Feature Ranking and Scoring of Gene Expression Data Using Associative Pattern Mining

[1]Prerna Sethi, [2]Alan Edward Alex, [3]Sathya Alagiriswamy
[1] Department of Health Information Management and Biological Sciences, [2]Computer Science Program, [3]Department of Biomedical Engineering, Louisiana Tech University
Ruston LA 71272, USA
{prerna, aalex, sal042}@latech.edu

Abstract. Exponentially growing gene expression data sets present a challenge to the understanding of biologically significant cellular mechanisms and to bioinformatics, in general. Feature (gene) selection is a crucial step in the classification of gene expression profiles. In this paper, we present a novel schema for gene filtering and ranking based on the schema's power to predictively classify samples into functional categories by applying three statistical impurity measures on a cancer dataset. Varied numbers of top-ranking genes, ranging from 50-400, are extracted and filtered based on their class presence and redundancy across feature ranking methods. Association dependency rules are discovered between these selected genes. We then present a scoring scheme for the featured genes by assigning weights based on the genes' degree of participation in the association rules. The featured genes are evaluated by comparing the classification accuracy using machine learning classifiers. The experimental results show a boost in the specificity and sensitivity by the proposed feature ranking and scoring schema.

Keywords: Feature Selection, Association Rule Mining, Classification, Gene Expression.

1 Introduction

Microarray analysis can yield novel information regarding cellular mechanisms, regulatory functions of genes, functions of genes and proteins, gene network structure, and pathways, and can relate the risk of being affected by diseases to gene expression profiles which are characteristics for phenomenon such as cancer [1,2]. With its thousands of uncharacterized variables, microarray data analysis presents one of the most daunting challenges facing bioinformatics. However, this technology can considerably contribute to the understanding of biologically significant cellular mechanisms. The latter objective can be achieved by identifying frequently occurring sets of marker genes, which are critical in tumor classification using gene expression data. A gene marker is a specific, unique sequence of DNA that can be used to identify the location on the chromosome. An approach to narrow the search for a gene marker is to select a set of features (discriminatory genes) based on some statistical or machine learning measure, which can distinguish between different types of samples according to their gene expression values. Another approach can be clustering performed on the genes or samples to identify clusters of genes that have similar gene expression patterns or clusters of samples that have similar expression profiles that can provide insight into therapeutic and pathogenic studies [1,3]. These approaches are usually combined with known classification schemes. Irrespective of the approach used, the challenge associated with the analysis of gene expression data is the large number of genes per sample, of which only a few genes may attribute to the development of a practically usable classifier. Hence, it is critical to select a set of discriminatory genes to boost the accuracy of the classification systems [4,5,6].

This paper compares three feature selection heuristics based on the statistical impurity measures, the Gini Index, Max Minority [7], and the Twoing Rule [8]. After the top 50-400 features are selected by these methods, the selected features are subjected to association rule discovery (ARD) [9] to find correlated pairs of genes occurring together. We also choose common sets of features across all three statistical measures to discover sets of correlated genes. Frequent-1, Frequent-2, and Frequent-3 sets of genes are discovered, and the genes are scored based on frequency of occurrence. The selected features are then studied for effectiveness by comparing the precision, recall, and F-measure of the classification algorithms, first with the selected features, then with all features.

1.1 Related Work

Classification is a well-studied discipline where the purpose is to build an efficient model that maps data items into several predefined categories. A model, consisting of data items and their class labels, is built using training data. The incoming data is filtered through this model and is expected to correctly predict the class labels on the training data as well as the incoming test data. Feature (gene) selection is an important preprocessing step to boost the specificity and sensitivity of the classifier, as it involves selecting a set of relevant genes, which behave as discriminatory features capable of distinguishing different types of samples according to their gene expression values.

In the past, several studies in cancer genomics have applied classification models to predict tumor samples. A survey on the methods for cancer classification comparing classification accuracy, computation time, and biologically relevant information is presented in [10]. The importance of feature selection methods in cancer classification is described in [11], while [12] presents a comparative study of various feature selection heuristics using two datasets. These papers involve both gene selection and classification of microarray data. However, few of these classification methods involve gene selection by assigning weights as an *apriori* step to improve the specificity and sensitivity of the classifier. This paper extends our previous work of feature selection and classification [13] by presenting a novel schema for gene filtering and ranking based on the schema's predictive power to classify samples into functional categories by applying three statistical impurity measures on a cancer dataset and further scoring the featured genes by assigning weights based on each gene's degree of participation in the association rules [14, 15]. The featured genes are evaluated by comparing the classification accuracy using various machine learning classifiers.

