Improving Classification of Microarray Data using Prototype-based Feature Selection

Blaise Hanczar[1], Mélanie Courtine[1], Arriel Benis[1], Corneliu Hennegar[1], Karine Clément[2], Jean-Daniel Zucker[1]

(1) EPML-CNRS IAPuces
LIM&BIO – University Paris13
74, rue Marcel Cachin
93017 Bobigny Cedex
hanczar_blaise@netcourrier.com

(2) INSERM « Avenir » and EA3502
University Paris VI, Hôtel-Dieu,
Service de Nutrition
1 place du Parvis Notre Dame
75004 PARIS – France

ABSTRACT

This paper addresses the problem of improving accuracy in the machine-learning task of classification from microarray data. One of the known issues specifically related to microarray data is the large number of inputs (genes) versus the small number of available samples (conditions). A promising direction of research to decrease the generalization error of classification algorithms is to perform gene selection so as to identify those genes which are potentially most relevant for the classification. Classical feature selection methods are based on direct statistical methods. We present a reduction algorithm based on the notion of prototype-gene. Each prototype represents a set of similar gene according to a given clustering method. We present experimental evidence of the usefulness of combining prototype-based feature selection with statistical gene selection methods for the task of classifying adenocarcinoma from gene expressions.

Keywords

Microarray data, Classification, Prediction, Feature Selection, Dimension Reduction, Gene Selection.

1. INTRODUCTION

In the last few years, the study of the transcriptome has made great progress thanks to the development of microarray technology. Today the number of scientific projects that include studies based on this possibility to measure simultaneously thousands of gene expressions across collections of samples is increasing dramatically. A major challenge in the context of microarray is the task of sample classification. This task is closely related to the ability to improve diagnosis of patients based on their gene expression profile. Based on this diagnosis it will become possible to improve the quality of treatment. The most accurate the model, the greater the confidence in the predictor results will be.

Our original motivation is related to the treatment of obesity related disease. One of our research goal is to build a model based on microarray data to better predict whether an obese patient will respond to a very low caloric diet (VLCD). Such a model would support the hypothesis that there is a signature of the caloric restriction in genes expression. Such a model would be of great use to improve the choice of treatment in obesity related disease [21].

To build a predictive model in supervised learning, it is assumed that a set of variables, which is denoted as *inputs*, has some influence on the *outputs*. The goal is to use the inputs to predict the value of the outputs. For microarray data processing, the gene expression profile is the input of the model and the output function may be any biological parameter. One of the major problems to be dealt with learning from microarray data is due to the fact they are noisy, scarce and expensive to produce. The great challenge for existing machine learning algorithms is to avoid overfitting the data and dealing with the very large number of dimension of the inputs.

Recent studies have reported successful application of different statistical and machine learning methods to classify microarray data [18-20]. For example Khan [20] creates a model which classifies cancers using Neural Networks. Brown [18] presents a method for microarray data classification using Support Vector Machines (SVM). Many Machine Learning algorithms have been experienced and thanks to a fine tuning they often reach comparable prediction accuracy on cross validation when based on the same initial outputs. There is nevertheless a constant need for improving the accuracy of existing algorithm so as to increase their acceptability as valid source of information to support complex diagnosis.

What makes this accuracy problem more difficult than in other fields of Machine Learning is clearly the discrepancy between the large number of genes and the small number of samples available. However, this situation is very unlikely to change for obvious reasons related to the cost of microarrays and the number of inputs that may reach tens of thousands on pangenomic chips. The problem addressed in this paper is to analyze whether reducing the number of genes preserving a maximum of discriminant information improves the learning accuracy. Several methods to reduce dimensions derived from machine learning are presented in the literature, including Genetic Algorithms [5], Wrapper approaches [4], Support Vector Machines [6], etc. Jaeger *et al*. [1] and Qi [2], focus on redundancy reduction and feature extraction. We propose a reduction method called *ProGene* that has two main objectives: first to reduce the cost and complexity of the classifier, and second to improve the accuracy of the model.

Section two introduces the principle of classification methods and the reason behind a reduction dimension step. It presents the state of the art in gene selection methods. The ProGene approach proposed is presented in section three. Section four presents two Datasets, and details the experimentation of different dimension reduction methods. Finally section five introduces experimental results and discusses the interest of ProGene.

