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

Raymond Wan rwan@kuicr.kyoto-u.ac.jp

Ichigaku Takigawa takigawa@kuicr.kyoto-u.ac.jp

Hiroshi Mamitsuka mami@kuicr.kyoto-u.ac.jp

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

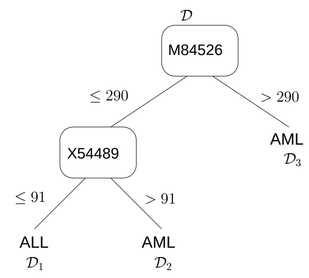
Abstract

Microarray data have a small number of samples and a large number of values collected per sample. While machine learning tools such as Support Vector Machines (SVM) work well for microarray data due to its higher classification accuracies, decision trees offer other advantages such as easily understood output. We report on our earlier work with decision trees which:

- *assumed that expression levels of each gene follows a Gaussian distribution;*
- *modified the decision tree splitting criteria to improve prediction accuracy for small data sets typical of microarrays; and*
- *showed that our modified splitting criteria improves over the original with seven microarray data sets*

1. Introduction

A decision tree algorithm recursively partitions an input data set $\mathcal{D}$ into smaller subsets $\mathcal{D}_1, ..., \mathcal{D}_k$. Central to constructing a decision tree is a splitting criterion which:



- adds a new node to the tree and
- for continuous values (such as microarray expression levels), provides the value v to split on (the cut point).

In general, most implementations create two children so that one branch has the samples which satisfy $A_k \leq v$; the remaining ones satisfy $A_k > v$. This is illustrated above where $(A_{M84526} \leq 290$ and $A_{X54489} \leq 91)$ for the subset $\mathcal{D}_1$.
A microarray data set is a table of:

- m samples (examples),
- n genes (attributes) for each sample, and
- a class for each sample (i.e., "tumor" or "normal").

In this work, we focus on the C4.5 Release 8 implementation of decision tree induction. The splitting criterion used is the *gain ratio*, which selects the attribute which creates subsets with the most uniformity in class. This method fails in one particular scenario: when the gain ratio of multiple genes *tie*. In this case, C4.5 arbitrarily selects the attribute on the furthest left. In the example table above of 5 examples and 2 attributes, attribute x_1 would be chosen even though x_2 is clearly the better choice.

	x_1	x_2	C
S_1	1	1	tumor
S_2	2	2	tumor
S_3	3	3	tumor
S_4	4	10	normal
S_5	5	11	normal

2. Method

The gain ratio scores each attribute based on how well the classes would be split. We augment to the gain ratio two other splitting criteria which assume that the expression levels of each attribute follow two separate Gaussian distributions representing the two classes.

2.1 Scoring

The additional criteria are:

1. the Student's t-test (two-tailed, two sample with unequal variance) and
2. the Kullback-Leibler (KL) divergence test.

The Student's t-test is well-known for reporting the difference in means between two distributions. In contrast, the KL divergence places importance in both the mean and the variance.

2.2 Ranking

The scores from all three metrics are combined to create a unified score which is then used to select the best splitting attribute. This is done through ranking as follows:

1. The scores for each metric is converted to ranks numbered from n down to 1.
2. Three parameters ($\{\alpha_1, \alpha_2, \alpha_3\}$) are used to weight the ranks of each attribute to get a final score.

3. Results

We compared our C4.5 modifications to the original using seven data sets. In Table 1 (left), we employed inverse cross-validation where the training size is smaller than the testing size. In Table 1 (right), we used normal cross-validation. Results are shown based on the training size for inverse cross-validation and the number of folds for normal cross validation. The results are:

- absolute accuracies for $\{1, 0, 0\}$ only;
- relative accuracies to the baseline for other cases;
- accuracies are averaged across 50 trials;
- below each accuracy is the *t*-statistic against the baseline; and
- *t*-statistic values highlighted in red are significant at the 99% confidence level (≥ 2.6800).

In general, our method performs best when the number of microarray data sets available is *small*. As the number of data sets increase, then the difference against the original C4.5 is negligible. Note that our changes to C4.5 never yield a *statistically-significant decrease* in accuracy.

In additional experiments (not shown) we determined that SVM (using the default parameters of the WEKA implementation) achieves accuracies which are 10% - 12% better than our method. However, the usefulness of decision trees lie in the ease of interpretation, which we plan to exploit in the future.

4. Future Work

We would like to:

- generalize our method for >2 classes,
- further evaluate the effectiveness of our method by examining the genes that it selects,
- combine our method with other methods such as decision forests, and
- relax the assumption of gene (attribute) independence and modify our method accordingly.

