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

Raymond Wan Ichigaku Takigawa
Hiroshi Mamitsuka
{rwan,takigawa,mami}@kuicr.kyoto-u.ac.jp

Bioinformatics Centre, Kyoto University, Japan

September 11, 2006

Overview

Decision trees are supervised machine learning techniques used for classification. However, they are rarely used for classifying high-throughput data sets such as microarrays due to its (generally) low accuracy compared with other techniques. However, decision trees possess other traits that make it potentially useful for microarray classification.

In this work, we identify one problem with applying decision trees to microarray data and show how it can be improved using Gaussian distribution-dependent criteria. Results show statistically-significant higher accuracies compared to the original algorithm was achieved in most cases.

Outline

Background

Microarray Data

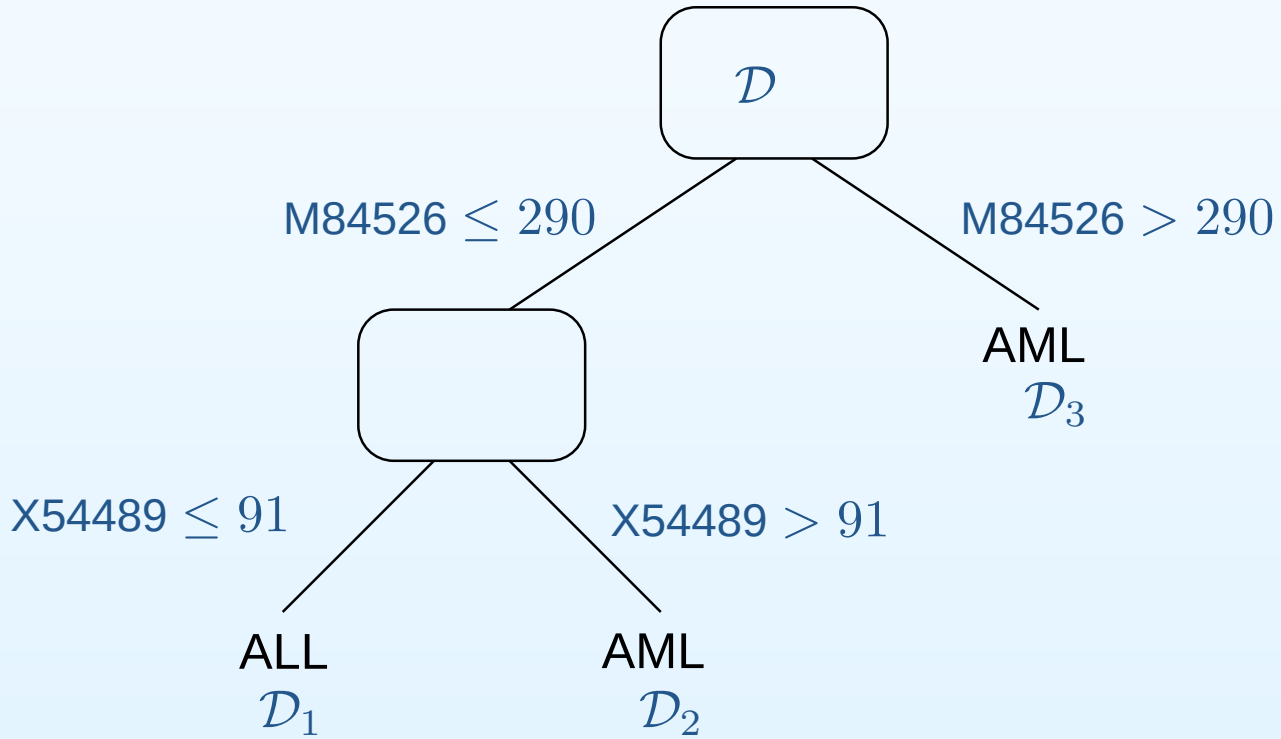
A microarray data set $\mathcal{D}$:

	Gene (Attribute, Feature)							Disease (Class)
Sample (Example)	A_1	A_2	A_3	A_4	A_5	...	A_n	
S_1								tumor
S_2								normal
S_3								tumor
S_4								tumor
S_5								normal
...								tumor
S_m								normal

The aim of microarray classification is to create a model which divides the samples into the two classes and to reuse the model as a "fingerprint".

Decision Trees (1)

Decision tree algorithms split $\mathcal{D}$ into subsets $\{\mathcal{D}_1,...,\mathcal{D}_k\}$ such that each subset is uniform in class. This is usually done in a top-down fashion. Each node represents an attribute and (for continuous values) a cut point.



