Machine Learning for Medical Decision Support: Evaluating Diagnostic Performance of Machine Learning Classification Algorithms

Patrick Herron
INLS 110 Data Mining
Spring 2004

Abstract

The present study explored dichotomic classification methods for medical diagnosis data through three experiments. A first experiment run in Weka used four different classification schemes on two different sets of medical test data thus permitting comparison of each scheme's performance. A second experiment tested the application of attribute selection, information gain, and boosting to Weka's support vector classification scheme (SMO). Finally, in the third experiment when a cost matrix was applied to breast cancer diagnostic data, false negatives were effectively reduced to under one percent while overall accuracy was slightly improved. The first experiment suggests that SMO may classify better than J48, IBk and Naïve Bayes with respect to medical test data from the UCI repository. The data of the first experiment also suggests that support vector classification-based diagnosis outperforms manual diagnosis of fine needle aspirate results. While the second experiment showed no enhancement of support vector performance, it did show that attribute selection may be useful for reducing the number of tests utilized in medical diagnosis. The third experiment showed that the application of a cost matrix lowered the rate of false negatives to < 1% while improving overall accuracy. Results indicate that data mining performance must be evaluated on a domain-by-domain basis in comparison to currently accepted clinical diagnosis practices. Data mining techniques can be successfully used for diagnostic support in medicine, not only for improving upon manual diagnosis but also both for identifying the most valuable medical diagnostic tests and for optimizing them. Ultimately, application of machine learning can only succeed if the domain of each data set is well understood.

Introduction

Data mining classification techniques may be useful for medical diagnosis decision support in a clinical setting. The utility of such methods may be measured using empirical means on real-world examples.

The author of the present study first attempted to determine what criteria should be used for analyzing the utility of data mining classification methods on two different clinical medical diagnostic data sets. Three experiments were run using the Weka Data Mining software package (version 3.4)—the first two using the Weka Experimenter and the third using the Explorer. The results of the three experiments were evaluated using the newly-established criteria.

The first experiment run tested four classification schemes on two different sets of medical test data found in the UCI Machine Learning Repository (Blake, 1998), the Cleveland 14 Heart Disease (slightly modified for present purposes) and the Wisconsin Breast Cancer. The second experiment, designed to evaluate the performance of boosting and attribute selection, was run on the same two data sets using the best-performing classification from experiment 1 modified by best attribute selection, information gain, and boosting. In the third and final Weka experiment, a cost matrix was applied to the overall best-performing classification scheme and run on the best-performing data set in order to see if false negatives could be reduced to zero while still returning a high percentage of true negatives.

The present aim is not to evaluate the complexities of each algorithm in light of the data but instead to test their efficacy in classifying the data in the context of clinical diagnosis. Algorithms are treated in a black-box fashion.

Rationale

In order to evaluate the performance of classification algorithms in the domain of medical test data we must choose evaluation criteria. Percent correct is frequently assumed to be the best indicator of performance on medical data. Yet, without a benchmark of manually-derived percent-correct (overall accuracy) values given a particular data set, our classifier's percent-correct result does not tell the whole story.

When medical clinical test data are generated for a diagnosis, they are usually done so in the context of making a medical decision, for the purpose of augmenting choices about further testing and/or treatment. For example, a minimally-invasive test on a skin tumor may provide information as to whether the tumor should be surgically excised or treated with chemotherapeutic agents, treatments which carry significant risk, risk that is dangerous if the patient does not actually require such treatment. A clinician may prefer to keep the number of false positives low. In our previous example, however, it is the false *negative* outcome (test analysis suggests "benign" when the cancer is actually malignant) that could be disastrous in the formulation of a diagnosis. It is undesirable to

tell a healthy patient that he or she is sick, but it is likely worse, particularly with respect to life-threatening ailments, to tell a sick patient that he or she is healthy and needs no further physical examination or treatment. We may certainly want our classification to perform well at discovering positive instances, but the over-examination and treatment of a healthy patient is less fortunate than the clean bill of health to an ill patient. Overall accuracy, then, is not the spirit of classification for medical diagnosis decisions; the spirit is not to use classification to issue diagnoses but instead to use it in order to assist the physician. We ideally want the rate of false negatives to be as close to zero as is reasonably possible. In other words we should be willing to sacrifice precision of positive classifications in exchange for improving the precision of negative classifications. We wish to issue a clean bill of health every time someone is healthy in order to circumvent unnecessary, dangerous, and costly treatment; and we want to never give someone who is sick a clean bill of health.

