IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB12606

This paper appears in the book, *Emerging Trends and Challenges in Information Technology Management, Volume 1 and Volume 2* edited by Mehdi Khosrow-Pour © 2006, Idea Group Inc.

Classification and Rule Generation for Colon Tumor Gene Expression Data

Shawkat Ali & Pramila Gupta
Central Queensland University, Melbourne, VIC 3000, Australia, {s.ali, p.gupta}@mel.cqu.edu.au

ABSTRACT

Microarray genome studies discover the relationship between gene expression profiles and various diseases. This relationship generally introduces valuable quantitative information from genome profiles. The information facilitates drugs and therapeutics development to provide better treatments. In this paper we suggest that the statistical learning algorithm, Support Vector Machine (SVM) is a useful classification technique to classify genome profiles. Performance and usefulness of SVM is verified with colon tumor genome data. A comparison of SVM's performance is made with another popular decision trees based classification technique C5.0. SVM is found to be superior to C5.0 in performance. However, SVM lacks the rule extraction capability. We extract rules to identify the responsible tissues for colon tumor using C5.0. The rules could be used with SVM to reduce the size of microarrays in future.

1. INTRODUCTION

Genome research has become a very attractive area for the pattern recognition community as well as biological scientists. Gene expression profiles are being used increasingly in the development of efficient cancer diagnosis procedures [1]. Microarrays may be used to identify tumor genes and targets for therapeutic drugs by comparing gene expressions of normal and tumor tissues.

Among the various types of cancers, colon cancer is the second most common cause of cancer mortality in Western countries [18]. Therefore, based on the critical importance of the issue and availability of important data on colon cancer [3] this research addresses the colon tumor microarray classification problem. The colon tumor microarray dataset [3] is frequently used by researchers in assessing their gene classification methods.

From the beginning of cancer research, biologists have used the traditional microscopic technique to assess tumor behavior for cancer patients. The critically important requirement for cancer diagnosis and treatment, i.e., a precise prediction of tumors [2] is not possible with the traditional technique. Modern data mining techniques are being used increasingly by biologists to obtain proper tumor information from genome databases [20].

Among the existing techniques, supervised learning methods (SLM) [2] and unsupervised learning methods (USLM) [3] are the most popular for microarray genome analysis. One of the most frequently used unsupervised learning methods is the clustering of samples. Alon *et al*. [3] are the pioneers in presenting and analyzing colon data with clustering. The initial expression levels of about 6500 genes were measured for 62 samples including 40 tumor and 22 normal colon tissues. They selected 2000 genes for clustering purposes. The data was grouped into two clusters with 8 wrong instances; three normal tissues were assigned to the "tumor" cluster and five tumor tissues were assigned to the "normal" cluster.

While USLM are indispensable SLM have substantial advantages as they work with a predetermined classification framework as a supervisor. Given a set of training samples, the gene expression levels and class label of each sample are known. The goal of the SLM is to build a unique model/classifier from training data by properly setting up the function parameters. The classifier is then used to assign accurate labels to new microarray samples. Finally, by comparing the predicted class with actual class, the best model is selected for tumor prediction.

However, microarray data normally contains several thousand cells of gene information within a single matrix. This may cause problems in constructing an appropriate model by using traditional SLM. Some state of the art methods can handle this situation, for example Fujarewicz and Wiench [4] have proposed a new Recursive Feature Replacement (RFR) algorithm for finding suboptimal gene subsets for tumor and normal colon tissue classification. They compared the experimental results with other techniques such as Recursive Feature Elimination (RFE) [5], Neighbourhood Analysis (NA) [5], and the pure Sebestyen [5]. A rule based SLM, decision trees is one of the most frequently used techniques in data mining for searching pattern information from microarray data [23]. Chi. *et al*., use the C4.5 decision trees algorithm to generate rules for different gene data [27]. Yu, *et al*., use a rule based decision trees algorithm for correlation-based feature selectors to classify gene data [28].

A comparatively new statistical SLM, the Support Vector Machine (SVM) [1] is found to be a more effective and robust technique for microarray data analysis as compared to decision trees. The main advantages of this technique are: less computational complexity, the reproducibility and scalability of the obtained data and suitability to handle voluminous datasets [2].