1.2 Motivation

Analysis of gene expression data for cancer classification can provide valuable information for early diagnosis and treatment. The computational extraction of derived patterns from microarray gene expression is a non-trivial task that involves sophisticated algorithm design and analysis for specific domain discovery. Moreover, the extraction of biologically significant knowledge from the gene expression data is a growing computational challenge, as the large number of genes, which can correspond to different time sequences or tissue types, has a dimensionality that is several orders of magnitude more than the evaluated samples. Moreover, many of these genes may not be significant for distinction, compared to the training samples. Feature extraction based approaches for classification lead to the extraction of a set of discriminatory genes that promise precise, accurate, and functionally robust analysis of gene expression data. Hence, our primary motivation for this work is to perform feature extraction by exploiting associative dependencies among the genes. The weights assigned to the selected features provide an additional boost to the classifier. Several experiments are conducted to show the efficacy of our proposed method.

The rest of the paper is organized as follows. In Section 2, we describe the methodology in detail; Section 3 presents the results of our experiments along with a comparison of feature extraction methods and classifiers. In Section 4, we present our conclusions.

2 Methodology

The overall methodology is illustrated in Figure 1. The framework consists of the following major computational steps: (1) data preprocessing, which involves standardization and normalization, (2) feature selection, which involves three statistical measures for gene ranking and selection, (3) association rule mining on the selective features to obtain the weights for the frequently occurring genes, (4) classifier building by giving additional boost to the selective features, (5) evaluation, which involves performance comparison of the classification schema using the test datasets and success metrics.

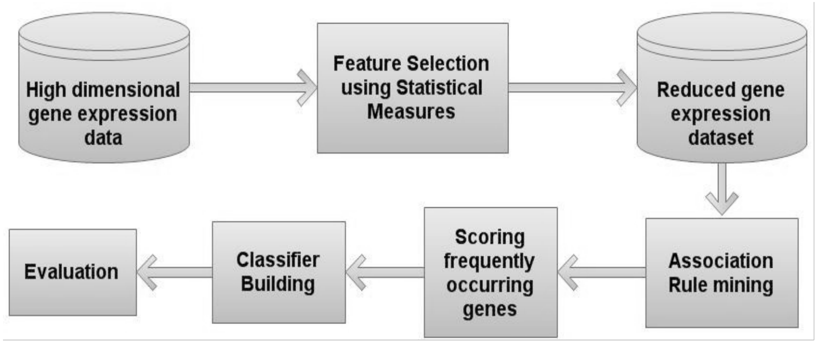

High dimensional gene expression data
Feature Selection using Statistical Measures
Reduced gene expression dataset
Association Rule mining
Scoring frequently occurring genes
Classifier Building
Evaluation

Fig. 1. Framework for feature selection and classification.

2.1 Target Dataset

Our target dataset is a collection of gene expressions values consisting of 7,129 genes from 72 leukemia samples reported by Golub et al. [2]. The data set is divided into an initial training set of 27 samples of Acute Lymphoblastic Leukemia (ALL) and 11 samples of Acute Myeloblastic Leukemia (AML) from bone marrow specimens. The testing data set is an independent set of 20 ALL and 14 AML samples, monitored under different experimental conditions and includes 24 bone marrow and 10 blood sample specimens. The datasets are available for download at http://www.genome.wi.mit.edu/MPR.

2.2 Preprocessing

The dataset is preprocessed using the techniques of standardization, normalization, and discretization. Normalization is performed using the z-score method, which transforms the features with mean 0 and standard deviation as 1. This process also standardizes the data. Finally, all the expression values are normalized in the range [0,1].