2. STATE OF THE ART IN REDUCTION METHODS

2.1 Reduction methods for machine learning

Input space reduction is often the key phase in the building of an accurate classifier [15]. Irrelevant and redundant features have negative effects on classification algorithms. To perform a statistically sound classification, experimental studies have showed that the number of samples must be larger than the number of features. But microarray Datasets contain several thousand genes for only a few samples, many of which are irrelevant for certain classification tasks. Therefore, reducing the number of genes ought to improve the performance of the classifier. By keeping the number of dimensions under control we aim at reducing the problem of the curse of dimensionality (classifier complexity grows exponentially with the number of dimensions).

The dimension reduction methods can be categorized using different criteria:

- Feature *selection* or feature *extraction*. Feature selection selects a subset of best original features. Feature extraction builds new features which are a transformation of the original space.
- *Wrapper* methods or *filter* methods. In wrapper methods, the reduction dimension is classifier-dependent, and in filter methods, it is classifier-independent [10-11]. More precisely in Wrapper methods, the choice of dimensions is related to the performance of the classifier when these particular dimensions are used.
- *Individual* methods or *collective* methods. In collective methods, the relevance value of a gene is also dependent on all other genes. In individual methods the relevance of a gene is computed independently of all others.
- *Specific* or *non-specific*. In specific methods, the reduction is dependent class, whereas it is not the case in non-specific methods.

2.2 Reduction dimension methods applied to microarray data

The classical approaches to reduce dimensions within the Microarray data context are feature selection, filter, individual, specific methods, which identify differentially expressed genes from a set of microarray experiments. A differentially expressed gene is a gene which has the same expression pattern for all examples of the same class, but different for examples belonging to different classes. The relevance value of a gene depends on its capacity of being differentially expressed. These methods rank genes depending on their relevance for discrimination. Then by setting a threshold, one can filter the less relevant genes among those considered. As such, these filtering methods may be seen as particular gene selection methods.

Several reduction dimension methods have been presented in the literature. Ben-Dor *et al.* present several rank-based methods to select potentially interesting genes [7]. Zheng describes two methods to find differentially expressed genes, GS and GSRobust that are detailed below [17]. Tusher *et al.* developed the SAM method (Significant Analysis of Microarrays), which combines t-test and permutations to calculate a False Discovery Rate (FDR) [3]. Today these methods are widely used by biologists. Nevertheless although very useful, there are intrinsic limits to these methods. A non differentially expressed gene will be considered to be irrelevant and will be removed from the classification process even though it might well contain information that would improve the classification accuracy. Thousands of genes and information about their expression are thus "lost". A feature extraction type method would make it possible to represent data in a reduced space using information from all genes. In addition, these methods are individual, the problem is that many machine learning algorithms do base their decisions on a combination of features. Information is contained from a combination of genes, and not only from a single gene. Moreover, the most relevant genes do not necessarily constitute the best subset of genes w.r.t. prediction accuracy. This paper explores the usefulness of an individual feature extraction method to improve prediction accuracy.

3. "PROGENE": A PROTOTYPE-BASED DIMENSION REDUCTION METHOD

We have conceived an individual feature extraction type dimension reduction algorithm. Our method will also be classifier-independent, which means it can be used with any classification algorithm. It will also be faster than wrapper methods which are computationally costly.

The dimension reduction technique we propose is in two steps. The first step is to identify equivalent classes inside the gene space w.r.t. to a given criterion be it the gene expression, the known gene function, or any biologically relevant criteria. The second step is to create gene prototypes that are good representatives of these classes. The classification task is performed using one or more prototype-genes that have been computed by an aggregation of genes that best represent the class.

3.1 Gene class identification

In this first step, we regroup similar genes in classes. The idea is that genes which belong to the same class contain *partially redundant information*, whereas the information held by each class is different.

In our experimentations we have used an unsupervised learning algorithm, K-mean, to form clusters of genes. As

mentioned above the methods used to cluster genes may be different. We mention in the discussion the possibility to base the definition of clusters upon biological function rather than the gene expression. One of the known drawbacks of the K-mean algorithm is that the number of clusters must be given before learning, and that the optimum number of clusters needed is unknown. If too many clusters are created, most information contained in prototype-genes will be redundant. On the other hand, if the number of clusters is not enough, useful information will be lost. In both cases, classifier performance will decrease. Although a method called Wrapper in Machine Learning [10] might be used to identify the best number of clusters we have favored a more straightforward approach by performing a set of tests with a variable number of clusters.