Decision Trees (2)

In this work, we focus on C4.5 Release 8 [Quinlan, 1996]. Other implementations exist (such as the J4.8 implementation in WEKA [Witten and Frank, 2005]) and operate in a similar way.

C4.5 uses the gain ratio to select the splitting attribute for each node.

- Each attribute is examined one-by-one.
- For continuous values, each attribute's values are sorted and every possible cut point is tested.

The attribute associated with the cut point that gives the best gain ratio is chosen.

Decision Trees (3)

But, if the number of genes (thousands) is far larger than the number of samples (10 or 20), then it is possible that multiple genes will yield the same gain ratio.

Which one should be chosen?

Decision Trees - Key Point

Problem: Decision tree implementations such as C4.5 break ties by selecting the "first" attribute.

Idea: Apply two Gaussian distribution-dependent criteria to break these ties.

Assume: Each gene is independent of each other and each gene A_k is made up of two normal distributions, one for each class.

	A_k	Class
$\mathcal{N}_1(\mu_1, \sigma_1)$	v_{1k}	tumor
	v_{2k}	tumor
	v_{3k}	tumor
$\mathcal{N}_2(\mu_2, \sigma_2)$	v_{4k}	normal
	v_{5k}	normal
	v_{6k}	normal

Normality of Microarrays

Normality assumption made based on previous work by others, including:

- Apply Student's *t*-test to validate results from clustering [Alon et al., 1999].
- Select genes from a microarray data set [Yeung et al., 2005].
- Others have suggested that the normality assumption is acceptable if the logarithms of the measurement values is used [Wit and McClure, 2004].
- Normality can be assumed for untransformed Affymetrix data [Giles and Kipling, 2003].

Decision Trees for Microarrays

Previous work:

- Zhang et al. [2001] used decision trees to classify microarray data. The splitting criterion was similar to the gain ratio but was further refined using cross-validation.
- They later built a forest of decision trees and took the top two levels of each to form a fingerprint for the tumor type [Zhang et al., 2003].
- The RankGene software [Su et al., 2003] includes several algorithms, including the gain ratio (information gain) and t-test to rank genes.

Unlike this previous work, we focus on the decision tree algorithm itself rather than how to apply it. Furthermore, instead of using the algorithms separately, we integrate t-test into the decision tree algorithm.

Method

Overview

Instead of selecting attributes based on the gain ratio alone, we augment it with two Gaussian distribution-dependent criteria. The two additional criteria are Student's *t*-test and Kullback-Leibler Divergence. Thus, for each gene A_k, we get three scores: $gain_s$, $ttest_s$, and KL_s.

We convert the scores into ranks: $gain_r$, $ttest_r$, and KL_r, respectively, and use these ranks to get a final score for each attribute. The gene with the best score is selected.

No changes to the gain ratio; each attribute's gain ratio is also its score.

Student's *t*-test

Several forms of Student's *t*-test exist. We employed a two-tailed, two sample *t*-test for unequal variance. The *t*-test score is simply its *t*-statistic.

Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence calculates the difference between two distributions.

$$D(p||q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} dx \,. \tag{1}$$

If both are normally distributed, their KL divergence can be simplified to:

$$D(p||q) = \frac{1}{2\sigma_2^2} (\mu_1 - \mu_2)^2 + \frac{(\sigma_1 - \sigma_2)(\sigma_1 + \sigma_2)}{2\sigma_2^2} + \log \left(\frac{\sigma_2}{\sigma_1} \right) . \tag{2}$$

Since the KL divergence is non-symmetric, we calculate KL_s as half of $D(p||q) + D(q||p)$ [Jeffreys, 1946].

Combining the Ranks

The ranks, $gain_r$, $ttest_r$, and KL_r, are assigned from n down to 1. Tied scores are assigned the same rank.

After ranking each attribute according to its gain ratio, *t*-statistic, and KL divergence, we introduce three parameters ($\{\alpha_1, \alpha_2, \alpha_3\}$) which weight the sum of the ranks:

$$\text{score}(A_k) = \alpha_1 \times \text{gain}_r + \alpha_2 \times \text{ttest}_r + \alpha_3 \times \text{KL}_r$$
$$\text{such that } \sum_{i=1}^{3} \alpha_i = 1. \tag{3}$$

Brief Example

Sample data set

Example data set with five examples and two attributes based on two classes, A and B:

	A_1	A_2	C
S_1	1	1	A
S_2	2	2	A
S_3	3	3	A
S_4	4	10	B
S_5	5	11	B

Sample calculation

Calculation of the three measures for attributes A_1 and A_2:

A_k	$gain_s$	$gain_r$	$ttest_s$	$ttest_r$	KL_s	KL_r
A_1	0.794	2	3.273	1	16.75	1
A_2	0.794	2	11.129	2	181.75	2

Combining the measures using two different sets of parameters:

$\{\alpha_1, \alpha_2, \alpha_3\}$	$score(A_1)$	$score(A_2)$
$1, 0, 0$	2.00	2.00
$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	1.33	2.00

Results

Experiments

Experiments were performed using both synthetic and real data sets.