The latter approach to evaluation is admittedly context-free; it does not take into account how well physicians perform at manually classifying an illness given the same data set. If we have a baseline of how the data are evaluated prior to their classification, we then can see whether classification offers any improvement over currently accepted practices. Given such baseline data we may lower our standards from the objective "great & ideal" zero-false-negative rate to "better than what we have now" comparisons.

Data and Methods

Experiment 1

The first experiment tested four classification schemes on two different sets of medical test data found in the UCI Machine Learning Repository (Blake, 1998), the Cleveland 14 Heart Disease (slightly modified for present purposes) and the Wisconsin Breast Cancer data set. The experiment was run using Weka 3.4's Experimenter interface with 10-fold cross-validation and three repetitions.

The Wisconsin Breast Cancer (WBC) data set contains data collected from the examination of cell samples obtained by fine needle aspiration biopsy from breast masses found in 699 patients. Aspirate samples were evaluated using a number of criteria and measures represented by the attributes and values of the data set:

Clump Thickness	1 - 10
Uniformity of Cell Size	1 - 10
Uniformity of Cell Shape	1 - 10
Marginal Adhesion	1 - 10
Single Epithelial Cell Size	1 - 10
Bare Nuclei	1 - 10
Bland Chromatin	1 - 10
Normal Nucleoli	1 - 10
Mitoses	1 - 10

The above attributes are regularly and reliably used by pathologists to differentiate between benign and malignant breast masses.

The WBC data set contains 699 instances described using 9 numeric attributes plus the class attribute (benign, malignant). 458 of those instances are benign and 241 are malignant. The data set contains no missing values.

The Cleveland-14-Heart Disease (CHD) data set[*] contains data collected from the examination of patients showing possible signs of heart disease as represented by a percentage of vascular occlusion.

The CHD set contains 303 instances. The classifier for the CHD was originally split into four classes: <50% occlusion, and three groups of >50% occlusion. All three groups of >50% occlusion have been lumped into one group for present purposes; the three groupings provided no additional information about the classification. 164 instances showed <50% occlusion while the other 139 showed occlusion of >50%. A number of

[*] Principal investigators: Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.; University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.; University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.; V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D

values are missing from the data set and are distinguished with the value -9.0. The data set is described by 13 attributes (7 numeric, 6 factors) with a class attribute indicating >50% occlusion or <50% occlusion. The attributes in the data set are as follows:

```
age: age in years (numeric)
sex: sex (factor; 1 = male; 0 = female)
cp: chest pain type (factor)
  -- Value 1: typical angina
  -- Value 2: atypical angina
  -- Value 3: non-anginal pain
  -- Value 4: asymptomatic
trestbps: resting blood pressure (numeric; in mm Hg on
admission to the hospital)
chol: serum cholesterol in mg/dl (numeric)
fbs: (fasting blood sugar > 120 mg/dl)  (1 = true; 0 =
false) (binarized numeric)
restecg: resting electrocardiographic results (factor)
  -- Value 0: normal
  -- Value 1: having ST-T wave abnormality (T wave
inversions and/or ST elevation or depression of > 0.05 mV)
  -- Value 2: showing probable or definite left
ventricular hypertrophy by Estes' criteria
thalach: maximum heart rate achieved (numeric)
exang: exercise induced angina (factor; 1 = yes; 0 = no)
oldpeak: ST depression induced by exercise relative to
rest (numeric)
slope: the slope of the peak exercise ST segment (factor)
  -- Value 1: upsloping
  -- Value 2: flat
  -- Value 3: downsloping
ca: number of major vessels (0-3) colored by flourosopy
(numeric)
thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
(factor)
classifier - num: diagnosis of heart disease (angiographic
disease status)
  -- Value 0: < 50% diameter narrowing
  -- Value 1: > 50% diameter narrowing
```

The two data sets were evaluated using four different classification algorithms: IBk, J48, Naïve Bayes, and SMO. IBk is the implementation of k-nearest neighbors used in Weka; J48 is Weka's equivalent of C4.5. SMO uses a sequential minimal optimization algorithm to train a support vector classifier; it is Weka's flavor of the support vector machine algorithm.

