Accurate Classification of SAGE Data Based on Frequent Patterns of Gene Expression

George Tzanis, Ioannis Vlahavas
Department of Informatics, Aristotle university of Thessaloniki
{gtzanis, vlahavas}@csd.auth.gr

Abstract

In this paper we present a method for classifying accurately SAGE (Serial Analysis of Gene Expression) data. The high dimensionality of the data, namely the large number of features, in combination with the small number of samples poses a great challenge and demands more accurate and robust algorithms for classification. The prediction accuracy of the up to now proposed approaches is moderate. In our approach we exploit the associations among the expressions of genes in order to construct more accurate classifiers. For validating the effectiveness of our approach we experimented with two real datasets using numerous feature selection and classification algorithms. The results have shown that our approach improves significantly the classification accuracy, which reaches 99%.

1. Introduction

The vast amounts of biological data that have been accumulated the last years due to the rapid progress in the field of biology have posed new questions and new demands. However, the parallel progress in the field of computer science compensates for these needs and assists the efficient management and analysis of these data. In particular, the fields of data mining and machine learning provide biologists a powerful set of tools to analyze these data fast, accurately and reliably.

Proteins are the main structural and functional units of an organism's cell. DNA and RNA, have the role to carry the genetic information of the organisms. In particular, the genetic information that is coded in the genes of DNA is transcribed into messenger (mRNA) and then is translated into a protein. The functions of an organism depend on the abundance of proteins which is partly determined by the levels of mRNA which in turn are determined by the the expression of the corresponding gene. Changes in gene expression underlie many biological phenomena.

The study of gene expression levels may guide to very important findings. One of the basic aims of gene expression data mining is to discover differences between the gene expression profiles of diseased and healthy tissues and use this knowledge to predict the health state of new samples. SAGE (Serial Analysis of Gene Expression) is a method that provides the quantitative and simultaneous analysis of the whole gene function of a cell [12]. The method works by counting short *tags* of all the mRNA transcripts of a cell. The set of all tag counts in a single sample is called a SAGE *library*, and describes the gene expression profile of the sample.

Some recent efforts have utilized data mining methods for analyzing SAGE data. Decision trees (C4.5) and support vector machines were used in [3] to classify the data according to cell state (normal or cancerous) and tissue type (colon, brain, ovary, etc.). Hierarchical clustering of SAGE libraries was also studied [3, 9]. In [11] hierarchical and partitional (*K*-Means) clustering algorithms as well as various cluster validation criteria were studied. Other approaches have also been applied on SAGE data, including mining of strong emerging patterns [10], association rules [2], and frequent closed itemsets [5]. The effect of dimensionality reduction methods was studied in [1]. Data cleaning was considered in [8] as well as the process of the attribution of a tag to a gene. Finally, various feature ranking, classification, and error estimation methods were presented in [7].

The small number of studies of SAGE data classification and the moderate classification accuracy of the so far proposed approaches motivated our work and guided us to an effort to define a better approach that improves prediction accuracy. An important advantage of the SAGE method is that the experimenter does not have to select the mRNA sequences that will be counted in a sample. This is quite important, since the appropriate sequences for studying various diseases such as cancer are not usually known in advance. This advantage of SAGE makes it a fairly promising method, especially for cancer studies as in this paper.

Our contribution is a new approach that uses frequent pattern mining for discovering any associations among the expressions of genes that can assist the construction of more accurate classifiers. As shown by the experimental results, our approach improves notably the prediction accuracy. The paper is outlined as follows. In the next section provide a detailed description of our approach. In section three, we present the datasets that were used and define our experimental setup. Then, in section four we

present our results and finally, in section five we conclude.

2. Our Approach

In this section we provide a detailed description of the proposed approach. Before presenting the basic steps of this approach (discretization, frequent pattern mining, feature selection, and classification) we will describe the structure of the input data.

The data are structured in a gene expression matrix A. The columns of the matrix represent the tags of the genes and the rows represent the different samples (SAGE libraries). The intersection of the i^{th} row with the j^{th} column, namely the element a_{ij}, is the gene expression level for the gene j in the sample i. A sample i is associated with a class label c_i. In our setup $c_i \in \{0, 1\}$, where 0 denotes the normal cell state and 1 denotes the cancerous cell state.

2.1 Discretization

The discretization procedure followed in our approach is important for two basic reasons:

1. For detecting the strong *under-expressions* (expressions of genes that are significantly lower than the mean gene expression) or *over-expressions* (expressions of genes that are significantly higher than the mean gene expression) of genes.
2. For transforming the data in a binary context, so that a frequent pattern mining algorithm can be applied.