2.3 Feature Selection and Scoring

The number of features is large compared to the small number of samples in the dataset. The program Rankgene [17] is used to rank the features in the dataset. The measures included in Rankgene have been widely used in machine learning or statistical learning theory. We use statistical impurity based measures, Gini Index (GI), Max Minority (MM) and the Twoing rule (TR) to extract the relevant features. These measures quantify the effectiveness of the feature by evaluating the predictability of a class by dividing the full range of the expression of a given gene into the two intervals of up-regulation and down-regulation. The prediction is based on the presence of all the samples belonging to a particular interval in the same class. We selected the Top-50, Top-100, Top-200, and Top-400 ranked genes from each of the three statistical measures, which formed our reduced feature datasets. We further selected the Top-50, Top- 100, Top-200, and Top-400 common genes across the three feature selection measures to observe whether a combined set of features from all three measures gives us a better performance in terms of classification accuracy. If a particular gene is highly ranked, then the other genes which are correlated with this gene are also likely to have high ranks [18]. We utilize the advantage of this correlation among the highly ranked genes by performing Association Rule Discovery (ARD) to find frequently occurring sets of genes. ARD was first introduced in [9] and has the following definition.

Let *I* be the set of items and *D* be the set of transactions. Each transaction *T* in *D* contains a set of items such as $T \subseteq I$. Association rules follow the form *X* => *Y*, where $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \emptyset$. X is called the *antecedent* (left hand side or LHS), and *Y* the *consequent* (right hand side or RHS) of the rule. The meaning of the rule X => Y is that data instances that contain X are likely to contain Y as well. To select the interestingness of the rules various measures of significance and interest can be applied, including *support* and *confidence*. The *support* of the rule is the percentage of transactions that contain both *X* and *Y*. The *confidence* of the rule is the conditional probability of Y given X, P(Y|X). The purpose of association rule mining is to find all the rules, which exceed the user specified threshold of support and confidence.

ARD is performed on each of these reduced feature datasets separately to find frequently occurring sets of genes. The frequently occurring genes establish patterns between them of the form $Gene_x \Rightarrow Gene_y$, which implies that when $Gene_x$ occurs, it is likely that $Gene_y$ also occurs. The Frequent-1, Frequent-2, and Frequent-3 patterns are discovered for all the sub-datasets.

The scores for the frequently occurring genes are obtained in the following manner. Let F_k be the set that contains *k* items occurring together. In our case *k* is [1,3]. Let $G \subseteq F$ such that $G = \{G_1, G_2, ... G_p\}$ be the featured genes that form the frequently occurring itemsets with a support score, s_i associated with them and let s_{ij} be the number of times genes, $G_i G_j$ occur together in all the samples. Hence, $\forall G_i \exists s_i$ such that,

$$F_1 = \{G_1(s_1), G_2(s_2), ...G_p(s_p)\} \tag{1}$$

$$F_2 = \{G_1G_2(s_{12}), G_1G_3(s_{13}), G_1G_4(s_{14})..........G_pG_n(s_{pn})\} \tag{2}$$

$$F_3 = \{G_1G_2G_3(s_{123}), G_1G_3G_4(s_{134}), ...G_{pnm}(s_{pnm)})\} \tag{3}$$

We observe that in F_1 , each gene has a support score which describes the number of times that gene has occurred in all the samples. However, in F_2 a gene G_j , which frequently occurs with gene G_k , may frequently occur with gene G_l , as well. Similarly, in F_3 a gene G_j , which frequently occurs with, $G_k G_l$ may frequently occur with $G_k G_m$, as well. Hence, we define the revised support score, s'_j for each gene G_j in F_2 and F_3 where s' is the average of the support scores of itemsets where G_j forms a frequent itemset with other genes. To elucidate further, consider G_1 in (2) where s' for G_1 will be the average of s_{12}, s_{13}, and s_{14}. Similarly, s' for G_1 in (3) will be the average of s_{123} and s_{134}. The genes occurring in F_3 and, consequently, in the F_2 itemset are more important than the genes, which can occur in F_2 and F_1 or F_1 alone, as it forms all possible combinations with other frequently occurring genes. Hence, we need to boost these genes for the classification process by assigning higher weights to the genes that occur in F_3, and then to F_2, and F_1, respectively. The weight W_x, for each gene G_k, in F_1, F_2, and F_3 is calculated using the following formula.

$W_x = \sum_{k=1}^{3} k(s_k)$, where, k is the number of itemsets depending on whether the gene belongs to the F_1, F_2, or F_3 itemset. These weights are then normalized using z-score normalization so that they fall in the range of [0,1].

2.4 Classification

A boost is provided to the selected features by multiplying them with the inverse of the weights in an effort to make them discriminatory, which can consequently increase the accuracy of the classification process. We train and test our new reduced feature sets using three well-known classifiers: AdaBoost, Naïve Bayes, and C4.5 [20]. A brief description of each of these classifiers follows.