We tested the real data sets in two ways. As the purpose of our work is to handle small data sets, we demonstrate our method with *inverse cross validation*, where the training size is smaller than the test size. This has the added advantage of providing us with ample examples for testing.

We also show our method using normal cross validation.

Comparisons with several algorithms from WEKA are also shown.

Real Data Sets

Name	Classes		Number of samples				Number of genes
	Class 1	Class 2	Proportion			Total	
Colon	Tumor	Normal	40	:	22	(62)	2,000
Leukemia	ALL	AML	47	:	25	(72)	7,129
Lung	ADCA	MPM	150	:	31	(181)	12,533
CNS	Success	Fail	39	:	21	(60)	7,129
Multiple	Tumor	Normal	190	:	90	(280)	16,063
Lymph	DLBCL	FL	58	:	19	(77)	7,129
Prostate	Tumor	Normal	52	:	50	(102)	12,600

Inverse Cross-Validation

	$\alpha_1, \alpha_2, \alpha_3$	Colon	Leukemia	Lung	CNS	Multiple	Lymph	Prostate	Avg
10	$\{1, 0, 0\}$	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	$\{1, 0, 0\}$	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)	

Each value is the accuracy after performing 50 repetitions. The corresponding two-tailed t-statistic is 2.6800 for a 99% confidence interval.

Cross-Validation

	$\alpha_1, \alpha_2, \alpha_3$	Colon	Leukemia	Lung	CNS	Multiple	Lymph	Prostate	Avg
5	$\{1, 0, 0\}$	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	$\{1, 0, 0\}$	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)	

Comparison with SVM/NB

	Training size / Folds	$\{1,0,0\}$	$\{\frac{1}{3},\frac{1}{3},\frac{1}{3}\}$	$\{0,\frac{1}{2},\frac{1}{2}\}$
Inverse CV	10	60.5	+5.4	+5.8
	20	70.4	+3.1	+2.8
CV	5	78.7	+1.1	+1.8
	10	79.0	+0.6	+2.1

	Training size / Folds	$\{1,0,0\}$	J4.8	Naive Bayes	SVM
Inverse CV	10	60.5	+0.1	+7.0	+12.2
	20	70.4	+0.2	+2.5	+10.5
CV	5	78.7	+0.0	-1.7	+10.5
	10	79.0	+0.3	-2.1	+10.6

Return to Weka page

Conclusion

Summary

We embedded two Gaussian distribution-dependent criteria within C4.5 Release 8 to improve its accuracy for high-dimensional data sets such as microarray data. Accuracy is not as good as algorithms like SVM, however, decision trees have inherent benefits such as being able to deal with non-numerical values and missing values.

In the future, we would like to:

- Consider combining our method with the decision forests of Zhang et al. [2003].
- Use the genes identified by our system to reveal some biological information from the microarray data.

Comparison

Characteristic	DT	SVM	Neural Nets	k-NN
Handling data of mixed type	●	○	○	○
Handling of missing values	●	○	○	●
Robustness to outliers	●	○	○	●
Insensitive to monotone transformations	●	○	○	○
Scalability (large N)	●	○	○	○
Dealing with irrelevant inputs	●	○	○	○
Linear combinations of features	○	●	●	◐
Interpretability	◐	○	○	○
Predictive power	○	●	●	●

Table 1: Good: ●; Fair: ◐; Poor: ○ (Taken from “Decision Tree Methods in Pharmaceutical Research” [Blower and Cross, 2006], which they adapted from Hastie et al. [2001].)

Acknowledgements

Motivation

Initial motivation for this project began through discussions with Dr. Craig Wheelock.