3.2 Construction of "prototype-genes"

For each class that has been created, we define a so-called *prototype gene. This prototype gene* is meant to represent the class. The role of the prototype is to reduce the redundancy of the Dataset, and the current approach used to build the prototype is to take the centre of each cluster. Thus a prototype is a vector whose dimension is the number of samples. Each expression value in a prototype corresponds to the mean of the gene expression in the cluster, for a certain sample (see figure below). This prototype-gene is considered to be a good representative of the whole cluster, as all members are considered to be similar.

Tab.1: Sketch of the ProGene Algorithm

1. CM ← select method of clustering (default : k mean)
2. NBCLUST ← select the desired number of clusters
3. For each iteration of the cross validation
   - 3.1. Define train set and test set
   - 3.2. Do NBCLUST clusters of genes on train set using CM
   - 3.3. For each cluster C_u
     - 3.3.1. Build Prototype P_u ← mean of this cluster
   - 3.4. Model ← training of SVM using prototype
   - 3.5. accuracy ← prediction on the test set
4. Compute the average accuracy

From a biological point of view, this gene prototype is characterized by an expression profile the most similar to all genes of a cluster. In other words, if the cluster reflects a particular function this gene prototype is the prototypic or typical expression profile. We extract the prototype P_u from cluster C_u, the vector $(y_1 \ldots y_M)$ represents the expression of the prototype P_u. This notion is different from the notion of Metagene [22]. Metagene are dominant singular factor (principal component) within a cluster.

$$P_u = (y_1 \cdots y_M) \text{ with } y_j = \frac{1}{Size(C_u)} \sum_{gene(i) \in C_u} x_{i,j}$$

The algorithm ProGene is described in table 1 and Figure1.

4. EXPERIMENTATION

4.1 Data

The current cancer classification methods have been based on histopathology for the last thirty years. However, the limits of these methods have been reaches of providing critical information which may influence treatment strategy. Microarray technology can contribute to improve cancer diagnosis. Two public Datasets have been used to test the ProGene method, one on lung cancers from Harvard [13], and the other on human acute leukemias [14] (cf. table 2).

Lung cancer Dataset

The current lung cancer classification is based on clinicopathological features. More fundamental knowledge of the molecular basis and classification of lung carcinomas could aid in the prediction of patient outcome, the informed selection of currently available therapies, and the identification of novel molecular targets for chemotherapy. The development of microarray methods for large-scale analysis of gene expression makes it possible to search systematically for molecular markers of cancer classification and outcome prediction in a variety of tumor types.

