Rule induction for classification of gene expression array data

Per Lidén[1], Lars Asker[1,2] and Henrik Boström[1,2]

[1] Virtual Genetics Laboratory AB, Fogdevreten 2A, SE-171 77 Stockholm, Sweden
{per.liden, lars.asker, henrik.bostrom}vglab.com
http://www.vglab.com
[2] Department of Computer and Systems Sciences, Stockholm University and Royal Institute of Technology, Forum 100, SE-164 40 Kista, Sweden
{asker, henke}@dsv.su.se

Abstract. Gene expression array technology has rapidly become a standard tool for biologists. Its use within areas such as diagnostics, toxicology, and genetics, calls for good methods for finding patterns and prediction models from the generated data. Rule induction is one promising candidate method due to several attractive properties such as high level of expressiveness and interpretability. In this work we investigate the use of rule induction methods for mining gene expression patterns from various cancer types. Three different rule induction methods are evaluated on two public tumor tissue data sets. The methods are shown to obtain as good prediction accuracy as the best current methods, at the same time allowing for straightforward interpretation of the prediction models. These models typically consist of small sets of simple rules, which associate a few genes and expression levels with specific types of cancer. We also show that information gain is a useful measure for ranked feature selection in this domain.

1 Introduction

Gene expression array technology has become a standard tool for studying patterns and dynamics of the genetic mechanisms of living cells. Many recent studies have highlighted its usefulness for studying cancer [1], [2], [3]. Gene expression profiling does not only provide an attractive alternative to current standard techniques such as histology, genotyping, and immunostaining for tumor classification, but also valuable insights into the molecular characteristics of specific cancer types. In clinical settings, diagnosis is of great importance for the treatment of cancer patients since the responsiveness for various drugs and prognostic outcome can vary between subtypes of cancers. Correct tumor classification will ideally optimize treatment, save time and resources, and avoid unnecessary clinical side effects.

To date, a number of methods have been applied to the problem of learning computers to classify cancer types based on gene expression measurements from microarrays. Alon et al used clustering as a means for classification in their original analysis of the colon cancer data set [2]. Subsequently, methods such as Support

Vector Machines (SVMs) [4], [5], Naïve Bayesian Classification [6], Artificial Neural Networks (ANNs) [3], and decision trees [7] have been employed to address this task. Some of these studies indicate that besides creating accurate prediction models, an important goal is to find valuable information about the system components that are being used as input to these models.

Previous studies have shown that classification accuracy can be improved by reducing the number of features used as input to the machine learning method [3], [1]. The reason for this is most likely that the high level of correlation between the expression levels of many genes in the cell makes much of the information from one microarray redundant. The relevance of good feature ranking methods in this domain has also been discussed by Guyon and colleagues [8].

Rule induction methods have been studied for more than two decades within the field of machine learning. They include various techniques such as divide-and-conquer (recursive partitioning), that generates hierarchically organized rules (decision trees) [9], and separate-and-conquer (covering) that generates overlapping rules. These may either be treated as ordered (decision lists) [10] or unordered rule sets [11]. Common for all these methods is that they are very attractive with regard to the analysis of input feature importance. Since the rule induction process in itself takes redundancy of input parameters into account, and that the process will seek to use the most significant features first, superfluous features are commonly left outside the prediction model. In this study we investigate three different rule induction methods for classifying cancer types based on gene expression measurements from microarrays, together with a simple method for feature selection based on information gain ranking.

Two public datasets are used in this study. The first data set is the colon cancer study published by Alon and co-workers [2]. Here the task is to separate between tumor tissue and normal colon tissue (a two class problem). This data set has been extensively studied by others [2], [4], [5], [6]. The prediction task for the second data set [3] is to discriminate between four types of small round blue cell tumors (SRBCTs): neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt's lymphoma (BL), and the Ewing family of tumors (EWS).