The discretization process works as follows. First, the initial data matrix A is divided into two new matrices A_0 and A_1 that contain the samples of the normal and cancerous cell states respectively. Then, for each matrix we calculate a 99% confidence interval for the expression levels of each gene. So, for each gene j we get two confidence intervals [left(j,0), right(j,0)] and [left(j,1), right(j,1)] for the normal and cancerous cell states respectively. Finally, we create two new matrices A_0' and A_1', where $a'_{0,ij}, a'_{1,ij} \in \{-1, 0, 1\}$. These values are assigned as follows:

- $a'_{c_i,ij} = -1$, if $a_{ij} <$ left(j,c_i) AND $a_{ij} \neq 0$
- $a'_{c_i,ij} = 0$, if $a_{ij} \in$ [left(j,c_i), right(j,c_i)] OR $a_{ij} = 0$
- $a'_{c_i,ij} = +1$, if $a_{ij} >$ right(j,c_i)

Assigning the value of -1 to $a'_{c_i,ij}$, means that gene j is significantly under-expressed in the sample i with respect to the expression levels of this gene in class c_i. Similarly, an assignment of +1 to $a'_{c_i,ij}$, means that gene j is significantly over-expressed in the sample i with respect to the expression levels of this gene in class c_i. A value of zero assigned to $a'_{c_i,ij}$ means that there is not a significant under-expression or over-expression.

The term "$a_{ij} \neq 0$" is used in order not to consider zero values as under-expressions. The rationale behind this is that a zero value means that a tag is not found in a sample, so the corresponding gene is not just under-expressed, but it is not expressed at all. According to biochemists, the vast majority of genes in the human genome are only expressed in one tissue type, and only some "housekeeping genes" are expressed in all cells [9]. In line with this consideration, it is very probable that a gene with zero expression level in a particular sample is never expressed in the tissue type from which the sample was taken. So it would be inaccurate if we considered it as an under-expression.

Matrices A_0' and A_1' are the input to the next step that involves mining for frequent gene expression patterns.

2.2 Frequent pattern mining

In this step we use the discretized gene expression matrices in order to find frequent patterns for each class (cell state) separately. We may use any of the known frequent pattern mining algorithms (i.e. FPGrowth [6]). After applying the frequent pattern mining algorithm we get two sets of frequent patterns F_0 and F_1 for matrices A_0' and A_1' respectively. Then, we create a new set of frequent patterns F, so that $F = F_0 \cup F_1 - F_0 \cap F_1$. F contains only the patterns that are frequent only in one of the two classes. The rationale behind this is that the patterns that are frequent in both classes do not provide adequate information for discriminating samples of different classes.

The set of patters F is the new set of features that will be used to describe the initial data. This means that the initial gene tags will be substituted by patterns of frequent gene under-expressions and/or over-expressions. This is done by creating a new data matrix A'. The columns of the matrix are the patterns in F and the rows represent the different samples. The intersection of the i^{th} row with the k^{th} column (the element $a'_{ik} \in \{0,1\}$) denotes the presence or the absence of pattern $k \in F$ in the sample i.

2.3 Feature selection and classification

In this final step we use the data matrix A' as an input to a classification algorithm in order to construct a model-classifier that will be used to classify new samples. Before applying the classification algorithm we may use a feature selection method in order to select a feature subset. Note that applying feature selection to the transformed data matrix A' requires much less time than applying it to the initial data matrix A. This is because the typical number of tags is in the order of tens of thousands, whereas

typical number of patterns contained in F is in the order of hundreds or thousands.

3. Experimental Setup

In this section we define our experimental setup. First, we present the datasets that we experimented with. Then, we present the feature selection, classification, and evaluation methods that were utilized in our experiments.

3.1 Datasets

We have used two real SAGE datasets in our study. The first one consists of 90 SAGE libraries and 27679 tags. The second one is a reduced dataset consisting of 74 SAGE libraries and 822 tags. From now on we will refer to these datasets as the 90x27679 and the 74x822 dataset respectively. Both datasets have been provided by Dr Olivier Gandrillon's team (Centre de Génétique Moléculaire et Cellulaire de Lyon, France) and have been studied and presented at the ECML/PKDD Discovery Challenge Workshops in 2004 and 2005. The SAGE libraries contained in these datasets are publicly available in the SAGEmap website (http://www.ncbi.nlm.nih.gov/SAGE/index.cgi) and have been prepared as of December 2002 [4]. They are collected from various human tissue types (colon, brain, ovary, etc.) and are labeled according to their cell state that is either normal or cancerous.