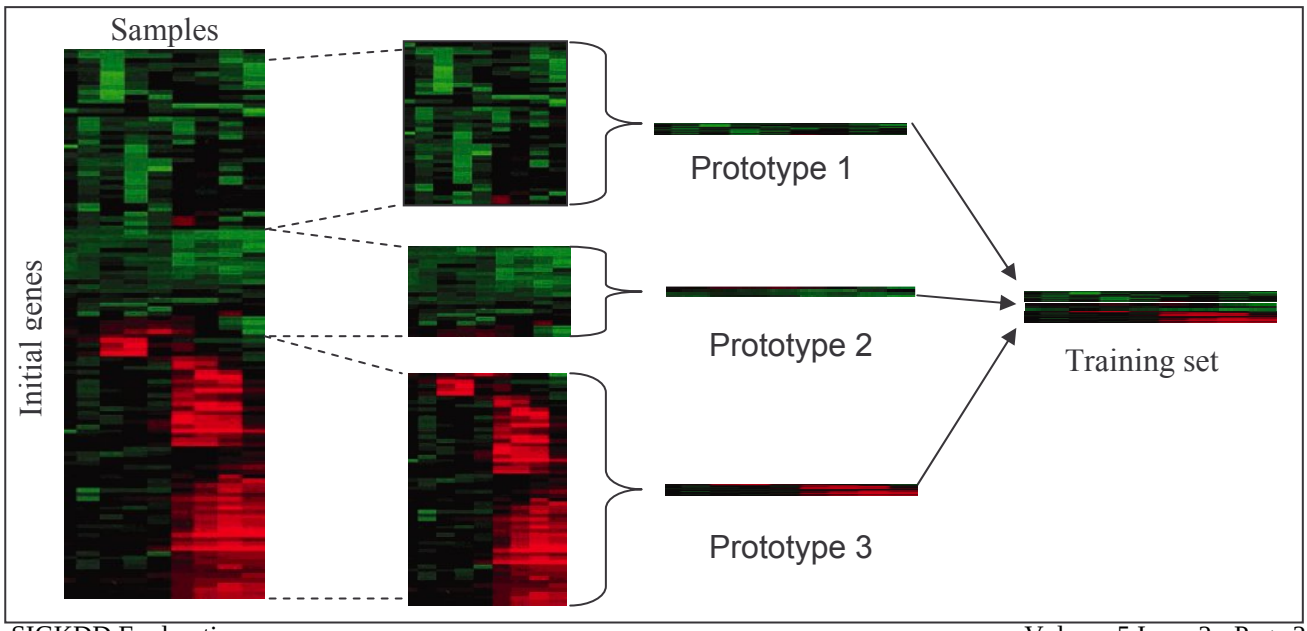

Fig.1: Algorithm ProGene

The Harvard Dataset contains 186 samples of lung cancer divided into subtypes: 139 adenocarcinomas, 21 squamous cell lung carcinomas, 20 pulmonary carcinoids, and 6 small-cell lung carcinomas, plus 17 normal lung samples. Total RNA extracted from samples was used to generate cRNA targets, subsequently hybridized to human U95A oligonucleotide probe arrays, according to the standard protocol [14], which measures the expression of a 12600 probe set. The goal in this task is to build a classifier which could accurately classify the cancer type, using only the gene expression profile. The Dataset of 202 patients was divided into 5 classes: one for each type of cancer.

Leukemia Dataset

Although the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) has been well made, no single test is currently sufficient to establish the diagnosis. Instead, current clinical practice involves an experienced hematopathologist's interpretation of the tumor's morphology, histochemistry, immunophenotyping, and cytogenetic analysis, each performed in a separate, highly specialized laboratory. Although usually accurate, leukemia classification remains imperfect and errors do occur. Distinguishing AML from ALL is critical for successful treatment. A more systematic approach to cancer classification have been developed based on the simultaneous expression monitoring of thousands of genes using DNA microarrays.

The Dataset published in [14] contains 72 samples of AML and ALL. RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix and containing 7129 probes for 6817 human genes. The goal in this task is to build a classifier which could accurately classify whether a patient had AML or an ALL, using only the gene expression profile.

Tab2 : The two Datasets used to experiment with ProGene reduction method

Dataset	Number of samples	Number of classes	Number of features
Lung cancer	202	5	12600
Leukemia	72	2	7129

4.2 Classical gene selection methods

The performance of ProGene have been compared to usual gene selection methods, all of which selects differentially expressed genes and rank the genes in order of importance. Three methods were selected SAM [3], GS, and GSRobust [17]. They are detailed hereafter.

SAM (Significant Analysis of Microarrays)

This is today a well-known method to identify differentially expressed genes. It is based on a classical statistical test (false discovery rate), and selects those genes that are the most useful in order to distinguish the classes. There are four important steps: first, to form a statistic for each gene; second to compute the null distribution for these statistics; third to choose the rejection regions; and fourth to assess or monitor the number of false positives by choosing a threshold. Given a set a genes and a parameter, SAM provides a subset of genes that are considered significant.

GS and GSRobust

GS and GSRobust are both presented by Zheng [17]. A statistic based on the sum of square error between classes and the sum of square error within classes is computed. GSRobust is a variant of GS which is not based on the sum of square error, but on the median absolute deviation (MAD). For each gene, a relevance value is computed:

$$GS_i = \frac{\sum_{j=1}^{k} \left(\overline{g_{ij}} - \overline{g_{i...}}\right)^2 / (k-1)}{\sum_{j=1}^{k}\sum_{l=1}^{n_{ij}} \left(g_{ijl} - \overline{g_{ij}}\right)^2 / (n_{ij}-1)}$$

$$GSRobust_i = \frac{MAD\left[median(\underline{g_{i1}}),...,median(\underline{g_{ik}})\right]}{\sum_{j=1}^{k} MAD(\underline{g_{ij}})}$$

where, k is the number of classes, $\underline{g_{ij}}$ is the vector of gene expression for the ith gene in the jth class, and

$$\overline{g_{ij}} = mean\left(g_{ij}\right)$$
$$g_i = mean\left\{mean\left(g_{ij}\right), j = 1,...,k\right\}$$

4.3 Study design

To assess the generalization error of the classifier learnt, the samples were iteratively and randomly divided into two Datasets: two thirds of the samples formed the training subset, and the last third formed the test subset. We used our dimension reduction method and the classifier learning on the training subset, and then tested the classifier on the test subset to compute the accuracy of the model. This cross validation is a widely accepted approach to assess classifier accuracy in Machine Learning.