2 Learning Algorithms

The rule induction system used in this study, Virtual Predict 1.0 [12], extends several rule induction algorithms developed in Spectre 3.0 [13]. The three methods that are used in the experiments are briefly described below. All three methods use a technique for discretizing numerical features during the induction process based on finding split points that separate examples belonging to different classes [14]. The technique is further optimized with respect to efficiency by using a sampling scheme that randomly selects 10% of the possible split points for each feature. All methods use the m-estimate [15], with m set to 2, for calculating class probabilities.

2.1 Divide-and-Conquer Using the Minimum Description Length Principle (DAC-MDL)

Divide-and-conquer (DAC), also known as recursive partitioning, is a technique that generates hierarchically organized rule sets (decision trees). In this work, DAC is combined with the information gain criterion [9] for selecting branching features. Furthermore, the minimum description length (MDL) criterion [16], modified to handle numerical attributes effectively, is used to avoid over-fitting. This method, referred to as DAC MDL, is preferred instead of splitting the training data into one grow and one prune set. Splitting into grow and prune set for this data is more likely to result in highly variable rule sets due to the limited size of the data sets.

2.2 Boosting Decision Trees (Boosting 50)

Boosting is an ensemble learning method that uses a weight distribution over the training examples and iteratively re-adjusts the distribution after having generated each component classifier in the ensemble. This is done in a way so that the learning algorithm focuses on those examples that are classified incorrectly by the current ensemble. New examples are classified according to a weighted vote of the classifiers in the ensemble [17]. The base learning method used in this study is divide-and-conquer using information gain together with a randomly selected prune set corresponding to 33% of the total weight. The number of trees in the ensemble is set to 50. Thus, the method generates an ensemble consisting of 50 individual base classifiers.

2.3 Separate-and-Conquer for Unordered Rule Sets (Unordered SAC)

Finally, a method that employs a totally different search strategy is compared to the previous methods, namely separate-and-conquer (SAC), also known as covering. SAC iteratively finds one rule that covers a subset of the data instead of recursively partitioning the entire data set, cf., [18]. The examples covered by this rule are then subtracted from the entire data set. This strategy is combined with incremental reduced error pruning [19], where each clause immediately after its generation is pruned back to the best ancestor. The criterion for choosing the best ancestor is to select the most compressive rule using an MDL coding scheme similar to the one in [16] but adapted to the single-rule case. The method generates an unordered set of rules, in contrast to generating a decision list [10]. This means that rules are generated independently for each class, and any conflicts due to overlapping rules are resolved during classification by using the naïve Bayes' inference rule (i.e., calculating class probabilities while assuming independent rule coverage).

2.4 Feature Selection Using Information Gain

Since the number of features in the two data sets in the current study is more than 25 times the number of data points, some dimensionality reduction scheme may prove useful for obtaining accurate models, in particular for the methods that generate single models. Although the use of the minimum description length criterion has shown to be quite effective for generating models that are tolerant against noise and random correlations, a large number of irrelevant and redundant variables may cause the rules to be over-pruned due to the additional cost of investigating these superfluous variables. Commonly used dimensionality reduction methods include principal component analysis, multi-dimensional scaling, and feature selection. Since the two former classes of methods do not necessarily lead to dimensions that are suited for discriminating examples belonging to different classes, a feature selection method based on the discriminative power (as measured by information gain) is preferred. This method has the additional benefit of allowing for direct interpretation of the generated rules (i.e., there is no need for transforming the rules back to the original feature space). The following formula, which is a Laplace corrected version of the formula in [9], is used to measure the information content for a numeric feature f and threshold value t:

$$I_{f,t} = \sum_{i=1}^{n} - l_i \log_2 \frac{l_i + 1}{l + n} + \sum_{i=1}^{n} - r_i \log_2 \frac{r_i + 1}{r + n}$$

where n is the number of classes, l_i is the number of examples of class i in the first subset (*i.e.*, examples with a value on f that is less than or equal to t), l is the total number of examples in the first subset, r_i denotes the number of examples of class i in the second subset (*i.e.*, examples with a value on f that is greater than t), and r is the total number of examples in the second subset. It should be noted that the above formula is restricted to evaluating binary splits (i.e., two elements in the partition), which is sufficient when dealing with numeric features that are divided into two intervals. For each numeric feature, all split points obtained from the examples were evaluated, and the k most informative features (*i.e.*, those resulting in the subsets with least information content) were kept, for some given k.

