 | ±9.04 | 20.00 | ±3.31 | 22.59 | ±4.94 |
| Su-2001 | 05.31 | ±2.90 | 05.22 | ±3.00 | 29.96 | ±1.52 | 23.81 | ±1.60 |
| West-2001 | 05.00 | ±2.59 | 04.38 | ±2.13 | 13.75 | ±3.11 | 14.06 | ±3.71 |
| Yeoh-2002-v2 | 04.63 | ±2.12 | 19.16 | ±4.22 | 28.52 | ±5.12 | 25.00 | ±4.64 |

**Table 3.** Average values of prediction error (in %) and its standard deviation ($\sigma$) of different conventional classifiers for microarray datasets

| Dataset | Mean & $\sigma$ of conventional classifier | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | SVM | | $K$-NN | | DT | | NB | |
| Armstrong-2002-v2 | 02.02 | ±1.12 | 02.45 | ±1.10 | 21.91 | ±1.62 | 12.13 | ±1.86 |
| Bhattacharjee-2001 | 02.61 | ±1.88 | 04.17 | ±2.38 | 10.49 | ±1.36 | 11.74 | ±1.11 |
| Chowdary-2006 | 02.28 | ±1.95 | 09.49 | ±2.65 | 14.12 | ±1.99 | 05.74 | ±1.49 |
| Laiho-2007 | 07.29 | ±4.68 | 03.13 | ±1.66 | 08.33 | ±1.39 | 21.88 | ±2.29 |
| Liang-2005 | 07.80 | ±2.94 | 10.40 | ±4.32 | 18.00 | ±4.58 | 26.80 | ±5.86 |
| Nutt-2003-v1 | 14.70 | ±5.34 | 15.61 | ±4.51 | 25.91 | ±5.41 | 33.79 | ±5.23 |
| Pomeroy-2002-v2 | 13.33 | ±4.89 | 07.96 | ±1.83 | 30.56 | ±5.56 | 32.78 | ±3.24 |
| Ramaswamy-2001 | 28.99 | ±2.86 | 27.62 | ±5.71 | 31.29 | ±4.73 | 33.71 | ±3.03 |
| Risinger-2003 | 12.78 | ±4.78 | 18.15 | ±4.27 | 19.44 | ±3.07 | 38.52 | ±5.87 |
| Su-2001 | 05.75 | ±3.35 | 05.80 | ±1.32 | 31.48 | ±1.48 | 27.24 | ±3.61 |
| West-2001 | 05.94 | ±2.46 | 11.25 | ±4.74 | 19.38 | ±2.46 | 18.44 | ±6.60 |
| Yeoh-2002-v2 | 07.48 | ±3.35 | 17.73 | ±5.18 | 23.11 | ±2.28 | 31.55 | ±4.17 |

changes of prediction errors with incremental feature subset numbers for the "Armstrong-2002-v2", "Bhattacharjee-2001", "Chowdary-2006", "Laiho-2007", "Liang-2005", "Nutt-2003-v1", "Pomeroy-2002-v2" and "Su-2001" datasets, respectively. The performance of $i$SVM, $iK$-NN, $i$DT and $i$NB for each dataset is shown in four sub figures. The best feature subset number $F$ for each dataset,