As for the Machine Learning algorithm, we have chosen a Support Vector Machine [9] classification method. Support vector machines are robust when applied on sparse and noisy data [23]. We ran this procedure several times to average results.

To highlight the importance of the clustering step in ProGene, we tested a variant of our method; "random clustering", where we created clusters randomly. This approach is the one advocated by Qi [2]. Qi reports significant improvement by a random gene selection before learning. The average error rate of these 100 tests for each k value was computed in order to assess the average performance of this method.

For the reduction methods presented (GS, GSRobust, ProGene, randomProGene) the error rate of the obtained classifier was also computed, varying the number of selected gene or prototype between 10 and 1000.

Because the SAM approach requires that the user fix a threshold manually, we limited ourselves to three typical selections: genes which are significant at 5%, genes which form a subset which has the smallest FDR (False discovery rate), and the hundred most significant genes. Then we built a classifier using only the selected genes, and we reported the error rate of the classifiers obtained.

All these methods were tested on both the Lung cancer and Leukemia Dataset, and implemented using the R statistical software environment [16].

4.4 Quality measure of the clustering

We computed the quality of the clustering in ProGene using a measure inspired by Ray and Turi [8]. Intra-cluster distance is represented by the average distance between the genes belonging to the same cluster. The higher the intra-cluster distance, the better the quality of the cluster.

$$Intra = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{N_i}\sum_{x\in C_i}\left|x - z_i\right|\right)$$

where k is the number of clusters and z_i the cluster center of the cluster i.

The inter-cluster distance represents the smallest distance between two clusters. The higher the inter-cluster distance, the more different the prototypes are.

$$Inter = \min\left(z_i - z_j\right), \left\{i \in \left[1, k-1\right], j \in \left[i+1, k\right]\right\}$$

We used the ratio between the intra- and the inter-distances as a clustering quality measure. To obtain a good clustering, inter-cluster distance must be maximized and intra-cluster distance must be minimized. Therefore, the lower the quality measure, the better the clustering.

$$ClusteringQuality = \frac{Intra}{Inter}$$

5. RESULTS AND DISCUSSION

5.1 Results of ProGene

Figure 2 shows the error rate of our method as a function of the number of clusters k on the lung cancer Dataset. ProGene yielded better performance when k was smaller than 300, in the best case an error rate of 8.3% for k=100. With a large number of clusters (k>300) the results are not so good the more the gene prototype the worst the performance become.

Figure 3 shows the error rate of our method as a function of the number of clusters k on the leukemia Dataset. We see a similar behavior with the best case an error rate of 4.5% for k=200, so a result five times better than without selection methods.

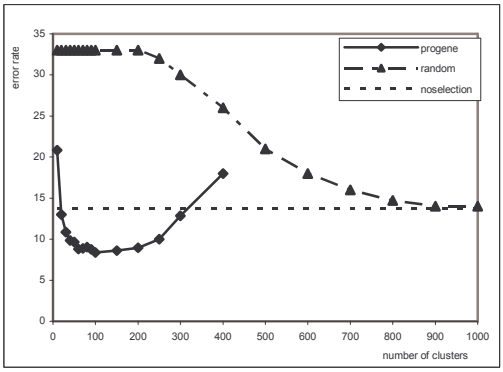

Fig.2: Comparative results of ProGene, Random clustering and no selection methods on the lung cancer Dataset

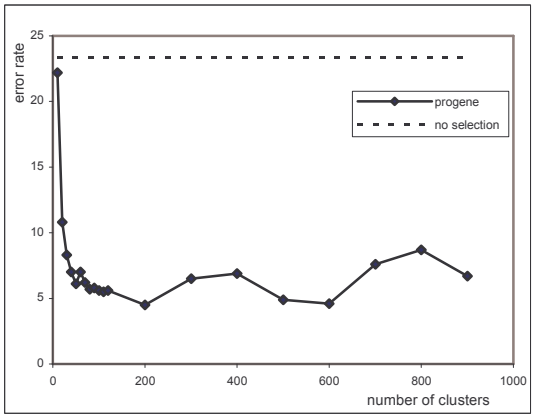

Fig.3: Comparative results of ProGene and no selection methods on the leukemia Dataset