Other feature selection methods could be used as well, but the above was chosen because of its simplicity and expected suitability.

3 Experimental Evaluation

3.1 Colon Cancer Data Set

For a set of 62 samples, 40 tumor samples and 22 normal colon tissue samples, the gene expression levels of 6500 genes were measured using Affymetrix oligonucleotide arrays [2]. Of these genes, the 2000 with the highest minimal intensity were selected by the authors for further analysis. The raw data was normalized for global variance between arrays by dividing the intensities of all genes by the average intensity of the array and multiplying by 50.

Feature selection. This is a two-class dataset, for which feature selection was done in one iteration. In this case, the 128 most highly ranked genes according to the information gain measure were selected for further analysis.

Classification results. Leave-one-out cross-validation was performed (Figure 1a) with the 2, 4, 8, 16, 32, 64, and 128 most highly ranked features. A few points can be made of the results of this analysis. The ensemble method Boosting 50, gave the best prediction accuracy using all 128 features, resulting in 7 misclassified examples. This accuracy does not significantly differ from other results reported for this data set. SVMs gave 6 errors [5], clustering gave 7 errors [2], [4], and Naïve Bayes classification gave 9 errors [6]. It is interesting to note that the boosting method works significantly better when applied to decision trees than decision stumps (i.e., one-level decision trees); 89% accuracy in our case vs. 73% for stumps, as evaluated by Ben-Dor and co-workers [4]. Zhang and co-workers report classification accuracy above 90% using a decision tree induction method similar to ours [7]. However, their analysis can be discussed from a methodological point of view, since the tree structure was induced using the entire data set, and the split-point values were the only parameters that were changed during the five-fold cross-validation for which this result was reported. This method thus takes advantage from a significant amount of information from the data it is going to be evaluated on, which is likely to result in an over-optimistic estimate.

Interestingly, the largest number of features resulted in best prediction accuracy for the ensemble method (Boosting 50). Figure 1a highlights a trend towards better classification with fewer attributes for the simple methods, and the opposite trend for the ensemble method, which is known to be more robust with respect to handling variance due to small sample sizes in relation to the number of features.

3.2 Small Round Blue Cell Tumor (SRBCT) Data Set

The expression levels for 6567 genes in 88 samples of both tissue biopsies and cell lines were measured using cDNA microarrays and 2308 of those genes were selected by a filtering step and divided into one training set (63 samples) and one test set (25

samples) [3]. Class labels were assigned by histological analysis. We have used the same division into test and training as in the original work. In the entire dataset, 29 examples were from the Ewing family of tumors (EWS) (of which 6 were test examples), 11 were Burkitt's lymphoma (BL) (3 test examples), 18 were neuroblastoma (NB) (6 test examples), and 25 were rhabdomyosarcoma (RMS) (5 test examples). The test set also included five non-tumor samples.

Feature selection. In order to select the best candidate features (genes) for this dataset, the information gain for each feature was calculated with respect to its usefulness for separating each of the four classes from the other three. The 32 top ranking features for each class were then selected, resulting in 125 unique genes out of a total of 128 selected (three genes occurred twice).

