Applying Gaussian Distribution-Dependent Criteria to Decision Trees for High-Dimensional Microarray Data

Raymond Wan[1] rwan@kuicr.kyoto-u.ac.jp
Ichigaku Takigawa[1] takigawa@kuicr.kyoto-u.ac.jp
Hiroshi Mamitsuka[1] mami@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan

Keywords: decision trees, continuous values, microarray data, splitting criteria

1 Introduction

Microarray data presents an interesting problem to machine learning algorithms due to their high-dimension and small number of samples. Since algorithms such as support vector machines (SVM) typically achieve high prediction accuracies, other methods garner a relatively small amount of attention even though they possess other characteristics which are useful for microarray analysis.

In this poster, we summarize our earlier work where we combine Gaussian distribution-dependent splitting criteria with the gain ratio found in most decision tree implementations [2]. The aim of this work is to improve the prediction accuracy of decision trees for microarray data. Experiments with seven microarray data sets demonstrate an overall improvement over the original.

2 Background

We denote a microarray data set as $\mathcal{D}$ with m samples (examples) and n genes (attributes). Each example is associated with a class such as "tumor" or "normal". When applied to microarray data, decision trees are used to find the gene(s) which split $\mathcal{D}$ into subsets of uniform classes. Thus, instead of merely providing a fingerprint for classifying future data sets, decision trees also state which genes are used to form the fingerprint. Figure 1 shows a sample decision tree.

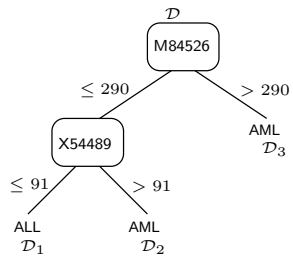

Figure 1: Decision tree.

The basis of most decision tree algorithms such as C4.5 Release 8 [1] and WEKA J4.8 [3] is the gain ratio. The gain ratio scores each attribute based on how well it splits the data set and selects a value to split on (i.e., 290 in Figure 1).

However, one problem with the gain ratio is that it is derived purely from the class distribution in the subsets. That is, the distribution of the values for the splitting attribute is ignored. If several genes yield the same gain ratio, then a decision tree implementation must pick one gene arbitrarily. In this work, we augment the gain ratio of one decision tree implementation (C4.5) with two Gaussian distribution-dependant splitting criteria: Student's t-test and Kullback-Leibler divergence.

3 Method

The two new splitting criteria depend on the assumption that the expression levels of each gene in a microarray data set follow two normal distributions – one for each class. For the Student's t-test, we elected to use a two-tailed, two sample t-test for unequal variance. The Kullback-Leibler divergence

tests the difference between two probability distributions. As a result, unlike the Student's t-test, it takes into consideration the mean and the variance instead of the mean alone.

Gain ratio and these two criteria yield three different values for each attribute. We convert these values into ranks so that the better a data set is separated by an attribute, the higher the rank of the attribute. The ranks are then weighted using three parameters ($\{\alpha_1, \alpha_2, \alpha_3\}$) which sum to 1.

4 Results

We compared our modifications to C4.5 to the original C4.5 and two algorithms within WEKA: Naive Bayes and SVM. Seven data sets of two classes each were used [2], ranging in size from 62 to 280 samples and 2,000 to 16,063 genes. Experiments were conducted within the framework of inverse cross-validation and normal cross-validation. In normal cross-validation, the training set is larger than the test set. Inverse cross-validation reverses this and allows us to examine the effectiveness of our method when the number of training samples is small. Each program was applied 50 times and then averaged across the 7 data files to create the table below.

	Training size / Folds	$\{1,0,0\}$	$\{\frac{1}{3},\frac{1}{3},\frac{1}{3}\}$	$\{0,\frac{1}{2},\frac{1}{2}\}$	Naive Bayes	SVM
Inverse CV	10 samples	60.5	+5.4	+5.8	+7.0	+12.2
	20 samples	70.4	+3.1	+2.8	+2.5	+10.5
CV	5 folds	78.7	+1.1	+1.8	-1.7	+10.5
	10 folds	79.0	+0.6	+2.1	-2.1	+10.6

The second column reports the number of samples in the training set or the number of folds created. The third column of this table shows the baseline C4.5 as an accuracy percentage. The remaining columns show the relative difference in accuracy. If all three criteria are given equal weight or if equal weight is given only to the two Gaussian distribution-dependent splitting criteria, improvements over the baselines are attained. Naive Bayes and SVM both outperform decision trees when the training set is small, but Naive Bayes is worse than the baseline for normal cross validation.

5 Discussions

Our modifications to C4.5 improve over the baseline for both types of cross-validations. While Naive Bayes and SVM perform better than either decision tree implementations, decision trees have the advantage of stating which genes are used to separate $\mathcal{D}$. This important point was not emphasized in this work, but we plan to illustrate this in our future work. Coupled with other techniques such as decision forests [4], decision trees can become an alternative microarray analysis tool.

Acknowledgements R. W. was supported by a Japan Society for the Promotion of Science (JSPS) fellowship.

References

[1] J. R. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996. Source available from: `http://www.rulequest.com/Personal/`.

[2] R. Wan, I. Takigawa, and H. Mamitsuka. Applying Gaussian distribution-dependent criteria to decision trees for high-dimensional microarray data. In M. M. Dalkilic, S. Kim, and J. Yang, editors, *Proc. Data Mining in Bioinformatics (VDMB) Workshop at the 32nd International Conference on Very Large Data Bases*, volume 4316 of *LNBI*, pages 40–49. Springer-Verlag, September 2006.

[3] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers, second edition, 2005.

[4] H. Zhang, C.-Y. Yu, and B. Singer. Cell and tumor classification using gene expression data: Construction of forests. *Proc. National Academy of Sciences USA*, 100(7):4168–4172, April 2003.