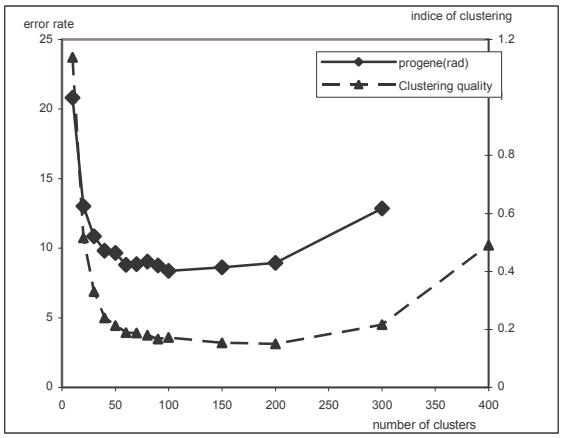

Fig.4: Quality measure of the clustering on the lung cancer Dataset

To understand and explain the variation in ProGene performance we plotted the quality measure of the clustering used by ProGene as a function of k on the lung cancer Dataset (see Figure 4). Minimum clustering quality occurs when k=100, which corresponds to our best prediction rate. There is clearly a correlation between the quality measure of the clustering and the error rate of our classifier. It suggests a heuristic approach to select the optimum k value based on this

quality measure. This measure could provide a significant speedup of performance although the link between the optimum and the quality measure has yet to be demonstrated.

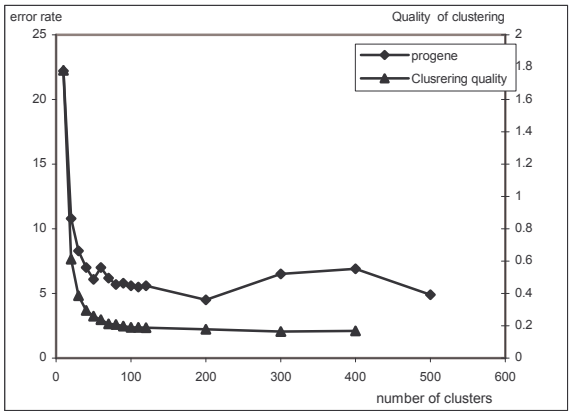

Fig.5: Quality measure of the clustering on the leukemia Dataset

We notice the same type of correlation between classification performances and the quality measurement of the cluster on the leukemia Dataset (cf. figure 5). This confirms the hypothesis that improvement of the clustering is associated with better extraction by the prototypes of the useful information existing within the data, and thus leading to a better classification.

5.2 Random clustering

The results obtained with the above mentioned method are presented in figure 2. The worst results are found for k<300 when the clusters become too large and heterogeneous. All samples come from the same class, the biggest one, resulting in an error rate of 31%. The closer k is to the number of genes, the closer the performances are to that of the "no selection" method.

5.3 Comparing ProGene with other methods

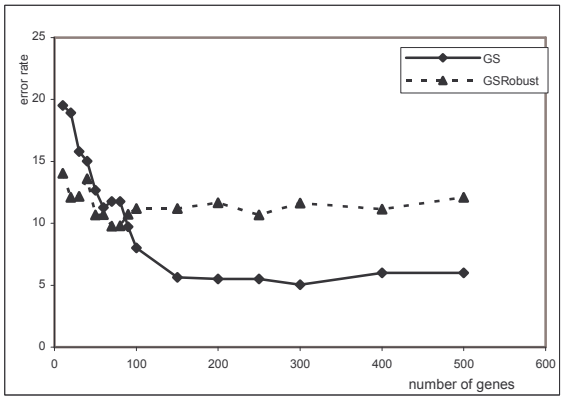

Fig.6: Results of GS and GSRobust methods on the lung cancer Dataset

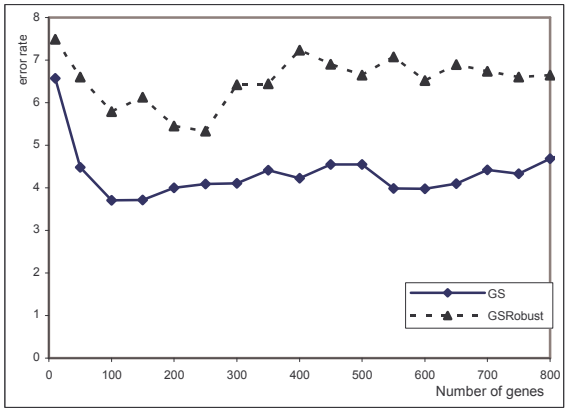

Fig.7: Results of GS and GSRobust methods on the leukemia Dataset

In figure 6 and 7 classification results obtained by using GS and GSRobust methods of genes selection are given. GSRobust has rather stable performances whatever the number of selected genes, with approximately 12% of error rate for lung cancer Dataset and 6% for leukemia Dataset. GS is more dependant of the number of selected genes but allows to obtain better performances, with 5.6% error rate for lung cancer Dataset and 3.7% for leukemias Dataset.