Classification results. The best classifier generated from the training set, Boosting 50, perfectly separates the 20 cancer samples in the test set. This separation is obtained using only the four attributes corresponding to the top ranked feature for each class. The same result is obtained for twelve and all selected features as well. Using the 96 features selected by Khan and co-workers [3], 100 % accuracy is obtained as well. One difference between our results and the ANN approach of Khan et al is the relative simplicity of the model generated here. The rule based prediction models that produce 100 % accuracy on test examples are typically based on about 200 rules, regardless of the number of features used. This means that every decision tree in the ensemble is on average composed of four rules, and that the entire classifier can be manually inspected (although with some difficulty). This can be compared to the 3750 ANN models created by Khan and colleagues. The other two methods performed slightly worse. At their best, Unordered SAC misclassified two test examples (for 32 features), while DAC MDL misclassified three test examples (also for 32 features). On the other hand, they generated significantly smaller models, consisting of five rules each.

We also performed leave-one-out cross validation of the entire dataset (both training and test examples) using the 4, 8, 16, 32, 64, and 128 most highly ranked features. The top $n/4$ ranked features for each class were selected in every round, where n is the total number of features selected. Error free classification was obtained for Boosting 50 when all the 128 features were selected, while one example was misclassified for 16, 32, and 64 features resulting in 99% accuracy (Figure 1b).

The trend of obtaining better classification with fewer attributes for the simple methods, and the opposite trend for the ensemble method that we noticed in the other experiment, can be observed also here, although this data set has four classes instead, where each class has its own ranked set of features.

3.3 Inspecting the Rules

In the previous section it was shown that the employed rule induction methods are useful for constructing accurate prediction models. In addition to this, rules can also give valuable insights into the studied systems. As an illustration, seven easily

interpretable rules are found when applying the unordered SAC method using the 16 highest-ranking features on the entire SRBCT data set (Table 1).

Table 1: Rules discovered by unordered SAC for SRBCT data set

		Coverage of examples			
Class	**Rule**	**EWS**	**BL**	**NB**	**RMS**
EWS	FVT1 > 1.35535	27	0	0	1
EWS	Caveolin 1 > 1.59365	26	0	0	1
BL	WASP > 0.61645	0	11	0	0
NB	AF1Q > 2.1795	0	0	17	0
NB	CSDA <= 0.69175	0	0	13	0
RMS	SGCA > 0.4218	0	1	0	24
RMS	IGF2 > 13.5508	0	0	0	4

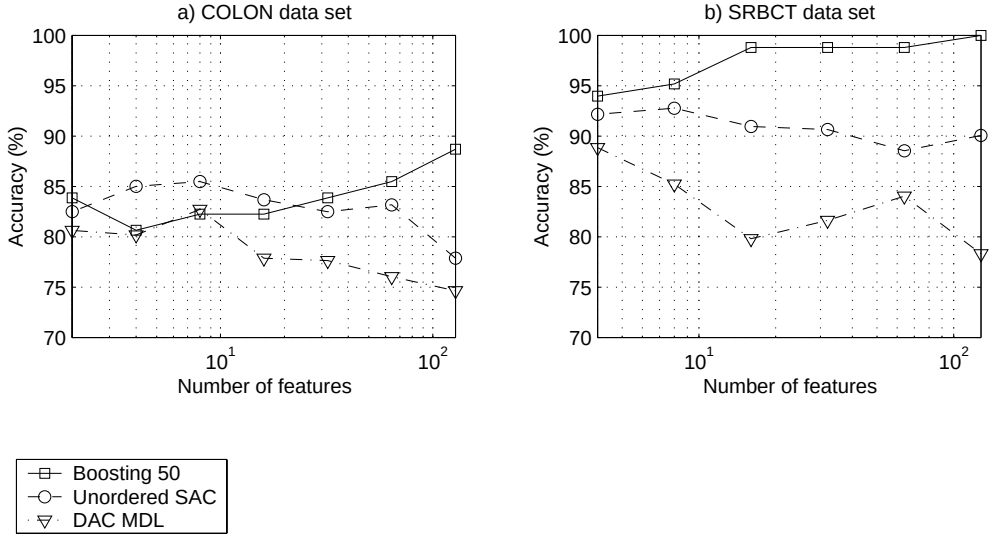

Fig. 1. Results from leave-one-out cross validation both data sets using DAC MDL, Unordered SAC, and Boosting 50. a) Results from the COLON data set using 2, 4, 8, 16, 32, 64, and 128 features. b) Results from the SRBCT data set using 4, 8, 16, 32, 64, and 128 features.