Experiment 2

The second experiment was designed to evaluate the performance of boosting and attribute selection. The experiment was conducted on the WBC and CHD data sets using the best-performing classification from experiment 1 (SMO) modified by best attribute selection, information gain, and boosting classification enhancements.

WBC and CHD were evaluated using three successive methods based in SMO to see if modification could improve upon SMO. First, boosting (using Weka's AdaBoostingM1 implementation) was added to SMO. Second, attribute selection was used to reduce dimensionality to seven attributes; a ranker evaluated each attribution by information gain and picked the best seven. The third and final phase of experiment 2 combined the methods of the previous two tests by boosting the attribute-selected SMO.

Experiment 3
In the third and final Weka experiment, a cost matrix was applied to the overall best-performing classification scheme (SMO) and run on the best-performing data set (WBC) in order to see if false negatives could be reduced to zero while still returning a high percentage of true negatives.

Results

Experiment 1
The overall results of Experiment 1 show the SMO algorithm outperformed the IBk algorithm by a slight margin (9 ranking wins versus 8). Both algorithms dramatically outperformed the other two algorithms tested—the Naïve Bayes algorithm and J48.

The Wisconsin Breast Cancer (WBC) data set shows an exceptionally high rate of success. The SMO algorithm was correct at classifying the WBC set (determining tumor malignancy) 97% of the time (+/- 2%). With the Cleveland Heart Disease (CHD) data, however, we see a much lower rate of success, with percent correct at 84% (+/- 6%). True negative results ("benign" in the WBC and "<50% vascular occlusion" in the CHD) showed similar performance across all algorithms.