**Table 4.** Average values of KI, MS and ARI over 20 runs of different integrated classifiers for microarray datasets

| Dataset | Integrated classifier | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $i$ SVM | | | $iK$ -NN | | | $i$ DT | | | $i$ NB | | |
| | KI | MS | ARI | KI | MS | ARI | KI | MS | ARI | KI | MS | ARI |
| Armstrong-2002-v2 | 0.84 | 0.21 | 0.82 | 0.81 | 0.31 | 0.80 | 0.69 | 0.39 | 0.68 | 0.79 | 0.35 | 0.79 |
| Bhattacharjee-2001 | 0.81 | 0.32 | 0.82 | 0.82 | 0.30 | 0.84 | 0.75 | 0.38 | 0.72 | 0.77 | 0.32 | 0.76 |
| Chowdary-2006 | 0.89 | 0.24 | 0.80 | 0.76 | 0.39 | 0.74 | 0.78 | 0.35 | 0.85 | 0.89 | 0.31 | 0.88 |
| Laiho-2007 | 0.78 | 0.36 | 0.76 | 0.79 | 0.37 | 0.79 | 0.76 | 0.42 | 0.74 | 0.67 | 0.35 | 0.72 |
| Liang-2005 | 0.70 | 0.39 | 0.77 | 0.80 | 0.36 | 0.77 | 0.82 | 0.35 | 0.89 | 0.77 | 0.37 | 0.81 |
| Nutt-2003-v1 | 0.72 | 0.32 | 0.76 | 0.79 | 0.39 | 0.81 | 0.76 | 0.31 | 0.74 | 0.73 | 0.44 | 0.77 |
| Pomeroy-2002-v2 | 0.88 | 0.29 | 0.84 | 0.87 | 0.26 | 0.85 | 0.73 | 0.46 | 0.76 | 0.70 | 0.48 | 0.74 |
| Ramaswamy-2001 | 0.64 | 0.40 | 0.69 | 0.67 | 0.47 | 0.71 | 0.70 | 0.40 | 0.73 | 0.70 | 0.35 | 0.76 |
| Risinger-2003 | 0.72 | 0.29 | 0.76 | 0.73 | 0.31 | 0.78 | 0.74 | 0.41 | 0.82 | 0.71 | 0.46 | 0.73 |
| Su-2001 | 0.77 | 0.26 | 0.75 | 0.73 | 0.40 | 0.70 | 0.73 | 0.44 | 0.79 | 0.69 | 0.55 | 0.71 |
| West-2001 | 0.78 | 0.31 | 0.78 | 0.75 | 0.36 | 0.76 | 0.75 | 0.39 | 0.74 | 0.72 | 0.31 | 0.75 |
| Yeoh-2002-v2 | 0.80 | 0.32 | 0.85 | 0.79 | 0.42 | 0.71 | 0.76 | 0.45 | 0.75 | 0.79 | 0.45 | 0.71 |

**Table 5.** Average values of KI, MS and ARI over 20 runs of different conventional classifiers for microarray datasets

| Dataset | Conventional classifier | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SVM | | | $K$ -NN | | | DT | | | NB | | |
| | KI | MS | ARI | KI | MS | ARI | KI | MS | ARI | KI | MS | ARI |
| Armstrong-2002-v2 | 0.89 | 0.25 | 0.87 | 0.86 | 0.24 | 0.86 | 0.77 | 0.30 | 0.75 | 0.79 | 0.39 | 0.81 |
| Bhattacharjee-2001 | 0.86 | 0.36 | 0.88 | 0.79 | 0.39 | 0.79 | 0.74 | 0.31 | 0.80 | 0.73 | 0.31 | 0.80 |
| Chowdary-2006 | 0.87 | 0.28 | 0.90 | 0.81 | 0.38 | 0.83 | 0.76 | 0.37 | 0.76 | 0.88 | 0.26 | 0.85 |
| Laiho-2007 | 0.78 | 0.42 | 0.71 | 0.82 | 0.38 | 0.80 | 0.73 | 0.31 | 0.78 | 0.72 | 0.40 | 0.75 |
| Liang-2005 | 0.75 | 0.39 | 0.70 | 0.74 | 0.40 | 0.73 | 0.73 | 0.43 | 0.77 | 0.75 | 0.43 | 0.78 |
| Nutt-2003-v1 | 0.71 | 0.37 | 0.75 | 0.73 | 0.39 | 0.75 | 0.67 | 0.43 | 0.63 | 0.63 | 0.41 | 0.64 |
| Pomeroy-2002-v2 | 0.72 | 0.34 | 0.76 | 0.81 | 0.37 | 0.83 | 0.76 | 0.43 | 0.71 | 0.77 | 0.43 | 0.72 |
| Ramaswamy-2001 | 0.76 | 0.33 | 0.70 | 0.76 | 0.45 | 0.79 | 0.68 | 0.41 | 0.71 | 0.65 | 0.46 | 0.65 |
| Risinger-2003 | 0.81 | 0.30 | 0.81 | 0.81 | 0.32 | 0.77 | 0.76 | 0.39 | 0.72 | 0.73 | 0.43 | 0.68 |
| Su-2001 | 0.85 | 0.31 | 0.81 | 0.89 | 0.33 | 0.88 | 0.74 | 0.46 | 0.78 | 0.72 | 0.45 | 0.78 |
| West-2001 | 0.87 | 0.31 | 0.81 | 0.72 | 0.37 | 0.79 | 0.78 | 0.45 | 0.79 | 0.76 | 0.41 | 0.75 |
| Yeoh-2002-v2 | 0.86 | 0.36 | 0.81 | 0.78 | 0.42 | 0.75 | 0.78 | 0.40 | 0.71 | 0.70 | 0.45 | 0.71 |

which are found from these figures, is reported in Table 6. In that table, best feature subset number, corresponding gain value and name of the classifier are also mentioned. The gain is computed according to the Eq. 2:

**Table 6.** Best "$F$" and gain (in %) values of different integrated classifiers for microarray datasets

| Dataset | $\boldsymbol{F}$ | $\mathcal{G}$ (%) | Name of the classifier |
|---|---|---|---|
| Armstrong-2002-v2 | 650 | 06.78 | $i$NB |
| Bhattacharjee-2001 | 322 | 00.86 | $i$SVM |
| Chowdary-2006 | 12 | 13.18 | $i$DT |
| Laiho-2007 | 662 | 06.07 | $i$SVM |
| Liang-2005 | 122 | 05.42 | $i$SVM |
| Nutt-2003-v1 | 308 | 13.73 | $i$NB |
| Pomeroy-2002-v2 | 104 | 11.57 | $i$NB |
| Ramaswamy-2001 | 122 | 11.37 | $i$NB |
| Risinger-2003 | 602 | 25.90 | $i$NB |
| Su-2001 | 587 | 04.73 | $i$NB |
| West-2001 | 524 | 07.75 | $i$$K$-NN |
| Yeoh-2002-v2 | 1042 | 09.57 | $i$NB |
| Summery | | **Avg. $\mathcal{G}$: 09.74 (%)** | **$i$SVM:3 times, $i$$K$-NN:1 times, $i$DT:1 times, $i$NB:7 times** |

**Table 7.** The Friedman ranks of all classifiers for microarray datasets

| Dataset | Integrated classifier | | | | Conventional classifier | | | |
|---|---|---|---|---|---|---|---|---|
| | **$i$SVM** | $i$$K$ -NN | $i$ DT | $i$ NB | SVM | $K$ -NN | DT | NB |
| Armstrong-2002-v2 | 1 | 3 | 8 | 7 | 4 | 5 | 2 | 6 |
| Bhattacharjee-2001 | 1 | 3 | 6 | 5 | 2 | 4 | 8 | 7 |
| Chowdary-2006 | 1 | 7 | 3 | 8 | 2 | 6 | 4 | 5 |
| Laiho-2007 | 1 | 2 | 4 | 6 | 5 | 3 | 8 | 7 |
| Liang-2005 | 1 | 3 | 6 | 5 | 2 | 4 | 7 | 8 |
| Nutt-2003-v1 | 1 | 2 | 5 | 6 | 3 | 4 | 7 | 8 |
| Pomeroy-2002-v2 | 1 | 3 | 5 | 6 | 4 | 2 | 7 | 8 |
| Ramaswamy-2001 | 4 | 5 | 3 | 6 | 2 | 1 | 7 | 8 |
| Risinger-2003 | 1 | 3 | 6 | 5 | 2 | 4 | 8 | 7 |
| Su-2001 | 2 | 1 | 7 | 8 | 3 | 4 | 5 | 6 |
| West-2001 | 2 | 1 | 5 | 8 | 3 | 4 | 6 | 7 |
| Yeoh-2002-v2 | 1 | 4 | 6 | 5 | 2 | 3 | 8 | 7 |
| Average rank | **1.417** | **3.083** | **5.333** | **6.250** | **2.833** | **3.667** | **6.417** | **7.000** |

$$\mathcal{G} = \left(\frac{PA\,of\,IC - PA\,of\,CC}{PA\,of\,CC}\right) \times 100 \tag{2}$$

Here predicted error is used to compute the Prediction Accuracy (PA) for gain computation. From Table 6, it can be seen that the best produced gain is