Naïve Bayes is one of the most successful learning algorithms for text categorization. It is based on the Bayes rule of assuming conditional-independence between classes. Based on the rule and using joint probabilities of sample observations and classes, the algorithm attempts to estimate the conditional probabilities of classes given an observation.

Adaptive Boosting (AdaBoost) is a meta algorithm that starts with one classifier and adds new classifiers. AdaBoost can be used with other algorithms to increase algorithm performance.

C4.5 is a widely used decision tree-based classifier. Pruned trees and subtree raising techniques are used in our experiments.

The results of each of these methods are detailed in Section 3. To compare the classification accuracy of various classifiers, we calculate the precision, recall, and F-measure as follows.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

In this case, TP = samples which are classified correctly belonging to a class; FP = samples incorrectly labeled as belonging to the class, and FN= samples which should be labeled as belonging to a class, but are not.

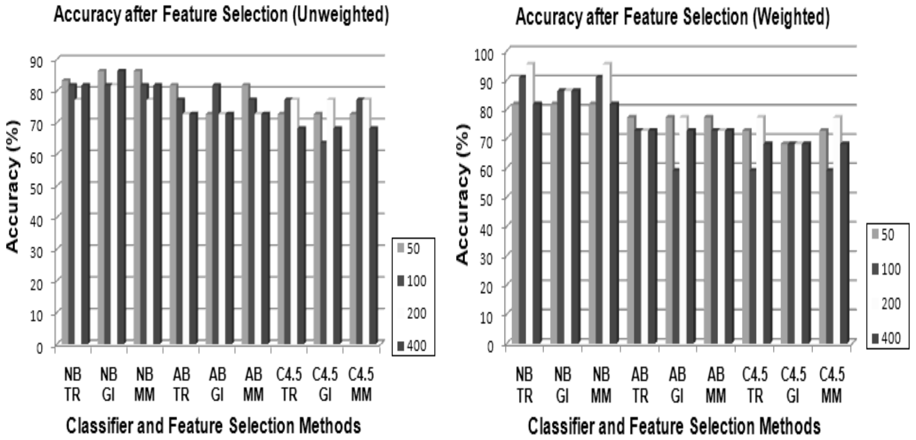
$$\text{F measure} = \frac{2(\Pr ecision * \mathrm{Re} call)}{\Pr ecision + \mathrm{Re} call}.$$

3 Results

We perform several experiments to evaluate the efficacy of the selected features by using our novel association rule based weighting scheme with an expectation to boost the accuracy of the classification methods. The sub-datasets formed from ranking measures are compared using different classifiers to observe the effectiveness of the selected features. The gene ranking algorithm is run using the Queen Bee LONI supercomputer [21], and all the other experiments are carried out on 3.2GHz Intel® Pentium® 4 processor with 1GB RAM.

3.1 Classification Accuracy of the Selected Classifiers with and without Weighted Scoring

In this experiment, we choose the top ranked genes based on the three feature selection measures TR, GI and MM and report the classification accuracies obtained with and without using the weighted scoring scheme. Classification, based on the three classifiers: Naïve Bayes, AdaBoost, and C4.5, is performed by applying a 70/30 split (70% training and 30% testing). It can be observed from Figure 2 that the Naïve Bayes classifier performs best among the three classifiers. Using just the top ranked genes without assigning any weights (unweighted), the Naïve Bayes classifier achieves a maximum classification accuracy of 86.36%, while our proposed weighted scoring scheme gives a classification accuracy of 91.91% and 95.45% using the top 100 and top 200 ranked genes.



(a) (b)

Fig. 2. Classification accuracies for the top ranked genes using the three classifiers (Naïve Bayes, AdaBoost, and C4.5) with and without the weighted scoring scheme.

We consider the performance of our system with all possible combinations of the three gene ranking measures with the top 50, 100, 200, and 400 genes for the three classifiers discussed above. The AdaBoost classifier gives a classification accuracy of 77.27% for the top 50 ranked genes selected using all three feature selection methods using our proposed scoring scheme whereas a classification accuracy of 81.81% is achieved for the top 100 ranked genes selected using GI and for the top 50 ranked genes selected using TR and MM without the assignment of weights. C4.5 performs the worst, obtaining a classification accuracy of 77% with the top 200 ranked genes selected using the TR and MM feature selection methods after the assignment of weights while the unweighted scheme achieves a maximum classification accuracy of 77% for the top 100 ranked genes selected using TR and MM and for the top 200 ranked genes selected using all the feature selection methods. It is evident from the results that feature selection methods used to select discriminatory features can be exploited to improve classification accuracy and speed.