$\{\alpha_1, \alpha_2, \alpha_3\}$	Colon	Leukemia	Lung	CNS	Multiple	Lymph	Prostate	Avg
			10 examples for training					
$\{1, 0, 0\}$	59.5	58.8	75.8	52.6	59.3	65.5	51.6	60.5
$\{0, 1, 0\}$	+1.4	+10.2	+4.6	+2.2	+2.0	+3.7	+12.1	+5.2
	(1.81)	(13.20)	(9.27)	(3.94)	(8.79)	(5.27)	(23.26)	
$\{0, 0, 1\}$	+0.8	+6.1	+6.9	+4.9	-0.7	+3.1	+3.4	+3.5
	(0.89)	(10.16)	(17.50)	(6.09)	(-2.94)	(4.04)	(6.92)	
$\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$	+2.3	+11.4	+5.3	+3.4	+1.5	+6.6	+7.6	+5.4
	(3.23)	(16.26)	(11.94)	(5.43)	(6.00)	(9.15)	(16.87)	
$\{0, \frac{1}{2}, \frac{1}{2}\}$	+2.0	+11.7	+5.6	+3.6	+1.7	+7.0	+9.4	+5.8
	(2.74)	(15.09)	(13.70)	(5.60)	(6.76)	(9.59)	(17.84)	
			20 examples for training					
$\{1, 0, 0\}$	68.2	82.7	81.1	55.2	63.3	74.4	68.0	70.4
$\{0, 1, 0\}$	-2.0	-3.0	+5.4	+0.5	+0.5	+3.8	+5.8	+1.6
	(-2.33)	(-3.11)	(11.92)	(0.65)	(1.55)	(5.28)	(7.55)	
$\{0, 0, 1\}$	-2.0	-6.0	+3.9	+1.9	-1.8	-2.2	-7.4	-1.9
	(-1.87)	(-5.70)	(9.13)	(2.29)	(-6.24)	(-2.58)	(-8.59)	
$\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$	+0.9	+2.5	+6.8	+0.6	+1.0	+4.3	+5.4	+3.1
	(0.93)	(2.51)	(15.72)	(0.68)	(3.79)	(5.96)	(7.97)	
$\{0, \frac{1}{2}, \frac{1}{2}\}$	-0.1	+1.2	+7.1	+0.8	+1.1	+4.3	+5.3	+2.8
	(-0.13)	(1.32)	(16.88)	(0.83)	(3.85)	(5.38)	(8.04)	

$\{\alpha_1, \alpha_2, \alpha_3\}$	Colon	Leukemia	Lung	CNS	Multiple	Lymph	Prostate	Avg
			5 folds					
$\{1, 0, 0\}$	76.5	84.8	93.4	56.2	78.5	79.2	82.4	78.7
$\{0, 1, 0\}$	-2.5	+2.5	+2.1	+0.8	-1.8	+11.4	+1.9	+2.1
	(-2.42)	(3.96)	(7.78)	(0.66)	(-3.96)	(15.43)	(2.90)	
$\{0, 0, 1\}$	-6.6	+2.9	-1.0	+2.5	-3.3	-5.1	-10.3	-3.0
	(-7.12)	(4.64)	(-2.70)	(2.07)	(-7.26)	(-5.91)	(-13.72)	
$\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$	-3.6	+3.6	+1.9	+4.0	-0.1	+2.5	-0.4	+1.1
	(-4.58)	(5.95)	(8.21)	(3.63)	(-0.13)	(3.62)	(-0.73)	
$\{0, \frac{1}{2}, \frac{1}{2}\}$	-3.6	+6.0	+3.6	+2.0	-0.2	+3.4	+1.2	+1.8
	(-4.73)	(10.41)	(14.41)	(1.99)	(-0.51)	(5.25)	(1.95)	
			10 folds					
$\{1, 0, 0\}$	77.5	80.4	93.7	58.3	78.7	81.1	83.3	79.0
$\{0, 1, 0\}$	+0.6	+6.4	+2.2	-2.8	-2.0	+12.1	+0.9	+2.5
	(0.68)	(10.20)	(9.44)	(-2.69)	(-5.14)	(21.72)	(1.52)	
$\{0, 0, 1\}$	-5.2	+10.1	-1.2	+3.5	-3.2	-6.0	-11.9	-2.0
	(-6.72)	(16.97)	(-4.03)	(3.73)	(-6.94)	(-9.59)	(-18.23)	
$\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$	-4.9	+7.8	+1.7	+1.5	-0.3	+0.3	-1.7	+0.6
	(-6.38)	(12.69)	(6.81)	(1.39)	(-0.88)	(0.40)	(-3.52)	
$\{0, \frac{1}{2}, \frac{1}{2}\}$	-2.7	+11.3	+3.7	-0.1	+0.3	+1.5	+0.6	+2.1
	(-3.89)	(23.32)	(16.14)	(-0.07)	(0.76)	(2.55)	(1.14)	

Table 1: Experiments with inverse cross-validation (left) and normal cross-validation (right). The size of the seven data sets are Colon (62), Leukemia (72), Lung (181), CNS (60), Multiple (280), Lymph (77), and Prostate (102).

Note: This work has appeared as 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.

The 17th International Conference on Genome Informatics (**GIW 2006**), December, 18-20, 2006, Yokohama, Japan