Furey et al. [6] have used SVM to classify the colon dataset. The experiments were performed twice for the whole dataset of 2000 features and for the top 1000 features. In both stages the result of the leave-one-out cross-validation was six misclassifications (3 tumor and 3 normal ones). Furey et al. [6] compared SVM with other methods such as Parzen Window [21] and Fisher's Linear Discriminant [22]. The conclusion was that SVM significantly outperformed all other methods. They used the Fisher discrimination kernel for their SVM technique [6].

In this paper, we investigate two popular SVM techniques using Sequential Minimal Optimization (SMO) [16] and Quadratic Programming (QP) [17] to classify the colon tumor data. The SVM implementations from Weka [16] and OSU [29] are used. First, we map the nature of the colon tumor data to choose a proper kernel for SVM. Following our previous research [7] to select a suitable kernel for SVM classification, traditional polynomial and radial basis function (rbf) kernels are used. We then compare the performance of SVM with the decision trees based learning technique C5.0. The decision trees tool is available on [15]. Since the number of samples is less than one thousand, the 10 fold cross validation method is used to evaluate the performance of SVM and C5.0 [8]. The investigation reported in this paper differs from [6] in the use of more efficient optimizer QP, selection of kernel and evaluation with the 10 fold cross validation.

SVM is function-estimation based learning technique but it does not support building rules to classify an example. In contrast, the decision trees technique can extract rules from a gene by identifying responsible tissues characterizing it. We use three different C5.0 methods: Rule,

Boosting and Winnow to generate rules to identify the responsible tissues that cause colon tumor. Finally, we summarize the best rule for each C5.0 method to find out the responsible tissues for tumor and normal colon gene. The rules could be used with SVM to reduce the size of microarrays in future.

The rest of the paper is organized as follows: Section 2 describes the SVM and C5.0 techniques. The final observation from the comparative studies to classify the genome profile and the rules generated to identify the responsible tissues that cause tumor are presented in Section 3. Conclusions from this research are presented in Section 4.

2. SUPERVISED LEARNING METHODS: SVM AND C5.0

2.1 Rule Based Method – C5.0

Rules based learning methods, especially decision trees (also called classification trees or hierarchical classifiers), are a top-down induction approach, that have been studied with much interest by the machine learning community. C5.0 is an advanced version of the ID3 and C4.5 decision trees algorithms [9]. ID3 is the third series of the 'interactive dichotomizer' procedures. It can classify nominal datasets only. For real value attributes, it is first binned into intervals to form unordered nominal values. It does not consider any standard pruning procedure. By minimizing the ID3 limitation, Quinlan [9] introduced the C5.0 algorithm to solve classification problems. C5.0 works in three phases similar to some other supervised learning methods. First, the root node at the top of the tree considers all samples and passes them to branch nodes. The branch nodes generate rules for a group of samples based on their entropy measure. In this stage C5.0 constructs the whole tree by considering all attribute values and then finalizes the decision rules by pruning. It uses a heuristic approach for pruning based on the statistical significance of splits. After fixing the best rule, the branch nodes send the final target value to leaf nodes [9, 10]. Detailed mathematical formulations for C5.0 are provided in [9]. Decision trees become problematic and their performance in solving classification problems deteriorates with large tree sizes [11]. Decision trees with large sizes can also sometimes be difficult to understand.

Decision trees techniques adopt the Rule, Boosting and Winnow methods to perform the classification task. An important feature of C5.0 is its ability to generate classifiers called "rule based method" which consist of unordered collections of simple if-then rules. Rule based methods are generally easier to understand than trees themselves since each rule describes a specific context associated with a class [19]. Moreover, a rule based method, generated from a tree, enhances comprehensibility as it usually has fewer rules than the number of leaves of the tree. Finally, it is observed that rules are often more accurate predictors than decision trees. Another innovation based on the classical method incorporated in C5.0 is called "Boosting" that follows voting [24]. The idea is to generate several classifiers rather than just one; it could either be decision trees or rulesets. When a new instance is to be classified, each classifier votes for its predicted class and the votes are counted to determine the final class. The decision trees and rulesets constructed by C5.0 do not generally use all of the attributes at the same time. For example, text classification describes a passage by the words that appear in it, so that there is a separate attribute for each different word in a restricted dictionary. When there are numerous alternatives for each test in the tree or ruleset, it is likely that at least one of them will appear to provide valuable predictive information. It can be useful to pre-select a subset of the attributes that will be used to construct the final decision tree or ruleset. The C5.0 mechanism to do this is called "Winnowing" [15].