3.2 Feature Selection and Weighted Scoring using Association Rule Mining

In this experiment, three feature selection methods, TR, GI and MM are used to rank the genes. The top 50, 100, 200, and 400 ranked genes are selected for further analysis. ARD is performed on the 12 sub-datasets of top ranked genes separately to find frequently occurring sets of genes. The support and confidence measures are set to 70% and 90%, respectively, for all sub-datasets in order to generate rules. Our experiment shows that a number of rules have common LHS but different RHS. We limit our selection to those rules that have only genes present on the LHS of the rule to identify a set of non-overlapping genes. Hence, a smaller number of genes qualify for the set support and confidence threshold in all sub-datasets. Table 1 shows the number of unique genes obtained after the ARD process has been completed for each sub-dataset.

Table 1. Number of unique genes obtained after ARD in all the sub-datasets.

Statistical Measure	Top 50	Top 100	Top 200	Top 400
Twoing Rule	8	18	24	42
Gini Index	9	17	27	43
Max Minority	8	18	24	42

The scoring method as described in Section 2.3 is used to obtain the scores for each gene in all sub-datasets. The scores are then normalized to the range [0, 1]. An example showing the scores obtained for the GI measure for the top 50 genes is shown in Table 2. Similarly, this scoring method is performed for the all sets of top ranked genes obtained using the feature selection methods. The normalized scores in the range [0, 1] are then used as weights to enhance the accuracy of the classifier by dividing the expression value of the gene present in the reduced feature set by the normalized score.

Table 2. Scores calculated for the set of nine genes forming the reduced feature set using the top 50 ranked genes selected based on Gini Index.

Name of Gene	F1 (%)	F2 (%)*2	F3 (%)*3	Scores	Normalized Scores
D88422_at	11.405	56.399	240.239	308.043	1
X62320_at	10.997	33.405	59.760	104.164	0.313
M27891_at	10.386	43.383	0	53.770	0.143
M63379_at	13.034	22.559	0	35.594	0.082
M19507_at	10.997	11.062	0	22.060	0.0372
M84526_at	10.386	11.062	0	21.449	0.035
M96326_rna1_at	10.386	11.062	0	21.449	0.035
M33195_at	11.405	0	0	11.405	0.001
M83667_rna1_s_at	10.997	0	0	10.997	0

The Precision and recall rates are noted for each of the classifiers and are averaged for two classes. The F-measure, which is the harmonic mean of precision and recall, is calculated for each of the above selection strategies and reported in Figure 3 for the top 50 genes.

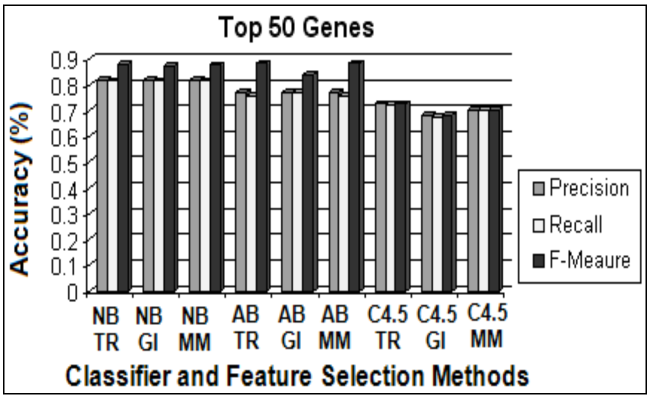


Fig. 3. Precision, recall, and F-measures calculated for the top 50 ranked genes obtained using the three feature selection methods (TR- Twoing Rule, GI- Gini Index, and MM-Max Minority) and the three classifiers (NB-Naïve Bayes, AB-AdaBoost, and C4.5).

The precision, recall, and F-measure for the top 100, 200, and 400 genes are shown in Figures 4, 5, and 6, respectively.