References

References (1)

U. Alon et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. National Academy of Sciences USA*, 96 (12):6745–6750, June 1999. Data: `http://microarray.princeton.edu/oncology/affydata/index.html`

P. E. Blower and K. P. Cross. Decision tree methods in pharmaceutical research. *Current Topics in Medicinal Chemistry*, 6 (1):31–39, 2006

P. J. Giles and D. Kipling. Normality of oligonucleotide microarray data and implications for parametric statistical analysis. *Bioinformatics*, 19(17):2254–2262, 2003

References (2)

T. R. Golub et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286 (5439):531–537, October 1999. Data: `http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`

G. J. Gordon et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, 62(17):4963–4967, September 2002. Data: `http://www.chestsurg.org/publications/2002-microarray.aspx`

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. Royal Society of London (A)*, 186:453–461, 1946

References (3)

S. L. Pomeroy et al. Prediction of central nervous system embryonal tumour outcome based on gene expresion. *Nature*, 415(6870): 436–442, January 2002. Data: `http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`

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/`

S. Ramaswamy et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. National Academy of Sciences USA*, 98(26):15149–15154, December 2001. Data: `http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`

D. Singh et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, March 2002. Data: `http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`

References (4)

Y. Su, T. M. Murali, V. Pavlovic, M. Schaffer, and S. Kasif. RankGene: identification of diagnostic genes based on expression data. *Bioinformatics*, 19(12):1578–1579, 2003. Software available from `http://genomics10.bu.edu/yangsu/rankgene/`

M. A. Shipp et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, January 2002. Data: `http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`

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

E. Wit and J. McClure. *Statistics for Microarrays*. John Wiley & Sons Ltd., 2004

References (5)

K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, 21(10): 2394–2402, 2005

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

H. Zhang, C.-Y. Yu, B. Singer, and M. Xiong. Recursive partitioning for tumor classification with gene expression microarray data. *Proc. National Academy of Sciences USA*, 98(12):6730–6735, June 2001

Appendix

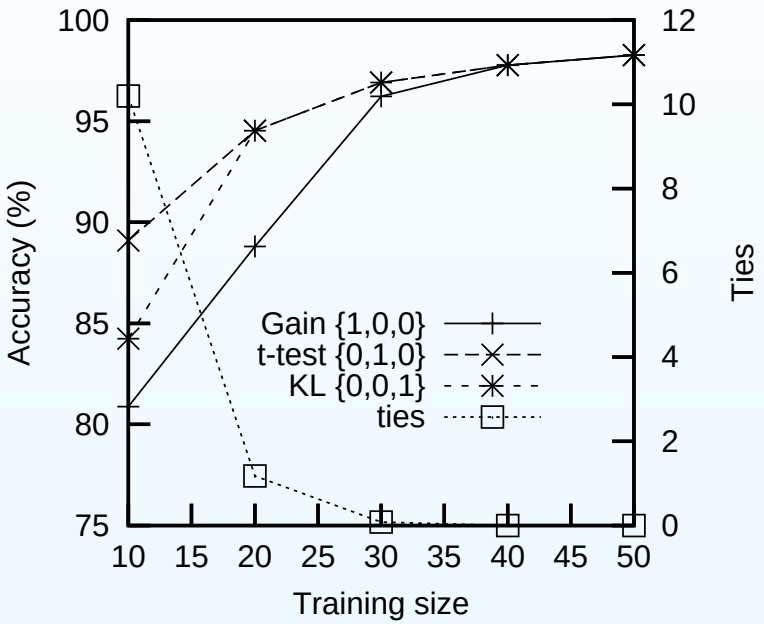
Artificial Data (1)

In order to simulate the ideal scenario for our method, we created data sets with two types of attributes:

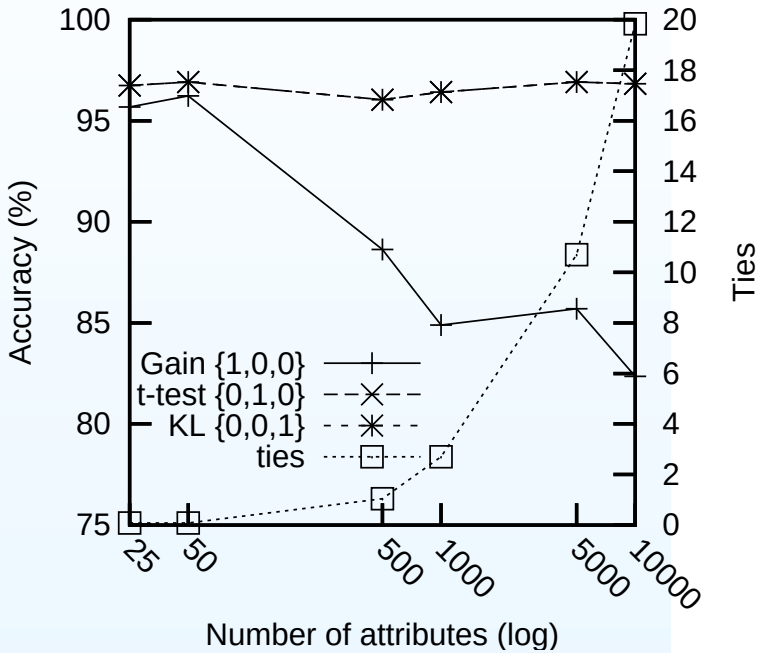
Type	Distributions		Number
	C_1	C_2	
1	$\mathcal{N}_1(0,1)$	$\mathcal{N}_2(1.5,1)$	25 to 10,000
2	$\mathcal{N}_1(0,1)$	$\mathcal{N}_2(100,1)$	1 (at the end)

Recall that C4.5 will select the first attribute with the best gain ratio, even if many are tied.