2.2 Statistical Learning Method - SVM

Statistical learning methods have received more attention from the pattern recognition community since the introduction of the Support Vector Machine (SVM) by Vapnik and his group in the mid 1990s [10]. SVM is the advanced version of the Generalized Portrait algorithm, which was developed in Russia in the late sixties [12]. SVM can be understood with its three working phases as is the case with C5.0. Cortes and Vapnik [26] mention the input or transformation phase, learning phase and decision phase for SVM. While C5.0 does not perform any significant work in the first phase SVM does it's most significant job of transforming data in this phase by using kernel mapping into a high dimensional feature space [19]. The kernel function can be polynomial, Gaussian, wavelet etc. The high dimensional space could theoretically be infinite, where linear discrimination is possible. In the learning phase, SVM starts to learn the data in the high dimensional feature space. It then minimizes the magnitude of the weight vectors to construct the optimal hyperplane [25]. In this stage SVM extracts the *support vectors* only. Based on the support vectors information SVM produces the final output function in the decision phase. Unlike C5.0, SVM does not consider all samples to construct the final decision function. Moreover, SVM always obtains the unique solution for the decision function unlike iterative approaches or pruning. Another feature of SVM is that it minimizes the structural risk rather than the empirical risk considered by most classical learning algorithms [10, 13]. Detailed mathematical formulations for SVM are provided in [13].

The statistical learning algorithm, SVM, has advantages over the well-established decision trees algorithms. It considers the dot product of the feature vectors to construct the optimal hyperplane rather than surface, clustering or interpolation as done in decision trees. This results in reduced probability of losing important information during modeling [14].

3. EXPERIMENTAL RESULTS: A COMPARISON AND RULES EXTRACTION

The data distribution nature of the colon tumor dataset is closely observed to select the proper kernel for SVM. We construct the histogram to map the data distribution as shown in Figure 1. The colon tumor data is highly positively declined, which suggests a trial-and-error approach to select a proper kernel for SVM classification [7].

3.1 A Comparison - Classification Performance

The major challenge of gene expression data is to handle the quite large number of genes in a single dataset. Classification of these datasets is very difficult with traditional learning algorithms because they contain a large number of genes (features) and thus the methods that search over subsets of features can be prohibitively expensive. We considered all features of the colon dataset in our experiment. Ten fold cross validation (10FCV) is a more appropriate method to get the prediction error in such situations [8]. We prefer 10FCV rather than the leave-one-out cross validation (LOOCV) method to measure the performance of SVM and C5.0.

We consider three different methods to construct decision trees including Rule based, Boosting and Winnow. The classification and computa-

Figure 1. Histogram based graphical presentation of colon tumor data. This figure shows data distribution positively declined rather than normal distribution.

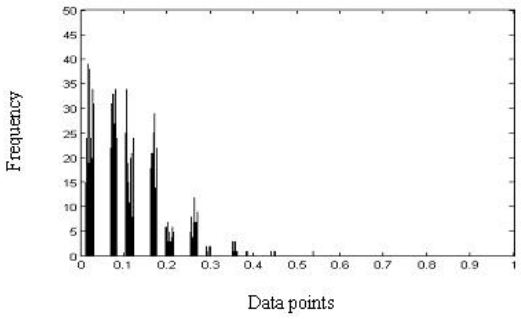

Table 1. Classification accuracy and computational time for colon tumor dataset with different decision trees and SVM methods

Algorithms	C5.0_ rule	C5.0_ boosting	C5.0_ winnow	SVM_ Smo_poly	SVM_ Smo_rbf	SVM_ qp_poly	SVM_ qp_rbf
Classification accuracy rate in %	90.20	85	85.70	85.48	79.03	99.90	99.9
Computational time in Sec.	2.3	10.2	0.9	0.56	0.52	5.09	5.02

tional performance for colon tumor data with the three decision trees methods is reported in Table 1.

The classification and computational performance for colon tumor data with the SMO and QP optimization methods for SVM using two classical kernel functions polynomial and rbf is also reported in Table 1. The suitable parameter for polynomial kernel and rbf kernel was 3 and 1.

We observe that the rule based decision trees method shows the best classification performance (90.20%) among all the trees based methods. The classification performances for the boosting and winnow methods are very close. The boosting method needs more computational time to finalize the most predictable decision trees. The winnow method is faster among all the decision tree methods. The SVM_smo_rbf technique shows the worst classification performance for colon tumor data. The SVM_smo_poly technique has shown classification performance very close to boosting and winnows decision trees methods. Both SVM_smo methods require shorter computational times as compared with other methods to classify the colon tumor dataset.