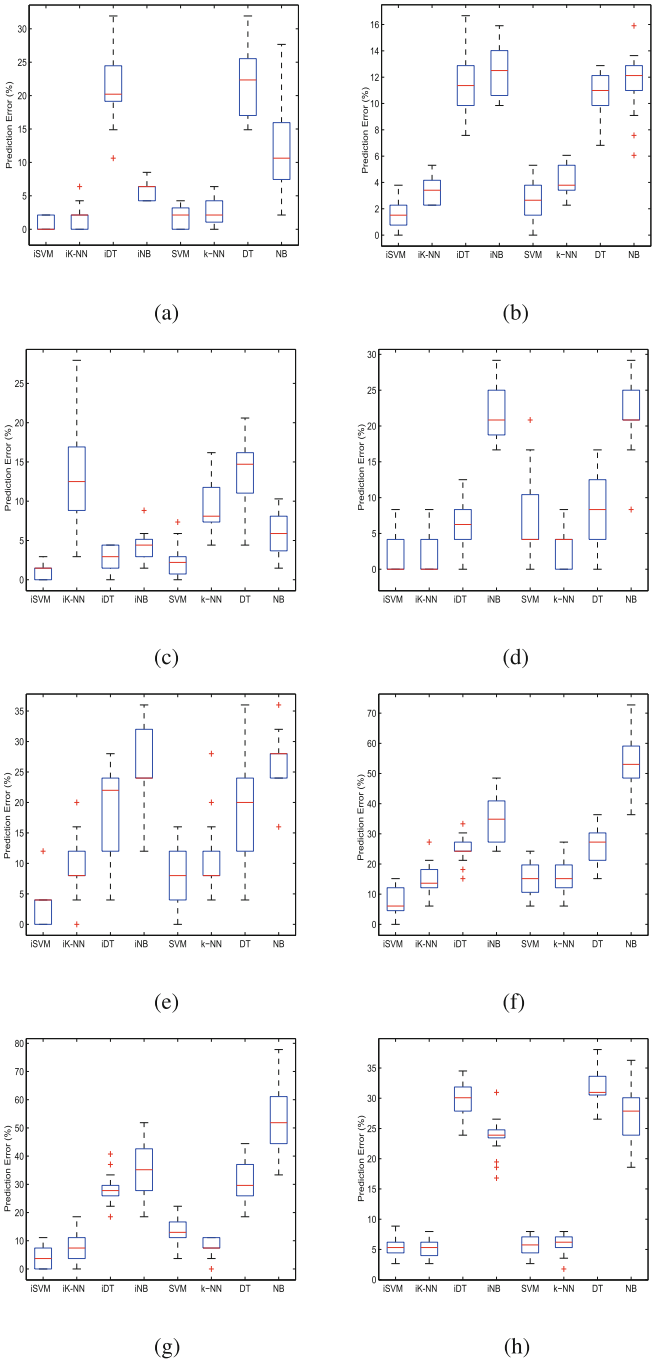


**Fig. 2.** Boxplot representation of prediction errors of different classification algorithms on (a) Armstrong-2002-v2 (b) Bhattacharjee-2001 (c) Chowdary-2006 (d) Laiho-2007 (e) Liang-2005 (f) Nutt-2003-v1 (g) Pomeroy-2002-v2 and (h) Su-2001 datasets