3.2 Feature selection, classification, evaluation

For the conduction of our experiments we have utilized the Weka library of machine learning algorithms [13]. For feature selection we have used X^2 *Statistic*, *Information Gain*, and *Relief F*. We have also used four classification algorithms, namely *C4.5*, *k-Nearest Neighbors* (*k-NN*), *Support Vector Machine* (*SVM*) with a linear kernel, and *RIPPER* a propositional rule learner. Moreover, for comparison purposes we have utilized a baseline classifier (*Majority*) that always predicts the majority class.

In order to evaluate our approach we used leave-one-out cross-validation (LOOCV). All the steps of our approach (discretization, frequent pattern mining, feature selection, and classification) were undergone LOOCV.

4. Results

In this section we present the results of the experiments that were conducted according to the setup described in the previous section. We compared the results of our approach with a common approach, which includes feature selection and classification of the original data and was also followed in other studies [3, 7]. We will refer to it as the baseline approach.

The six graphs in Figure 1 present the classification accuracy that was achieved on the 99x27679 dataset by all the feature selection and classification algorithms for both the baseline and our approach. The y-axis represents the classification accuracy and the x-axis represents the number of the top ranked features that were selected for constructing the classifiers. A first remark that can be made is that the accuracies of the classifiers of the baseline approach are near to the accuracy of the majority classifier. In particular, the accuracy of RIPPER is almost always worse than the majority classifier's accuracy, whereas the accuracies of the other classifiers sometimes get worse than the majority classifier's accuracy. The best accuracy of the baseline approach (84.44%) was achieved when the 1000 top-ranked (Relief F) features were used by the SVM classifier.

In contrast, the classifiers of our approach achieved accuracies that are far higher than the accuracy of the majority classifier. In particular, the best accuracy (98.89%) was achieved in most cases by the SVM classifier (C4.5 and k-NN also achieved 98.89%), when at least 1000 features were used. The worst accuracy was achieved by k-NN (1000 features), but is 11 percentage points higher than majority classifier's accuracy.

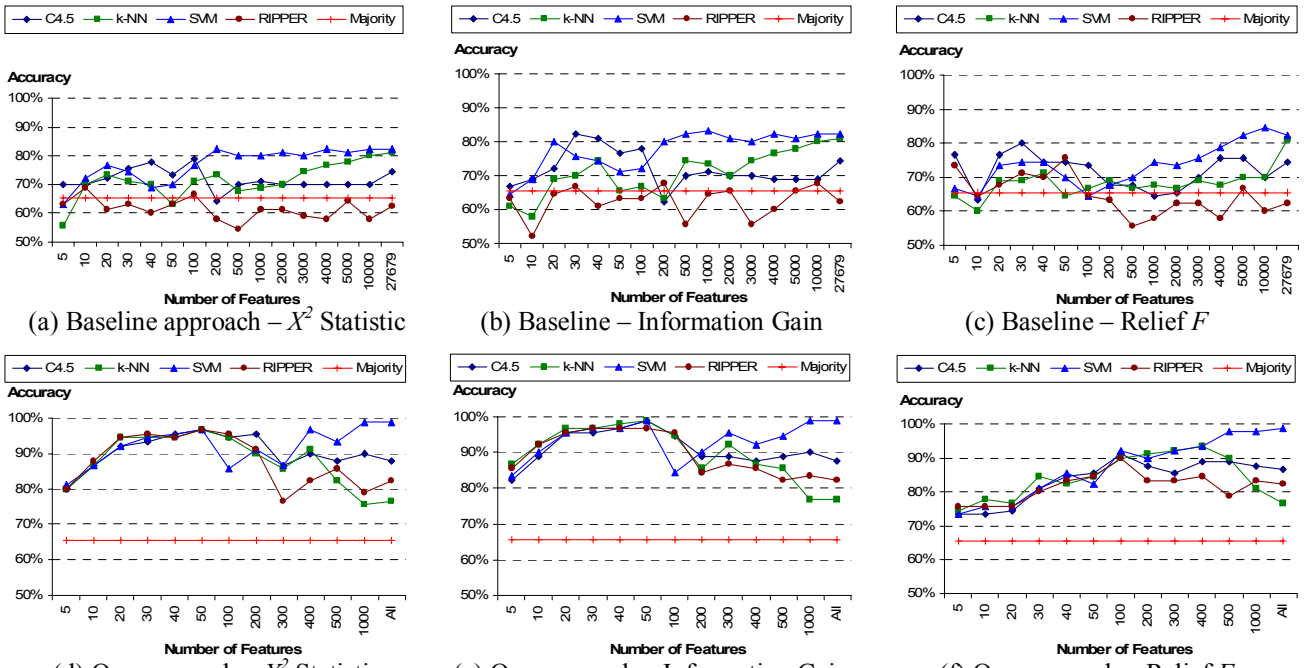
An important observation for the results of X^2 and Information Gain is that all the classifiers of our approach achieve very high (around 95%) and almost constant accuracy for the top 20 up to 50 selected features (recall that the features in our approach are frequent gene expression patterns). In the case of Relief F the accuracy increases with the number of features (from the 20 to the 100 top selected features) reaching approximately 91% at 100 features. This is particularly interesting and indicates that only a small number of patterns are required in order to build very accurate and efficient (using a few features the computational cost is considerably reduced) classifiers.