Negative classifications were more difficult for CHD (SMO: 77% +/- 1%) than positive ones (89% +/- 1%). The Wisconsin Breast Cancer data performed about the same with respect to positive and negative diagnoses. The frequency for false negatives in classifying both data sets using each algorithm was lower than the frequency of false positives. IBk and SMO performed best with respect to false negatives across both data sets.

```
-----------------
Experiment 1 Data
-----------------
Weka Experiment Environment
All tests run using two tailed confidence of 0.05


Percent_correct
Dataset                   (1)           SMO  | (2)       J48    (3)NaiveBayes    (4)       IBk
-------------------------------------------------------------------------------------------------
wisconsin-breast-cancer   (30)   96.71( 1.96) |   95.04( 3.05)     95.99( 2.08)     96.57( 2.24)
cleveland-14-heart-diseas (30)   83.6 ( 6.08) |   77.56( 5.84) *   83.37( 6.88)     82.26( 6.18)
-------------------------------------------------------------------------------------------------
Percent_correct
>-<   >   < Resultset
  1   1   0 IBk -K 10
  1   1   0 SMO
  0   0   0 NaiveBayes
 -2   0   2 J48


Kappa_statistic
Dataset                   (1)           SMO  | (2)       J48    (3)NaiveBayes    (4)       IBk
-------------------------------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.93( 0.04) |    0.89( 0.07)      0.91( 0.05)      0.92( 0.05)
cleveland-14-heart-diseas (30)    0.67( 0.13) |    0.55( 0.12) *    0.66( 0.14)      0.64( 0.12)
-------------------------------------------------------------------------------------------------
Kappa_statistic
>-<   >   < Resultset
  1   1   0 IBk -K 10
  1   1   0 SMO
  0   0   0 NaiveBayes
 -2   0   2 J48



Root_mean_squared_error
Dataset                   (1)           SMO  | (2)       J48    (3)NaiveBayes    (4)       IBk
-------------------------------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.16( 0.08) |    0.2 ( 0.07)      0.19( 0.07)      0.15( 0.05)
cleveland-14-heart-diseas (30)    0.4 ( 0.07) |    0.44( 0.05)      0.35( 0.08) *    0.35( 0.05) *
-------------------------------------------------------------------------------------------------
Root_mean_squared_error
>-<   >   < Resultset
  3   3   0 J48
  2   2   0 SMO
 -1   1   2 NaiveBayes
 -4   0   4 IBk -K 10


True_positive_rate
Dataset                   (1)           SMO  | (2)       J48    (3)NaiveBayes    (4)       IBk
-------------------------------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.97( 0.03) |    0.96( 0.04)      0.95( 0.03) *    0.98( 0.02)
cleveland-14-heart-diseas (30)    0.89( 0.06) |    0.81( 0.09) *    0.87( 0.08)      0.86( 0.09)
-------------------------------------------------------------------------------------------------
True_positive_rate
>-<   >   < Resultset
  2   2   0 SMO
  1   1   0 IBk -K 10
 -1   0   1 J48
 -2   0   2 NaiveBayes


False_positive_rate
Dataset                   (1)           SMO  | (2)       J48    (3)NaiveBayes    (4)       IBk
-------------------------------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.04( 0.05) |    0.07( 0.05)      0.03( 0.04)      0.05( 0.06)
cleveland-14-heart-diseas (30)    0.23( 0.13) |    0.26( 0.11)      0.21( 0.12)      0.22( 0.1 )
-------------------------------------------------------------------------------------------------
False_positive_rate
>-<   >   < Resultset
  0   0   0 IBk -K 10
  0   0   0 NaiveBayes
  0   0   0 J48
  0   0   0 SMO
```

```
True_negative_rate
Dataset                    (1)           SMO  | (2)       J48    (3)NaiveBayes    (4)       IBk
                          ------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.96( 0.05) |    0.93( 0.05)       0.97( 0.04)      0.95( 0.06)
cleveland-14-heart-diseas (30)    0.77( 0.13) |    0.74( 0.11)       0.79( 0.12)      0.78( 0.1 )
                          ------------------------------------------------------------------------

True_negative_rate
>-<   >   < Resultset
  0   0   0 IBk -K 10
  0   0   0 NaiveBayes
  0   0   0 J48
  0   0   0 SMO


False_negative_rate
Dataset                    (1)           SMO  | (2)       J48    (3)NaiveBayes    (4)       IBk
                          ------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.03( 0.03) |    0.04( 0.04)       0.05( 0.03) v    0.02( 0.02)
cleveland-14-heart-diseas (30)    0.11( 0.06) |    0.19( 0.09) v     0.13( 0.08)      0.14( 0.09)
                          ------------------------------------------------------------------------

False_negative_rate
>-<   >   < Resultset
  1   1   0 SMO
  1   1   0 IBk -K 10
 -1   0   1 J48
 -1   0   1 NaiveBayes


IR_precision
Dataset                    (1)           SMO  | (2)       J48    (3)NaiveBayes    (4)       IBk
                          ------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.98( 0.02) |    0.97( 0.02)       0.99( 0.02)      0.97( 0.03)
cleveland-14-heart-diseas (30)    0.83( 0.08) |    0.79( 0.07)       0.84( 0.08)      0.83( 0.07)
                          ------------------------------------------------------------------------

IR_precision
>-<   >   < Resultset
  1   1   0 NaiveBayes
  0   0   0 IBk -K 10
  0   0   0 SMO
 -1   0   1 J48

IR_recall
Dataset                    (1)           SMO  | (2)       J48    (3)NaiveBayes    (4)       IBk
                          ------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.97( 0.03) |    0.96( 0.04)       0.95( 0.03) *    0.98( 0.02)
cleveland-14-heart-diseas (30)    0.89( 0.06) |    0.81( 0.09) *     0.87( 0.08)      0.86( 0.09)
                          ------------------------------------------------------------------------

IR_recall
>-<   >   < Resultset
  2   2   0 SMO
  1   1   0 IBk -K 10
 -1   0   1 J48
 -2   0   2 NaiveBayes

Overall rankings: (Scoring: *4 for false negative, *3 for true negatives, *2 for percent correct, *1 for all others)
-----------------
>-<   >   < Resultset
  9   9   0 SMO
  8   8   0 IBk -K 10
 -6   0   6 J48
-11   1  12 NaiveBayes
```