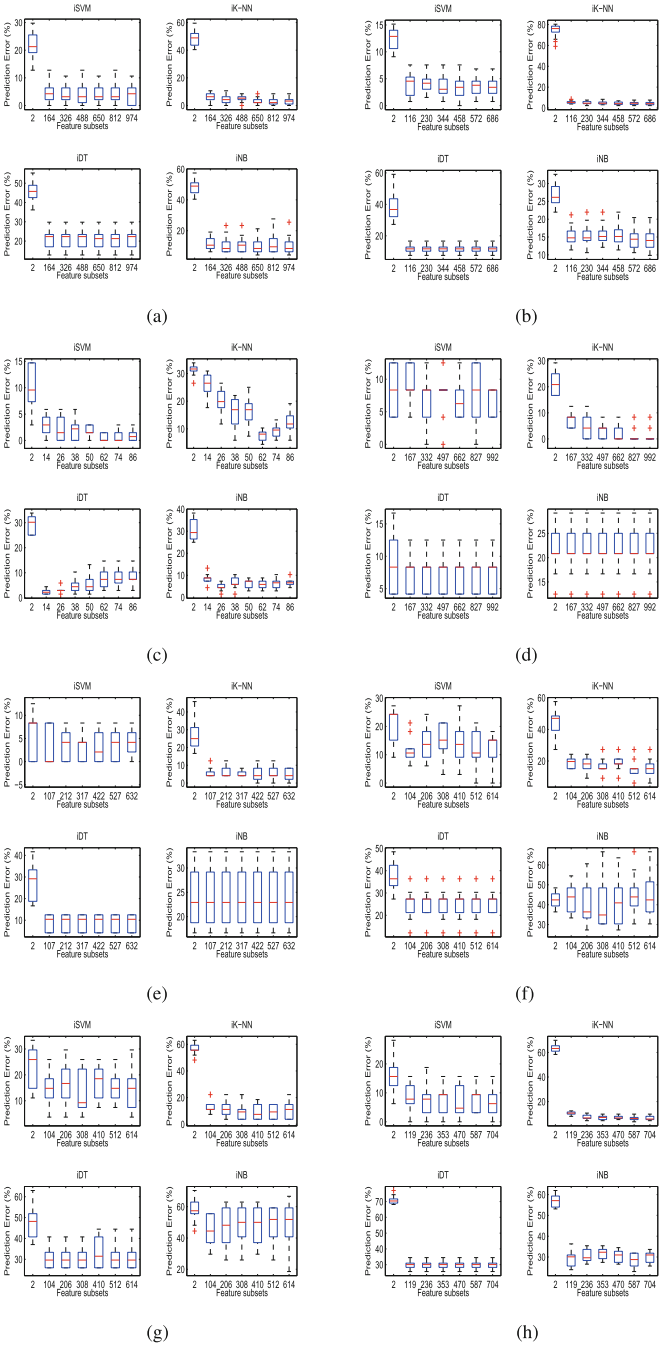


**Fig. 3.** Boxplot representation of the changes in prediction errors with feature subset numbers of different integrated classification algorithms on (a) Armstrong-2002-v2 (b) Bhattacharjee-2001 (c) Chowdary-2006 (d) Laiho-2007 (e) Liang-2005 (f) Nutt-2003-v1 (g) Pomeroy-2002-v2 and (h) Su-2001 datasets

25.90% for "Risinger-2003" dataset and the average of the best gain is 9.74%. It gives a better understanding about the superiority of the ICs over CCs. Moreover, the best gain produced by $i$SVM, $iK$-NN, $i$DT and $i$NB are 3, 1, 1 and 7 times respectively, which also reveal the same fact for ICs. Therefore, it indicates the superior performance of ICs for proper classification of microarray data.

Statistical test like Friedman test has been conducted for the used classifiers and the rank of these classifiers are reported in Table 7. The rank is determined based on the average prediction error values produced by the ICs and CCs. From Friedman test, the average rank of the classifiers, $i$SVM, $iK$-NN, $i$DT and $i$NB, is computed as 1.417, 3.083, 5.333 and 6.250. Based on average rank, the *chi-square* value: 60.861 and $p$ value: $0.13 \times 10^{-4}$ at $\alpha = 0.05$ significance level is obtained. This is also a strong evidence to reject the null hypothesis. Therefore, the results produced by the ICs are statistically significant.

## 4 Conclusion

Microarray expression analysis generates millions of data related to the biological interpretation of genes and their functions. However, sophisticated computational methods are required in order to successfully analyze these microarray data. In this regard, the developed method shows promising results. The present study can be viewed as a comparative analysis of integrated and conventional classifiers where 12 microarray datasets are used. The integrated classifier is developed based on feature selection scheme. In feature selection scheme, bootstrap samples are used to create diverse and informative set features using principal component analysis. Thereafter, such features are multiplied with the original data to construct the training and testing data for the Support Vector Machine, $K$-Nearest Neighbor, Decision Tree and Naive Bayes classifiers separately. Finally, the posterior probability is computed for each classifier to get the classification result. For microarray datasets, the values of prediction errors, Kappa-Index, Minkowski Score, Adjusted Rand Index as well as the statistical significant test, indicate the superior performance of integrated classifiers. Moreover, the gain produced by integrated classifiers over conventional classifiers has also verified the goodness of this integration. Therefore, judging all the results, it can be concluded that the proposed integrated classifiers are quantitatively, visually and statistically superior to their conventional counterparts for microarray data analysis.

The application of the proposed method could be beneficial for binding activity prediction of protein-peptide [20,21]. Additionally, the developed method can also be used for miRNA marker [22–24] and gene selection [25].

