Evaluating the SVM Component in Oracle 10g Beta

Dept. of Computer Science and Statistics
University of Rhode Island
Technical Report TR04-299

Lutz Hamel and Angela Uvarov
Department of Computer Science and Statistics
University of Rhode Island

Susie Stephens
Oracle Corporation

4/15/04

Table of Contents

Overview

The objective of this beta test was to test the data mining functionality of the Oracle 10g release. Particular attention was given to the newly included Support Vector Machine (SVM) component (Vapnik, 1998). We used two datasets from the UCI machine learning repository[1] in order to test this data mining functionality:

- Wisconsin breast cancer data (Mangasarian & Wolberg, 1990).
- Email spam database.[2]

Both datasets are binary classification problems particularly chosen to be compared to more traditional data mining algorithms. Here we compare the performance of the Oracle SVM implementation with the performance of the C4.5 decision tree algorithm (Quinlan, 1993). Details about the individual datasets appear in the corresponding sections of this document.

We found that the Oracle SVM implementation compares very favorably to the traditional C4.5 decision tree algorithm.

Test I: Wisconsin Breast Cancer Data

This dataset consists of 645 records, once duplicates have been eliminated. Of these records 512 records are reserved for training and 133 for testing. The attributes are defined as follows:

Sample code number: id number
Clump Thickness: 1,2,3,4,5,6,7,8,9,10.
Uniformity of Cell Size: 1,2,3,4,5,6,7,8,9,10.
Uniformity of Cell Shape: 1,2,3,4,5,6,7,8,9,10.
Marginal Adhesion: 1,2,3,4,5,6,7,8,9,10.
Single Epithelial Cell Size: 1,2,3,4,5,6,7,8,9,10.
Bare Nuclei: 1,2,3,4,5,6,7,8,9,10.
Bland Chromatin: 1,2,3,4,5,6,7,8,9,10.
Normal Nucleoli: 1,2,3,4,5,6,7,8,9,10.
Mitose: 1,2,3,4,5,6,7,8,9,10.
Class: classes: 2-benign, 4-malignant

For details on these attributes please refer to the literature. The class distribution in the dataset:

Benign: ~ 65%
Malignant: ~ 35%.

This is a binary classification problem where all the independent attributes are categorical attributes.

[1] http://www.ics.uci.edu/~mlearn

[2] Creators: Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt, Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304.

Results for SVM using a Linear Kernel

Model settings:
SVMS_CONV_TOLERANCE = .001
SVMS_KERNEL_CACHE_SIZE = 50000000
SVMS_COMPLEXITY_FACTOR = .16666666666666699

Confusion Matrix:

		Predicted	
		2 (B)	4 (M)
Actual	2 (B)	93	2
	4 (M)	2	37

Error Rate: 3%
Accuracy: 97%

Results for SVM using a Gaussian Kernel

Model settings:
SVMS_CONV_TOLERANCE = .001
SVMS_KERNEL_CACHE_SIZE = 50000000
SVMS_STD_DEV = 3.7416573867739413
SVMS_COMPLEXITY_FACTOR = 1.1959376673823801

Confusion Matrix:

		Predicted	
		2 (B)	4 (M)
Actual	2 (B)	94	1
	4 (M)	0	39

Error Rate: 0.7%
Accuracy: 99.3%

Results for C4.5

Model Settings:
Pruning Confidence: 25%
Minimum Support: 5

Confusion Matrix:

		Predicted	
		2 (B)	4 (M)
Actual	2 (B)	86	9
	4 (M)	1	38

Error Rate: 7.5%

Accuracy: 92.5%

Simplified Decision Tree:

Uniformity of Cell Size = 1: 2 (268.0/6.2)
Uniformity of Cell Size = 2: 2 (34.0/9.3)
Uniformity of Cell Size = 3: 4 (37.0/19.6)
Uniformity of Cell Size = 4: 4 (31.0/10.3)
Uniformity of Cell Size = 5: 4 (27.0/1.4)
Uniformity of Cell Size = 6: 4 (19.0/1.3)
Uniformity of Cell Size = 7: 4 (16.0/2.5)
Uniformity of Cell Size = 8: 4 (22.0/2.5)
Uniformity of Cell Size = 9: 4 (4.0/2.2)
Uniformity of Cell Size = 10: 4 (55.0/1.4)

Remarks:
Both SVM models are more accurate than the decision tree model. It is interesting to note that the SVM with a Gaussian kernel can model this dataset almost perfectly. As can be seen by the pruned decision tree most of the predictive power of this dataset is due to one attribute: the uniformity of cell size.

Test II: The Spam Database

Each row in this dataset represents an email message that is either considered spam (Cranor & LaMacchia, 1998) or not. The 57 continuous attributes of the dataset describe word and character frequencies in the email messages. The dataset has 4601 records with a class distribution: spam: ~39%, non-spam: ~61%. This is a binary classification problem where all the independent attributes are continuous. The dataset was divided into 3520 training records and 811 records for testing.

Results for SVM using a Linear Kernel

Model Settings:
SVMS_CONV_TOLERANCE = .001
SVMS_KERNEL_CACHE_SIZE = 50000000
SVMS_COMPLEXITY_FACTOR = .11417207662774299

Confusion Matrix:

		Predicted	
		0	1
Actual	0	516	26
	1	9	260

Error Rate: 4.3%
Accuracy: 95.7%

Results for SVM using a Gaussian Kernel

Model Settings:
SVMS_CONV_TOLERANCE = .001
SVMS_KERNEL_CACHE_SIZE = 50000000
SVMS_STD_DEV = 4.812661641473027
SVMS_COMPLEXITY_FACTOR = .75904342468903196

Confusion Matrix:

		Predicted	
		0	1
Actual	0	522	20
	1	15	254

Error Rate: 4.3%
Accuracy: 95.7%

Results for C4.5

Model Settings:
Pruning Confidence: 25%
Minimum Support: 10

Confusion Matrix:

		Predicted	
		0	1
Actual	0	441	33
	1	28	309

Error Rate: 7.5%
Accuracy: 92.5%

<u>Remarks:</u>
Again we observe that the SVM models are more accurate essentially cutting the error rate in half. It is also interesting to note there is essentially a tradeoff between the linear and the Gaussian model at the same error rate: the linear model produces more false positives and the Gaussian model produces more false negatives.

Conclusions

The SVM component of the Oracle 10g database allowed us to perform several data mining exercises without major problems. The obtained models compared very favorably with the models constructed by the C4.5 decision tree algorithm, in fact they appear to be superior to the C4.5 models. Of course, a more detailed statistical analysis is necessary to study if the difference between the model accuracy is statistically significant and how much it depends on the particular selection of test vs. training set.

References

Cranor, L. F., & LaMacchia, B. A. (1998). Spam! *Communications of the ACM, 41*(8), 74-83.

Mangasarian, O. L., & Wolberg, W. H. (1990). Cancer diagnosis via linear programming. *SIAM News, 23*(5), 1-18.

Quinlan, J. R. (1993). *C4.5 : programs for machine learning*. San Mateo, Calif.: Morgan Kaufmann Publishers.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.