The accuracies achieved by our approach applied on the 99x27679 dataset with all feature selection methods, with all classification algorithms and for each selected feature set are significantly better than the majority classifier's accuracy at a 95% confidence level, with an exception of some classifiers when Relief F was used for feature selection and the 5, 10 and 20 top ranked features were selected.

Figure 2 presents the results obtained on the 74x822 dataset. The highest accuracy that was achieved by the baseline approach is 85.14% (Relief F, SVM, 500 features), whereas the accuracy of the classifiers of our approach reaches 98.65% in many cases. The most remarkable fact is the early degradation of k-NN classifier constructed in our approach, especially for X^2 and information gain. This can be explained, if we consider that the input data of the classification algorithms in our approach are binary and if we take into account that the distance function used by k-NN is not suitable for binary

samples. Another important remark is that as with the other dataset, a small number of features (10-20) that is demanded for achieving very high accuracies in our approach.

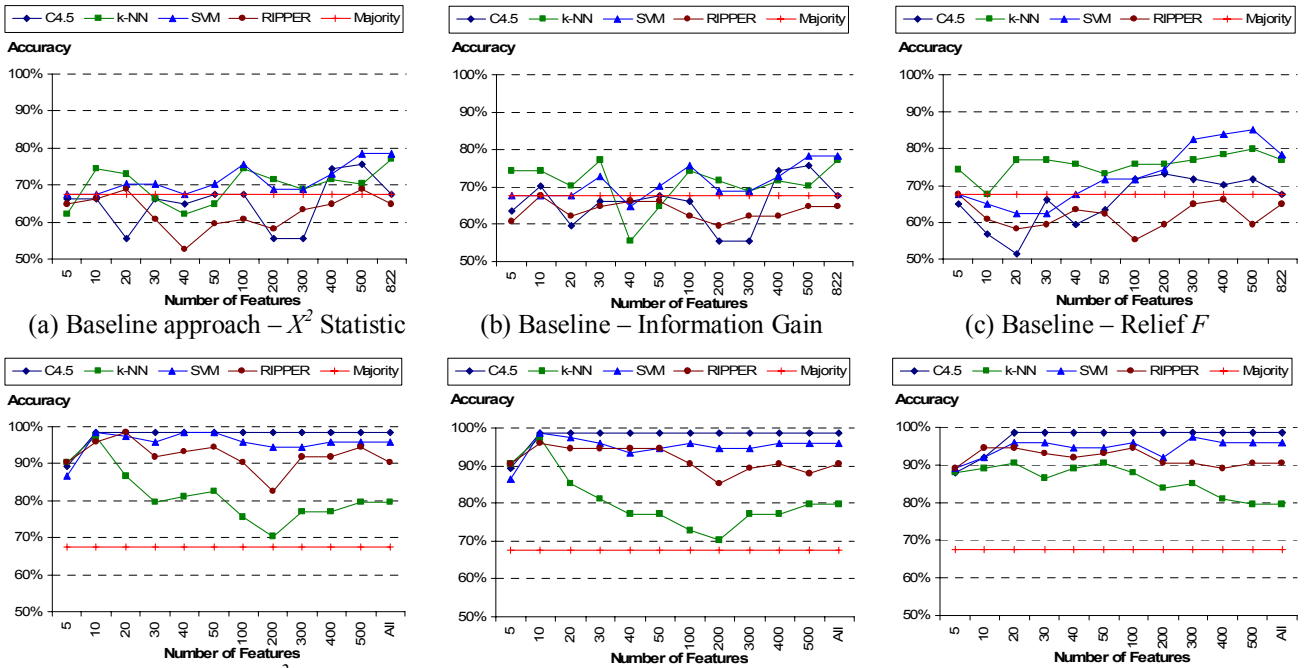
The accuracies achieved by our approach on the 74x822 dataset except for *k*-NN with large feature sets are significantly better than the majority classifier's accuracy at a 95% confidence level.