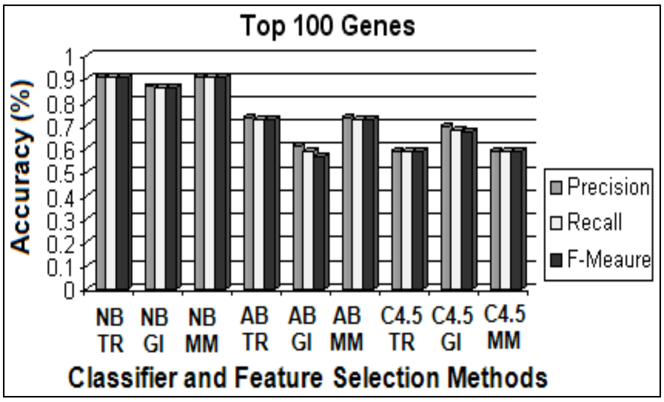


Fig. 4. Precision, recall, and F-measures calculated for the top 100 ranked genes obtained using the three feature selection methods (TR- Twoing Rule, GI- Gini Index, and MM-Max Minority) and the three classifiers (NB-Naïve Bayes, AB-AdaBoost, and C4.5).

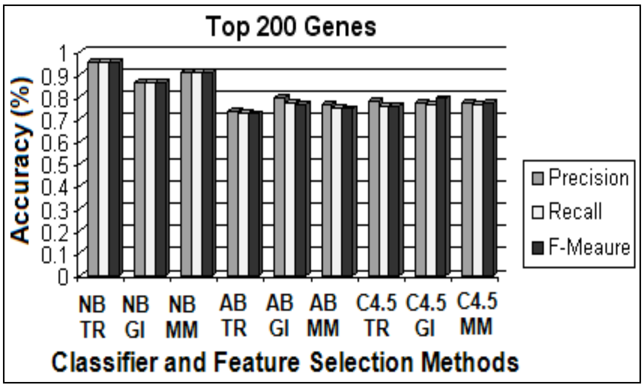


Fig. 5. Precision, recall, and F-measures calculated for the top 200 ranked genes obtained using the three feature selection methods (TR- Twoing Rule, GI- Gini Index, and MM-Max Minority) and the three classifiers (NB-Naïve Bayes, AB-AdaBoost, and C4.5).

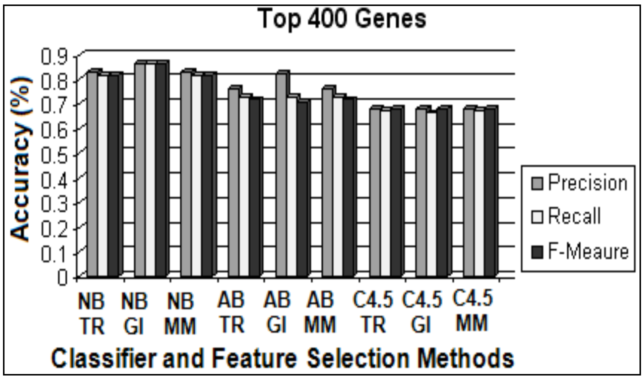


Fig. 6. Precision, recall, and F-measures calculated for set of top 400 ranked genes obtained and the three feature selection methods (TR- Twoing Rule, GI- Gini Index, and MM-Max Minority) using the three classifiers (NB-Naïve Bayes, AB-AdaBoost, and C4.5).

3.3 Classification Accuracy with Common Genes Across All Measures

This experiment is an extension of a previous experiment in which we scored and selected a set of genes from the top 50, 100, 200, and 400 sets of genes using the gene ranking measures and performing ARD on them. We then found the sets of common genes across all three gene ranking measures and the corresponding ranked genes in the top 50, 100, 200, and 400 sets. We observed that there were eight common genes in different sub-datasets of the top 50 genes and 16, 24, and 38 common genes in sub-datasets of 100, 200, and 400 genes, respectively. Figure 7 shows the classification accuracy of the common genes using the three classifiers. There is, however, no appreciable difference in the accuracy here as all the three gene ranking measures had similar ranks for the top 50, 100, 200, and 400 genes.