Experiment 2

Experiment 2, as evidenced by the rankings in the results data, effectively demonstrates that boosting and/or attribute selection via information gain does little or nothing to alter performance. The only marginal performance gains found were with respect to error, so marginal that they are of little noteworthiness. A reduction in the number of attributes does not detract from the performance of SMO classification for either the WBC data set or the CHD data set.

```
-----------------
Experiment 2 Data
-----------------
Weka Experiment Environment
All tests run using two tailed confidence of 0.05


Percent_correct
Dataset                    (1)           SMO  | (2) AdaBoost     (3) AttSel-7d    (4) Ada+AttSel-7d
-------------------------------------------------------------------------------------------------------
wisconsin-breast-cancer   (30)   96.71( 1.96) |   96.61( 2.08)     96.23( 2.21)     96.23( 2.21)
cleveland-14-heart-diseas (30)   83.6 ( 6.08) |   83.38( 6.74)     83.92( 5.99)     83.92( 5.99)
-------------------------------------------------------------------------------------------------------

Percent_correct

>-<   >   < Resultset
  0   0   0 metaAdaBoostM1 of (AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO)
  0   0   0 AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO
  0   0   0 AdaBoostM1 of SMO
  0   0   0 SMO



Kappa_statistic
Dataset                    (1)           SMO  | (2) AdaBoost     (3) AttSel-7d    (4) Ada+AttSel-7d
-------------------------------------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.93( 0.04) |    0.93( 0.05)      0.92( 0.05)      0.92( 0.05)
cleveland-14-heart-diseas (30)    0.67( 0.13) |    0.66( 0.14)      0.67( 0.12)      0.67( 0.12)
-------------------------------------------------------------------------------------------------------

Kappa_statistic
>-<   >   < Resultset
  0   0   0 metaAdaBoostM1 of (AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO)
  0   0   0 AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO
  0   0   0 AdaBoostM1 of SMO
  0   0   0 SMO



Root_mean_squared_error
Dataset                    (1)           SMO  | (2) AdaBoost     (3) AttSel-7d    (4) Ada+AttSel-7d
-------------------------------------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.16( 0.08) |    0.17( 0.07)      0.18( 0.08)      0.18( 0.07)
cleveland-14-heart-diseas (30)    0.4 ( 0.07) |    0.36( 0.07) *    0.39( 0.07)      0.36( 0.07) *
-------------------------------------------------------------------------------------------------------
Root_mean_squared_error
>-<   >   < Resultset
  0   1   1 SMO
  1   0   1 AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO
  0   1   1 AdaBoostM1 of SMO
  0   1   1 metaAdaBoostM1 of (AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO)



True_positive_rate
Dataset                    (1)           SMO  | (2) AdaBoost     (3) AttSel-7d    (4) Ada+AttSel-7d
-------------------------------------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.97( 0.03) |    0.97( 0.03)      0.97( 0.03)      0.97( 0.03)
cleveland-14-heart-diseas (30)    0.89( 0.06) |    0.88( 0.07)      0.89( 0.07)      0.89( 0.07)
-------------------------------------------------------------------------------------------------------
True_positive_rate
>-<   >   < Resultset
  0   0   0 metaAdaBoostM1 of (AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO)
  0   0   0 AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO
  0   0   0 AdaBoostM1 of SMO
  0   0   0 SMO
```

```
False_positive_rate
Dataset                    (1)           SMO  | (2) AdaBoost     (3) AttSel-7d    (4) Ada+AttSel-7d
                        ----------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.04( 0.05) |    0.04( 0.05)      0.05( 0.05)      0.05( 0.05)
cleveland-14-heart-diseas (30)    0.23( 0.13) |    0.23( 0.12)      0.22( 0.13)      0.22( 0.13)
                        ----------------------------------------------------------------------------
>-<   >   < Resultset
  0   0   0 metaAdaBoostM1 of (AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO)
  0   0   0 AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO
  0   0   0 AdaBoostM1 of SMO
  0   0   0 SMO



True_negative_rate
Dataset                    (1)           SMO  | (2) AdaBoost     (3) AttSel-7d    (4) Ada+AttSel-7d