## References


1. DeRisi, J.L., Iyer, V.R., Brown, P.O.: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278**(5338), 680–686 (1997)
2. Stears, R.L., Martinsky, T., Schena, M., et al.: Trends in microarray analysis. Nat. Med. **9**(1), 140–145 (2003)
3. Valentini, G., Masulli, F.: Ensembles of learning machines. In: Marinaro, M., Tagliaferri, R. (eds.) WIRN 2002. LNCS, vol. 2486, pp. 3–20. Springer, Heidelberg (2002). doi:10.1007/3-540-45808-5_1
4. Mitra, S., Mitra, P., Pal, S.K.: Evolutionary modular design of rough knowledge-based network using fuzzy attributes. Neurocomputing **36**, 45–66 (2001)
5. Khotanzad, A., Chung, C.: Application of multi-layer perceptron neural networks to vision problems. Neural Comput. Appl. **7**(3), 249–259 (1998)
6. Freund, Y., Schapire, R.E.: A desicion-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995). doi:10.1007/3-540-59119-2_166
7. Jordan, M.I., Jacobs, R.A.: Hierarchical mixture of experts and the EM algorithm. Neural Comput. **6**, 181–214 (1994)
8. Hashem, S.: Optimal linear combination of neural networks. Neural Comput. **10**, 519–614 (1997)
9. Boser, B.E., Guyon, I.M., N.Vapnik, V.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152 (1992)
10. Sun, S.: Ensembles of feature subspaces for object detection. In: Yu, W., He, H., Zhang, N. (eds.) ISNN 2009. LNCS, vol. 5552, pp. 996–1004. Springer, Heidelberg (2009). doi:10.1007/978-3-642-01510-6_113
11. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco (1993)
12. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
13. Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., Korsmeyer, S.J.: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat. Genet. **30**(1), 41–47 (2002)
14. Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J., Meyerson, M.: Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl Acad. Sci. **98**(24), 13790–13795 (2001)
15. Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., Yu, J., Wang, Y., Mazumder, A.: Prognostic gene expression signatures can be measured in tissues collected in rnalater preservative. J. Mol. Diagn. **8**(1), 31–39 (2006)
16. Friedman, M.: A comparison of alternative tests of significance for the problem of m rankings. Ann. Math. Stat. **11**, 86–92 (1940)
17. Cohen, J.A.: Coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
18. Jardine, N., Sibson, R.: Mathematical Taxonomy. Wiley, New Jersey (1971)
19. Yeung, K.Y., Ruzzo, W.L.: An empirical study on principal component analysis for clustering gene expression data. Bioinformatics **17**, 763–774 (2001)

20. Saha, I., Rak, B., Bhowmick, S.S., Maulik, U., Bhattacharjee, D., Koch, U., Lazniewski, M., Plewczynski, D.: Binding activity prediction of cyclin-dependent inhibitors. J. Chem. Inf. Model. **55**(7), 1469–1482 (2015)
21. Mazzocco, G., Bhowmick, S.S., Saha, I., Maulik, U., Bhattacharjee, D., Plewczynski, D.: MaER: a new ensemble based multiclass classifier for binding activity prediction of HLA Class II proteins. In: Kryszkiewicz, M., Bandyopadhyay, S., Rybinski, H., Pal, S.K. (eds.) PReMI 2015. LNCS, vol. 9124, pp. 462–471. Springer, Cham (2015). doi:10.1007/978-3-319-19941-2_44
22. Bhowmick, S.S., Saha, I., Maulik, U., Bhattacharjee, D.: Identification of miRNA signature using next-generation sequencing data of prostate cancer. In: Proceedings of the 3rd International Conference on Recent Advances in Information Technology, pp. 528–533 (2016)
23. Lancucki, A., Saha, I., Bhowmick, S.S., Maulik, U., Lipinski, P.: A new evolutionary microRNA marker selection using next-generation sequencing data. In: 2016 IEEE Congress on Evolutionary Computation (CEC), pp. 2752–2759 (2016)
24. Saha, I., Bhowmick, S.S., Geraci, F., Pellegrini, M., Bhattacharjee, D., Maulik, U., Plewczynski, D.: Analysis of next-generation sequencing data of mirna for the prediction of breast cancer. In: Panigrahi, B.K., Suganthan, P.N., Das, S., Satapathy, S.C. (eds.) SEMCCO 2015. LNCS, vol. 9873, pp. 116–127. Springer, Cham (2016). doi:10.1007/978-3-319-48959-9_11
25. Bhowmick, S.S., Saha, I., Maulik, U., Bhattacharjee, D.: Biomarker identification using next generation sequencing data of RNA. In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 299–303 (2016)