Tables 3 and 4 show the comparison of the best performances of all dimension reduction method on the two Datasets. These results show the key role played by reduction method on performances. Indeed all methods are causing a drop in the error rate. We note that GS produced best results, followed closely by ProGene and SAM.

Tab.3: Performance on lung cancer Dataset

	Error rate	Number of features
No Selection	13.8% ± 1.3	12600
GS	5.6%±1.4	200
GSRobust	9.8%±2.6	70
SAM	8% ± 1.3	100
ProGene	8.3% ± 1.2	100
Progene+GS	4.8% ± 1.1	100

Tab.4: Performance on leukemia Dataset

	Error rate	Number of features
No Selection	23.4%	7129
GS	3.7%±1.4	100
GSRobust	5.3%±3.2	250
ProGene	4.5%±1.5	200
ProGene+GS	1.4% ± 1	50

5.4 Combining ProGene with GS

Ranking methods and ProGene have similar performances. However, their objectives are very different. Ranking methods select genes which can statistically distinguish two classes and ProGene attempts to eliminate redundancy within the data. This suggests

that it could be useful to combine these two types of method so as to take advantage of both aspects.

We tried to combine ProGene algorithm with a selection by GS. We started by using ProGene for building the prototype genes. Then we made a selection of the prototype genes based on the GS, as if prototype genes where normal genes. Finally, we carried out the classification by using only the prototypes selected by GS. We tested this combined procedure on the two Datasets.

We tested this combination of algorithms on the leukemia Dataset, by varying the number of clusters from 200 to 800, which corresponds to the value for which the best results were obtained. The number of selected prototypes varies between 25 and 200.

Tab.5: Error rate of combination of ProGene and GS on the leukemia Dataset. Numbers in bold correspond to best results for a given number of clusters.

Error rate when combining ProGene and GS							
		Number of genes					
Nb clusters		**25**	**50**	**75**	**100**	**150**	**200**
	200	2.90%	**1.45%**	4.35%	5.80%	7.08%	4.18%
	400	8.06%	3.83%	3.83%	5.28%	3.83%	**2.64%**
	600	9.74%	8.41%	4.51%	**2.92%**	4.51%	4.46%
	800	2.65%	2.65%	**1.67%**	2.65%	3.52%	**1.67%**

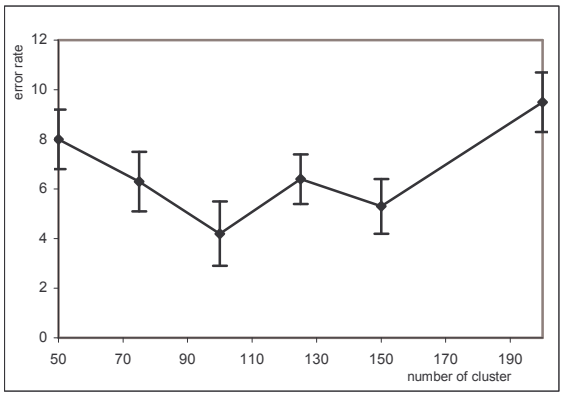

Fig.8: Results of combination of prgene and GS on the lung cancer Dataset

On the lung cancer Dataset, we chose only one number of clusters: the maximum before the performance drop of ProGene, which is 250. The number of selected prototypes varies between 50 and 200.

There is an improvement in performance by using the combination of these two methods. The error rate drops down to 1.4% for the leukemia Dataset, obtained by creating 200 prototypes and by selecting only 50 of them (cf. table 5). The error rate drops down to 4.8% on the lung cancer Dataset by building 250 prototypes and by selecting 100 of them. ProGene eliminates the redundancy within the data by creating prototypes (cf. figure 8). And because not all the prototypes are relevant for a classification, they are filtered out by the GS algorithm. For each number of clusters there exist a combination of ProGene and GS that outperforms each one taken independently.

6. CONCLUSION AND FUTURE WORK

To improve the accuracy of machine learning based classifier algorithm, reduction methods play a key role. We have developed a somewhat original dimension reduction method, which experimentally increases classification accuracy of a Support Vector Machine based classifier. Although the performance gain is comparable to that of classical reduction methods combining them outperforms both methods. It suggests that their filtering approach is complementary.