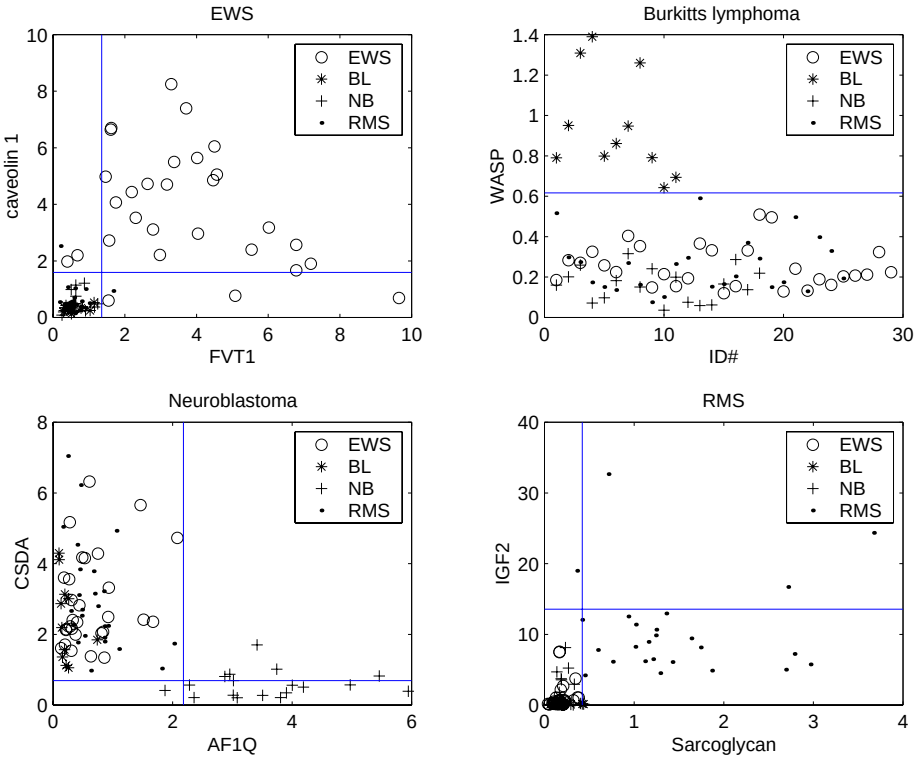

Fig. 2. Graphical representation of the seven rules discovered for the SRBCT data set. The lines mark thresholds for the expression levels of discovered genes. a) The two rules that sepa-rate EWS from all other cancer types. b) BL is perfectly separated from all other examples by one gene. c) NB is distinguished by high expression of AF1Q and low expression of CSDA. d) RMS is separated by high expression of sarcoglycan alpha and IGF2.