On the other hand, the experiment showed that the SVM_qp methods have the highest classification accuracy (99.9%). The two SVM_qp methods require longer computational time as compared with all other methods except for C5.0_boosting.

We closely observed the individual performance of SVM_smo and SVM_qp based on the confusion matrix. The best fold performance from 10FCV for both the SVM techniques is reported in Table 2.

We observe that SVM-qp shows 100% classification accuracy in both the cases of tumor and normal colon genes. However, SVM_smo shows better performance for classifying normal cells rather than tumor cells. The average performance of SVM_smo_poly is better than SVM_smo_rbf classifier.

3.2 Rules for Colon Tumor

This section outlines the extraction of rules that could help detect the responsible tissues causing colon tumor. We used all the different decision trees methods to extract the appropriate rules. The best rules are reported here.

Rules from Rule Based Method:

Rule 1: If attribute 1671 <= 56.91875 or If attribute 682 <= 107.4425 or If attribute 822 > 3307.498 then tumor.

Table 2. Confusion matrix based on the best fold colon tumor dataset classification performance for different SVM optimizations and kernels methods.

Algorithms	True Positive	False Negative	False Positive	True Negative		% Accuracy
SVM_qp_poly	1	0	0	1		100.00%
SVM_qp_rbf	1	0	0	1		100.00%
SVM_smo_poly	0.77	0.23	0.1	0.9		83.50%
SVM_smo_rbf	0.5	0.5	0.05	0.95		72.50%

Rule 2: If attribute 682 > 107.4425 and attribute 822 <= 3307.498 and attribute 1671 > 56.91875 then normal colon.

The accuracy of the rule is 100%.

Rules from Boosting Methods:

Rule 1: If attribute 1671 <= 56.91875 or If attribute 1671 > 56.91875 and gene822 > 3307.498 or If attribute 822 <= 3307.498 and attribute 1466 <= 24.90357 then tumor.

Rule 2: If attribute 1466 > 24.90357 then normal colon.

The accuracy of the rule is 100%.
Rules from Winnow Methods:

Rule 1: If attribute 249 <= 1627.27 then normal colon.

Rule 2: If attribute 249 > 1627.27 then tumor.

The accuracy of the rule is 100%.

We discovered three individual rules to identify the tumor and one single combined rule to recognize the normal colon with the rule based method. The parameter *m* value was 4 and the confidence interval was 80%. We extracted two separate combined rules with the boosting method; one to recognize the tumor and one single rule for normal colon. The parameter *m* value was 2 and the confidence interval was 85%. Finally we discovered two single rules to identify the normal colon and tumor with the winnow method. The parameter *m* value was 2 and the confidence interval was 90%. These rules showed 100% accuracy to classify the test dataset. Therefore all rules are acceptable to identify the responsible tissues for colon tumor using C5.0.

4.0 CONCLUSIONS

The present research addressed an important problem of colon tumor identification and provided a solution to classify a genome profile and identify the tissues that cause colon cancer. The experimental results indicate that SVM is able to classify the genome profile most efficiently and accurately. SVM with the single kernel rbf and the qp optimization method have shown equal or better performance as compared with C5.0. The decision trees algorithm C5.0 was able to discover proper rules to identify the responsible tissues for tumor. The rules have shown higher accuracy and performance, which is useful to verify new genome profile. The entropy based rule generation could be useful to deduct less significant features for the high dimensional gene expression data for future expansions to SVM. The research could be very useful and beneficial for medical practitioners and drug developers. This research could be explored further in future to study other types of cancer as well as different types of chronic diseases.

5.0 REFERENCES

[1] Brown, P.O. and Botstein, D., Exploring the new world of the genome with DNA microarrays, Nature Genetics Supplement, vol. 21, pp. 33-37, Jan. 1999.

[2] Kyung-Joong, K. and Sung-Bae, C. Prediction of colon cancer using an evolutionary neural network**,** *Neurocomputing*, v 61, n 1-4, pp. 361-379, October, 2004.

[3] Alon U., Barkai N., Notterman D. A., Gish K., Ybarra S., Mack D. and Levine A. J., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed

by oligonucleotide arrays. Proc. Natl. Acad. Sci., Vol. 96, pp. 6745–6750, 1999.