Our future research is focused along three lines. Because ProGene seems to be rather sensitive to the number of clusters, we have started to explore various approaches to select its optimal value as a hyper-parameter. We are also investigating the use of other types of prototype and testing the use of enriched descriptions of prototype (such as standard deviation of clusters in prototype construction). This method should also be compared more throughly the Metagene approach that extracts principal components of each cluster [22]. Finally, we are exploring new type of clustering approach based on the biological function of genes and not only on their expression. We have carried out preliminary experiments using the biological function of the Gene Ontology [12] to form biologically relevant clusters of genes As per now we have used second and third level GO terms and the results we have obtained do not outperform clustering based on expression. The GO terms used might have been too general gene they were from the second and third level of generality. We are now investigating the use of more specific biological function and assessing the accuracy improvements.

REFERENCES

[1] Jaeger J., Sengupta R., Ruzzo W.L., Improved Gene Selection for Classification of Microarrays. Pacific Symposium on Biocomputing, (2003), 53-64.

[2] Qi H., Feature selection and kNN fusion in molecular classification of multiple tumor types, International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS'02), (June 2002).

[3] Tusher V.G., Tibshirani R., Chu G., Significance analysis of microarrays applied to the ionizing radiation response, PNAS (2001), 98:5116-5121.

[4] Inza I., Sierra B., Blanco R., Larrañaga P., Gene selection by sequential wrapper approaches in microarray cancer class prediction, Journal of Intelligent and Fuzzy Systems, (2002), 25-34.

[5] Li L., Darden T.A., Weinberg C.R., Levine A.J., Pedersen L.G., Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method, Combinatorial Chemistry and High Throughput Screening, (2001), 727-739.

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

[7] Ben-Dor A., Friedman N., & Yakhini Z., Scoring genes for relevance, Agilent Technologies Technical Report AGL-2000-13, (2000).

[8] Ray S., Turi R.H., Determination of number of cluster in k-means clustering and application in colour image segmentation, Proceeding in the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99), (1999), 137-143.

[9] Cristianini N., Shawe-Taylor J., an introduction to support vector machines (and other kernel-based learning methods), Cambridge University Press (2000).

[10] Blum A, Langley P., Selection of relevant features and examples in machine learning, Artificial Intelligence, (1997), 245-271.

[11] Wilson D.R., Martinez T.R., Reduction Techniques for Exemplar-Based Learning Algorithms, Machine Learning (1998).

[12] Gene ontology™ consortium : http://www.geneontology.org

[13] Bhattacharjee A. *et al.,* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, PNAS, (2001), 13790-13795.

[14] Golub T. R., *et al.*, Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Sciences, (1999), 286:531-537

[15] Cakmakov D., Bennani Y., Feature selection for pattern recognition, (2002).

[16] Becker R.A., Chambers J.M. and Wilks A.R., The New S Language, Chapman & Hall, New York. (1988).

[17] Zheng G., Olusegun E., Narasimhan G., Neural Network Classifiers and Gene Selection Methods for Microarray Data on Human Lung Adenocarcinoma, CAMDA03, (2003), 63-67.

[18] Brown M.P.S., Grundy W., Lin D., Cristianni N., Sugnet C., Furey T., Ares M., Hauussler D., Knowledge based analysis of microarray gene expression data using support vector machine, Proc Natl. Acad. Sci., (2000), 97(1):263-267

[19] Shipp M.A. et al. Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. Nature Medecine (2002), 8:68-74.

[20] Khan J., Wei J.S., Ringner M., Saal L.H., Ladanyi M., Westermann F., Berthold F., Schwarb M., Antonescu C.R., Peterson C., Meltzer PS., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medecine (2001) 7(6):673-679

[21] Clement K., « Monogenic forms of obesity: from mice to human ». Ann Endocrinol, (2000); 61(1)

[22] Huang E., Cheng S.H., Dressman H., Pittman J., Tsou M., Horng C., Bild A., S Iversen E.S., Liao M., Chen C., West M., Nevins J.R., Huang A.T., .Gene Expression Predictors of Breast Cancer Outcomes, Lancet, CAMDA03, (2003).

[23] Brown M.P.S., Grundy W.N., Lin D., Cristianini N., Sugnet C.W., Furey T.S., Ares M., Haussler D., Knowledge-based analysis of microarray gene expression data by using support vector machines, PNAS, (2000), 97: 262-267.