The rules discovered for EWS involve two genes: Caveolin 1 and follicular lymphoma variant translocation 1 (FVT1). Caveolin 1 encodes a protein that is known to play an important role in signal transduction and lipid transport. It has been associated with prostate cancer [20] and adenocarcinoma of the colon [21]. FVT1 has been proposed to be associated with follicular lymphoma by its close localization with Bcl-2 [22]. The single rule for Burkitt's lymphoma (BL) shows how this cancer type can be singled out based on a high expression level for the gene encoding the Wiskott-Aldrich syndrome protein (WASP) only. Likewise, neuroblastoma (NB) is separated from all the other tumor types by two independent rules involving the expression levels of the genes for AF1Q and cold shock domain protein A (CSDA). Specific expression of a fusion between the AF1Q gene and the mixed lineage leukemia (MLL) gene has been associated with leukemia [23], and this finding suggests an involvement in NB, possibly indicating that the fusion is present in NB as well. CSDA is a transcriptional regulator involved in stress response, and is believed to act as a repressor of human granulocyte-macrophage colony stimulating factor (GM-CSF) transcription [24]. Its down regulation may indicate an involvement in tumorogenesis in NB. Finally, RMS is separated from the other tumor types by the specific expression of sarcoglycan alpha (SGCA), a muscle specific protein associated with muscular dystrophy [25]. High expression of this gene is probably more indicative of the tissue origin of this tumor type than related to the molecular background of RMS. The second rule for RMS involves insulin-like growth factor II (IGF2), which is an already known oncogene associated with this cancer type [26]. Figure 2 shows a graphic representation of the coverage of all the rules.

4 Concluding Remarks

We have shown that rule induction methods are strong candidates for microarray analysis. One attractive property of this class of methods is that they do not only generate accurate prediction models, but also allow for straightforward interpretation of the reasons for the particular classification they make. Rule induction represents a whole class of methods, of which decision trees is perhaps the best known, but not necessarily the best-suited method for this particular type of task, as demonstrated in this study. Common for this class of methods is that they allow for a trade off between increased accuracy versus low complexity (i.e. high interpretability) of generated models. We have evaluated three of these methods, DAC-MDL, SAC, and Boosting for two different tumor tissue classification tasks. The classification accuracy was shown to be on level with the best current methods while exhibiting a much higher level of interpretability. Moreover, as opposed to many other methods employed for microarray data classification, rule induction methods can be applied in a straightforward manner to multi-class problems, such as the SRBCT data set.

From a histological point of view, the four tumor types represented in the SRBCT data set are rather similar. However, we found that the four classes can be distinguished quite easily due to a number of more or less obvious differences in their

respective expression patterns. From a molecular genetics point of view, the cancer types are thus rather disparate. The extensive literature regarding cancer-associated genes has allowed us to verify relatedness between genes and cancer described by a small set of rules.

Inspection of classification rules derived from numerical attributes typically gives the impression of the rules being very specific. However, since most rule sets generated only employ one split point for every gene used, the rules can easily be translated into qualitative conditions, *i.e.* whether a particular gene is relatively up- or down- regulated, when distinguishing between different classes, such as tumor types.

One major goal of gene expression array analysis is to discover new and interesting pathways describing causal dependencies underlying characteristic cellular behavior. We believe that the methods described in this paper are useful tools can contribute to a complete understanding of these pathways. We also believe that this approach can be applicable to neighbouring areas of gene expression array classification where phenotypes are to be correlated with global gene expression patterns.

References

1. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286, 531-537
2. Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,S.Y.D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.*, **96**, 6745-6750.
3. Khan,J., Wei,J.S., Rignér,M., Saal,L.H., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C. and Meltzer,P.S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 7, 673-679.
4. Ben-Dor,A., Bruhn,L., Friedman,N., Nachman,I., Schummer,M. and Yakhini,Z. (2000) Tissue classification with gene expression profiles. In Proceedings of the 4th International Conference on Computational Molecular Biology (RECOMB) Universal Academy Press, Tokyo.
5. Furey,T.S., Cristianini, N., Duffy,N., Bednarski,D.W., Schumm,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16, 906-914.
6. Keller,A.D., Schummer,M., Hood,L. and Ruzzo,W.L. (2000) Bayesian Classification of DNA Array Expression Data. Technical Report, University of Washington.
7. Zhang,H., Yu,C.Y., Singer,B. and Xiong,M. (2001) Recursive partitioning for tumor classification with gene expression microarray data. Proc. Natl. Acad. Sci., 98, 6730-6735
8. Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002) Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46 (1-3): 389 – 422.
9. Quinlan,J.R. (1986) Induction of decision trees. Machine Learning, 1, 81-106
10. Rivest,R.L. (1987) Learning Decision Lists. Machine Learning, 2, 229-246