[4] Fujarewicz, K. and Wiench, M., Selecting differentially expressed genes for colon tumor classification**,** Int. J. Appl. Math. Comput. Sci., Vol. 13, No. 3, pp. 327-335, 2003.

[5] Nguyen D.V. and Rocke D.M., Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics*, Vol. 18, No. 1, pp. 39-50, 2002.

[6] Furey T. S., Christianini N., Duffy N., Bednarski D. W., Schummer M. and Haussler D., Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, Vol. 16, No. 10, pp. 906-914, 2000.

[7] Ali, S. and Smith, K. A. Automatic Parameter Selection for Polynomial Kernel, *Proceedings of the IEEE International Conference on Information Reuse and Integration*, USA, pp. 243-249, 2003.

[8] Henery, R. J. 'Methods of comparison', in D. Michie, D. J. Spiegelhalter and C. C. Taylor (eds.), Machine Learning, Neural and Statistical Classification, *Ellis Horwood Limited*, New York Chapter 7, 1994.

[9] R. Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufman Publishers, San Mateo, CA, 1993.

[10] Duin, R. P. W. A note on comparing classifier, *Pattern Recognition Letters*, vol. 1, pp. 529-536, 1996.

[11] Duda, R. O., Hart, P. E. and Stork, D. G. Pattern Classification, *Wiley*, New Yourk, 2nd edition, 2001.

[12] Smola A. and Schölkopf, B. A tutorial on support vector regression, *Statistics and Computing*, 2003.

[13] Vapnik, V. N. An overview of statistical learning theory, *IEEE Transaction on Neural Networks*, Vol.10, No.5, September 1999.

[14] Lodhi, H., Saunders, C., Cristianini, N., Watkins, C., and Shawe-Taylor, J., Text classification using string kernels, Appeared in *Journal of Machine Learning Research*, 2003.

[15] Quinlan, R. See5: An informal tutorial, http://www.rulequest.com/see5-win.html, 2005.

[16] Witten, I. H. and Frank, E., Data Mining: practical machine learning tool and technique with Java implementation, *Morgan Kaufmann*, San Francisco, 2000.

[17] Vapnik, V. N. Statistical learning theory, *John Wiley & Sons*, Inc., 2000.

[18] Chung-Faye, G. A., Kerr, D. J., Young, L. S., and Searle, P. F., Gene therapy strategies for colon cancer, Mol. Med. Today 6 (2), pp. 82–87, 2000.

[19] Ali, S. and Smith, K. A., On Learning Algorithm Selection for Classification, Accepted by *International Journal of Applied Soft Computing*, Elsevier Science, 2004.

[20] Huang, X. and Pan, W. Linear regression and two-class classification with gene expression data. Bioinformatics 19(16), pp. 2072-2078, 2003.

[21] Babich, G. A., and Camps, O. I., *Weighted Parzen windows for pattern classification*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18(5), pp.567-570, 1996.

[22] Cooke, T., *Two variations on Fisher's linear discriminant for pattern recognition* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24(2), pp. 268-273, 2002.

[23] Shu-Tzu, T. and Chao-Tung, Y., Decision tree construction for data mining on grid computing, *Proceedings of IEEE International Conference on e-Technology, e-Commerce and e-Service*, pp. 441-447, 2004.

[24] Grossmann, E., AdaTree: Boosting a Weak Classifier into a Decision Tree, Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Workshop, pp. 105 – 105, 2004.

[25] Yan, G., Wang, H., Ding, G. and Lin, L., Augmented Lagrange multiplier based fuzzy evolutionary algorithm and application for constrained optimization, Proceedings of the 4th IEEE World Intelligent Control and Automation, vol. 3, pp. 1774-1778, 2002.

[26] Cortes C. and Vapnik. V., Support-vector network. Machine Learning, vol.20, pp.273— 297, 1995.

[27] Chi Z., Weimin X., Tirpak, T. M. and Nelson, P. C. Evolving accurate and compact classification rules with gene expression programming, *IEEE Transactions on Evolutionary Computation*, vol. 7(6), pp. 519-531, 2003.

[28] Yu, W., Igor, V. T., Mark, A. H., Eibe, F., Axel, F., Klaus F. X. M. and Hans W. M., Gene selection from microarray data for cancer classification-a machine learning approach, *Computational Biology and Chemistry*, Vol. 29(1), pp. 37-46, 2005.

[29] Ma, J., Zhao, Y. and Anhalt, S., OSU SVM Classifier Matlab Toolbox, http://sourceforge.net/projects/svm, 2005.