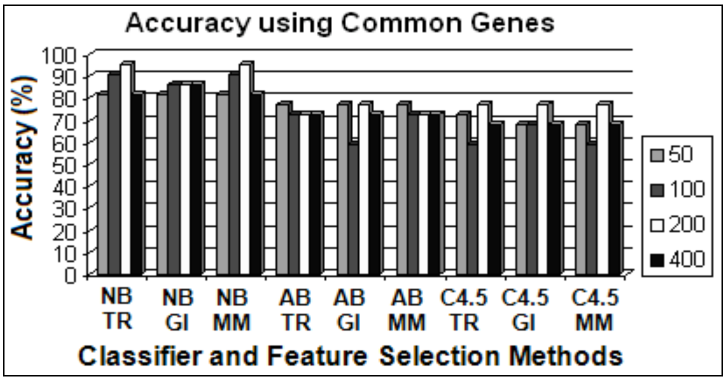


Fig. 7. Classification accuracy for the 12 sets of top ranked genes (50, 100, 200, and 400) obtained using the three feature selection methods (TR- Twoing Rule, GI- Gini Index, and MM-Max Minority) using the three classifiers (NB-Naïve Bayes, AB- AdaBoost, and C4.5).

4 Conclusion

This paper introduces a novel approach based on ARD to assign weights to the featured genes and to subsequently boost the classification accuracy. Gene selection improves class prediction more by the proposed weighted scoring scheme, which provides a boost to the selected set of genes for classification. The accuracy of the Naïve Bayes classifier for the top ranked genes using our proposed method outperforms the accuracy of itself for the top ranked genes where no without additional boost is provided. This result implies that boosting the feature selection effectively reduces the insignificant dimensions and noise to improve classification accuracy. The gene ranking measures rank genes based on their predictive power to classify the genes into samples. We observed that the common sets of genes obtained from the top 50-400 genes did not show appreciable difference in classification accuracy. This observation highlighted the correlation between the top ranked genes, and we exploited this attribute to find the associations between them and to assign scores. We expect that this method can be effectively extended to build multiclass classifiers for the analysis of gene expression data.

References

1. Beer, D.G., et al.: Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat. Med. 8:816-824 (2002).
2. Golub T.R., et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286: 531-537 (1999).
3. Ramaswamy, S. et al.: Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl Acad. Sci. USA. 98:15149–15154 (2001).
4. Alon U., et al.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA. 96:6745–6750 (1999).
5. Ben-Dor, A. et al.: Tissue classification with gene expression profiles. J. Comp. Biol. 7:559–583 (2000).
6. Furey, T.S. et al.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinfor. 16:906–914 (2000).
7. Breiman L., et.al.: Classification and Regression Trees, CRC Press, (1984).
8. Murthy S. K., et.al.: A system for induction of oblique decision trees. J. Art. Intell. Res. 2:1–33 (1994).
9. Agrawal, R.; Imielinski, T.; Swami, A.: Mining Association Rules between Sets of Items in Large Databases. SIGMOD Conference, 207-216 (1993).
10. Y.C. Tang et. al .: A Hybrid CI-Based Knowledge Discovery System on Microarray Gene Expression Data, IEEE CIBCB Conference, 25-30 (2005).
11. Dudoit S., Fridlyand J. and Speed T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc., 97, 77–87 (2002).
12. Liu H., et.al.: A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Inform., 13:51–60 (2002).
13. Sethi P.; Gene Selection through Association Rule Filtering for Supervised Classification. BIOT, 5:65-72 (2008).
14. Sethi P., Leangsuksun C.: A Novel Computational Framework for Fast Distributed Computing and Knowledge Integration for Microarray Gene Expression Data Analysis. IEEE AINA, 2:613-617 (2006).
15. Sethi P. Leangsuksun C.: Fast Knowledge Integration in Gene Expression Databases using High-performance Parallel Computing. EITC, 2:164-165 (2006).
16. Simon, R., et al. Gene-expression profiles in hereditary breast cancer (2001).
17. Su,Y., Murali,T.M., Pavlovic,V. and Kasif,S.: RankGene: identification of diagnostic genes based on expression data. Bioinformatics, 1578–1579 (2003).
18. Hanczar B., et.al.: Improving classification of microarray data using prototype-based feature selection. SIGKDD Explor. Newslett., 5, 23–30 (2003).
19. A. Grove and D. Schuurmans.: Boosting in the Limit: Maximizing the Margin of Learned Ensembles. Conf. Artif. Intell., 15:692–699 (1998).
20. The top ten algorithms in data mining. CRC press, 2009.
21. Louisiana Optical Network Initiative. http://www.loni.org.

Acknowledgements

This research project was made possible by National Institutes of Health Grant Number P20 RR16456 from the INBRE Program of the National Center for Research resources and the supercomputing resources provided by the Louisiana Optical Network Initiative.