11. Clark,P. and Niblett,T. (1989) The CN2 Induction Algorithm. Machine Learning, 3, 261-283
12. Boström,H. (2001) Virtual Predict User Manual. Virtual Genetics Laboratory AB, available from http://www.vglab.com
13. Boström,H. and Asker,L. (1999) Combining Divide-and-Conquer and Separate-and-Conquer for Efficient and Effective Rule Induction. Proc. of the Ninth International Workshop on Inductive Logic Programming, LNAI Series 1634, Springer, 33-43
14. Fayyad,U. and Irani,K. (1992) On the Handling of Continuous Valued Attributes in Decision Tree Generation. Machine Learning, 8, 87-102
15. Cestnik,B. and Bratko,I. (1991) On estimating probabilities in tree pruning. Proc. of the Fifth European Working Session on Learning, Springer, 151-163
16. Quinlan and Rivest (1989) "Inferring Decision Trees Using the Minimum Description Length Principle", Information and Computation 80(3) (1989) 227-248
17. Freund,Y. and Schapire,R.E. (1996) Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference, 148-156
18. Boström,H. (1995) Covering vs, Divide-and-Conquer for Top-Down Induction of Logic Programs. Proc. of the Fourteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann 1194-1200
19. Cohen,W.W. (1995) Fast Effective Rule Induction. Machine Learning: Proc. of the 12th International Conference, Morgan Kaufmann, 115-123
20. Tahir,S.A., Yang,G., Ebara,S., Timme,T.L., Satoh,T., Li,L., Goltsov,A., Ittmann,M., Morrisett,J.D. and Thompson,T.C. (2001) Secreted caveolin-1 stimulates cell survival/clonal growth and contributes to metastasis in androgen-insensitive prostate cancer. Cancer Res., 61, 3882-3885
21. Fine,S.W., Lisanti,M.P., Galbiati,F. and Li,M. (2001) Elevated expression of caveolin-1 in adenocarcinoma of the colon. Am. J. Clin. Pathol., 115, 719-724
22. Rimokh,R., Gadoux,M., Berthéas,M.F., Berger,F., Garoscio,M., Deléage,G., Germain,D. and Magaud,J.P. (1993) FVT-1, a novel human transcription unit affected by variant translocation t(2;18)(p11;q21) of follicular lymphoma. Blood, 81, 136-142
23. Busson-Le Coniat,M., Salomon-Nguyen,F., Hillion,J., Bernard,O.A. and Berger,R. (1999) MLL-AF1q fusion resulting from t(1;11) in acute leukemia. Leukemia, 13, 302-6
24. Coles,L.S., Diamond,P., Occhiodoro,F., Vadas,M.A. and Shannon,M.F. (1996) Cold shock domain proteins repress transcription from the GM-CSF promoter. Nucleic Acids Res., 24, 2311-2317
25. Duclos,F., Straub,V., Moore,S.A., Venzke,D.P., Hrstka,R.F., Crosbie,R.H., Durbeej,M., Lebakken,C.S., Ettinger,A.J., van der Meulen,J., Holt,K.H., Lim,L.E., Sanes,J.R., Davidson,B.L., Faulkner,J.A., Williamson,R. and Campbell,K.P. (1998) Progressive muscular dystrophy in alpha-sarcoglycan-deficient mice. J. Cell. Biol., 142, 1461-1471
26. El-Badry,O.M., Minniti,C., Kohn,E.C., Houghton,P.J., Daughaday,W.H. and Helman,L.J. (1990) Insulin-like growth factor II acts as an autocrine growth and motility factor in human rhabdomyosarcoma tumors. Cell Growth Differ., 1, 325-331