Artificial Data (2)



(1)



(2)

1. Number of attributes with overlapping distributions: 50.
2. Number of training examples: 30.

Return to experiments

Inverse Cross-Validation - Weka

	WEKA	Colon	Leukemia	Lung	CNS	Multiple	Lymph	Prostate	Avg
10	$\{1, 0, 0\}$	59.5	58.8	75.8	52.6	59.3	65.5	51.6	60.5
	J4.8	+0.2	+0.0	+0.1	+0.2	-0.1	+0.4	+0.2	+0.1
		(0.35)	(0.11)	(0.20)	(0.66)	(-0.79)	(1.32)	(0.60)	
	NB	+3.2	+10.6	+8.0	+7.9	+5.2	+9.9	+4.0	+7.0
		(5.29)	(18.04)	(21.20)	(11.33)	(21.91)	(16.43)	(10.40)	
	SVM	+8.7	+19.7	+11.9	+3.8	+8.6	+13.4	+19.3	+12.2
		(13.62)	(31.61)	(32.81)	(5.70)	(36.96)	(22.49)	(41.12)	
20	$\{1, 0, 0\}$	68.2	82.7	81.1	55.2	63.3	74.4	68.0	70.4
	J4.8	+1.2	+0.2	-0.2	+0.1	+0.1	+0.1	+0.2	+0.2
		(3.09)	(0.52)	(-0.66)	(0.36)	(1.15)	(0.39)	(0.72)	
	NB	-4.8	+6.1	+6.2	+7.3	+5.3	+3.8	-6.4	+2.5
		(-6.53)	(7.10)	(15.89)	(8.79)	(19.04)	(5.45)	(-9.26)	
	SVM	+7.9	+7.4	+11.0	+4.6	+11.0	+16.2	+15.2	+10.5
		(9.47)	(9.27)	(27.03)	(5.89)	(48.71)	(23.05)	(22.77)	

Return to Weka page

Cross-Validation - Weka

	WEKA	Colon	Leukemia	Lung	CNS	Multiple	Lymph	Prostate	Avg
5	$\{1, 0, 0\}$	76.5	84.8	93.4	56.2	78.5	79.2	82.4	78.7
	J4.8	+0.4	-0.1	-0.0	-0.5	+0.2	-0.1	+0.3	+0.0
		(0.62)	(-0.35)	(-0.48)	(-1.23)	(1.45)	(-0.22)	(1.76)	
	NB	-19.5	+13.3	+4.6	+3.4	+4.8	+1.3	-19.5	-1.7
		(-25.39)	(31.90)	(20.36)	(3.29)	(13.95)	(2.22)	(-38.81)	
	SVM	+5.0	+13.0	+5.8	+9.2	+13.0	+17.6	+9.6	+10.5
		(6.20)	(29.57)	(26.88)	(12.12)	(43.94)	(29.16)	(18.55)	
10	$\{1, 0, 0\}$	77.5	80.4	93.7	58.3	78.7	81.1	83.3	79.0
	J4.8	+1.7	+0.1	+0.1	+0.1	+0.1	-0.2	+0.1	+0.3
		(4.61)	(0.45)	(0.72)	(0.30)	(0.84)	(-0.60)	(0.88)	
	NB	-21.5	+18.2	+4.3	+1.4	+4.8	-1.3	-20.5	-2.1
		(-37.80)	(38.46)	(19.16)	(2.43)	(17.58)	(-2.30)	(-46.07)	
	SVM	+5.6	+17.8	+5.5	+7.1	+14.0	+15.8	+8.4	+10.6
		(9.33)	(40.89)	(25.07)	(9.47)	(59.72)	(37.02)	(18.45)	

Return to Weka page

Data Set Sources

Name	Citation
Colon	Alon et al. [1999]
Leukemia	Golub et al. [1999]
Lung	Gordon et al. [2002]
CNS	Pomeroy et al. [2002]
Multiple	Ramaswamy et al. [2001]
Lymph	Shipp et al. [2002]
Prostate	Singh et al. [2002]

Return to data list