(a) Baseline approach – X^2 Statistic (b) Baseline – Information Gain (c) Baseline – Relief *F*

(d) Our approach – X^2 Statistic (e) Our approach – Information Gain (f) Our approach – Relief *F*

Figure 1. Classification accuracy on 99x27679 dataset



(a) Baseline approach – X^2 Statistic (b) Baseline – Information Gain (c) Baseline – Relief *F*

(d) Our approach – X^2 Statistic (e) Our approach – Information Gain (f) Our approach – Relief *F*

Figure 2. Classification accuracy on 74x822 dataset

Both datasets contain samples collected from various tissue types. Since the genes that are expressed in each tissue type are different, an increased number of genes should be affect classification. As is shown by the experimental results of the baseline approach, about 500-1000 features seem to be important. The remaining genes are rather irrelevant, but do not affect negatively the classification accuracy. However, in our approach, where we transform the original feature space from genes to patterns of gene expressions (each pattern may contain more than one gene) the number of the important features-patterns decreases dramatically, but the achieved accuracy increases remarkably.

Table 1 presents the highest accuracies reported by previous approaches as well as ours. As shown in the table, the difference between the highest accuracy that was achieved by our approach on the 90x27679 is 13.29 percentage points greater than the highest one reported in the literature for the same dataset. Similarly, the corresponding difference on the 74x822 dataset is 12.47 percentage points.

Table 1. Classification results reported so far

Study	Dataset	Accuracy
Gamberoni and Storari [3]	90x27679	82.20%
Lin and Li [7]	90x27679	85.60%
Alves et al. [1]	74x822	86.18%
Our approach	90x27679	98.89%
	74x822	98.65%

5. Conclusions

In this paper we proposed an approach for effectively and efficiently classifying gene expression data collected with the SAGE method. We have utilized the most prominent patterns of the expressions of genes in order to construct more accurate classifiers. The experiments shown that a small number of patterns (less than a hundred) is adequate to construct very accurate classifiers with accuracy around 95%. So, the benefit is twofold. First, we manage to significantly improve classification accuracy, and second we drastically reduce the data dimensionality and consequently the computational cost. All these are done in the cost of the use of a frequent pattern mining algorithm that is very efficient in datasets like SAGE; such an algorithm is usually devoted in finding frequent patterns in huge datasets with millions of samples and many thousands of features.

Our future plans include the adaptation and application of our approach in gene expression data that were collected with other techniques, like DNA arrays. Also, we intend to study more in-depth the impact of sequencing errors and other possible sources of noise on the effectiveness of gene expression classification.

6. References

[1] A. Alves, N. Zagoruiko, O. Okun, O. Kutnenko, and I. Borisova. "Predictive Analysis of Gene Expression Data from Human SAGE Libraries". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, pp. 60-71.

[2] C. Becquet, S. Blachon, B. Jeudy, J.F. Boulicaut, and O. Gandrillon. "Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data". *Genome Biol*ogy, 3(12), 2002.

[3] G. Gamberoni and S. Storari. "Supervised and unsupervised learning techniques for profiling SAGE results". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, 2004, pp. 121–126.

[4] O. Gandrillon. "Guide to the gene expression data". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, 2004, pp. 116–120

[5] G. Gasmi, T. Hamrouni, S. Abdelhak, S. Ben Yahia, and E. Mephu Nguifo. "Extracting Generic Basis of Association Rules from SAGE Data". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, 84–89.

[6] J. Han, J. Pei, Y. Yin. "Mining frequent patterns without candidate generation". In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas, Texas, USA, 2000, pp. 1-12.

[7] H.-T. Lin and L. Li. "Analysis of SAGE Results with Combined Learning Techniques". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, pp. 102–113.

[8] R. Martinez, R. Christen, C. Pasquier, and N. Pasquier "Exploratory Analysis of Cancer SAGE Data". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, 2005, pp. 72–77.

[9] R.T. Ng, J. Sander and M.C. Sleumer. "Hierarchical cluster analysis of SAGE data for cancer profiling". In *Proceedings of Workshop on Data Mining in Bioinformatics*, 2001, pp. 65-72.

[10] F. Rioult. "Mining strong emerging patterns in wide SAGE data". In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, 2004, pp. 484–487.

[11] G. Tzanis and I. Vlahavas. "Mining High Quality Clusters of SAGE Data". In *Proceedings of the 2nd VLDB Workshop on Data Mining in Bioinformatics*, Vienna, Austria, 2007.

[12] V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler. "Serial analysis of gene expression", *Science*, 270 (5235), 1995, pp. 484-487.

[13] I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco, 2000.