                        ----------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.96( 0.05) |    0.96( 0.05)      0.95( 0.05)      0.95( 0.05)
cleveland-14-heart-diseas (30)    0.77( 0.13) |    0.77( 0.12)      0.78( 0.13)      0.78( 0.13)
                        ----------------------------------------------------------------------------
True_negative_rate
>-<   >   < Resultset
  0   0   0 metaAdaBoostM1 of (AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO)
  0   0   0 AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO
  0   0   0 AdaBoostM1 of SMO
  0   0   0 SMO



False_negative_rate
Dataset                    (1)           SMO  | (2) AdaBoost     (3) AttSel-7d    (4) Ada+AttSel-7d
                        ----------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.03( 0.03) |    0.03( 0.03)      0.03( 0.03)      0.03( 0.03)
cleveland-14-heart-diseas (30)    0.11( 0.06) |    0.12( 0.07)      0.11( 0.07)      0.11( 0.07)
                        ----------------------------------------------------------------------------
False_negative_rate
>-<   >   < Resultset
  0   0   0 metaAdaBoostM1 of (AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO)
  0   0   0 AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO
  0   0   0 AdaBoostM1 of SMO
  0   0   0 SMO


IR_precision
Dataset                    (1)           SMO  | (2) AdaBoost     (3) AttSel-7d    (4) Ada+AttSel-7d
                        ----------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.98( 0.02) |    0.98( 0.02)      0.97( 0.03)      0.97( 0.03)
cleveland-14-heart-diseas (30)    0.83( 0.08) |    0.83( 0.08)      0.84( 0.08)      0.84( 0.08)
                        ----------------------------------------------------------------------------
IR_precision
>-<   >   < Resultset
  0   0   0 metaAdaBoostM1 of (AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO)
  0   0   0 AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO
  0   0   0 AdaBoostM1 of SMO
  0   0   0 SMO


IR_recall
Dataset                    (1)           SMO  | (2) AdaBoost     (3) AttSel-7d    (4) Ada+AttSel-7d
                        ----------------------------------------------------------------------------
wisconsin-breast-cancer   (30)    0.97( 0.03) |    0.97( 0.03)      0.97( 0.03)      0.97( 0.03)
cleveland-14-heart-diseas (30)    0.89( 0.06) |    0.88( 0.07)      0.89( 0.07)      0.89( 0.07)
                        ----------------------------------------------------------------------------
IR_recall
>-<   >   < Resultset
  0   0   0 metaAdaBoostM1 of (AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO)
  0   0   0 AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO
  0   0   0 AdaBoostM1 of SMO
  0   0   0 SMO


(1) SMO
(2) AdaBoostM1 of SMO
(3) AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO
(4) AdaBoostM1 of (AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO)


Overall Rankings
--------------------
>-<   >   < Resultset
  2   2   0 SMO
  1   1   0 AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO
 -1   0   1 AdaBoostM1 of SMO
 -2   0   2 metaAdaBoostM1 of (AttributeSelectedClassifier using InfoGainAttributeEval -N 7 of SMO)
```

Experiment 3

The use of a cost matrix in an attempt to reduce false negatives succeeded in reducing the number of false negatives to a significant degree. The cost matrix used in this final experiment ultimately reduced the false negative rate below 1%. Despite the application of the cost matrix to the SMO classification the overall accuracy remained 97%

```
-----------------
Experiment 3 Data
-----------------
Weka Explorer


=== Run information ===

Scheme:       weka.classifiers.meta.CostSensitiveClassifier -S 1 -W weka.classifiers.functions.SMO -- -C 1.0 -E 1.0 -G
0.01 -A 1000003 -T 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1
Relation:     wisconsin-breast-cancer
Instances:    699
Attributes:   10
              Clump_Thickness
              Cell_Size_Uniformity
              Cell_Shape_Uniformity
              Marginal_Adhesion
              Single_Epi_Cell_Size
              Bare_Nuclei
              Bland_Chromatin
              Normal_Nucleoli
              Mitoses
              Class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

CostSensitiveClassifier using reweighted training instances

weka.classifiers.functions.SMO -C 1.0 -E 1.0 -G 0.01 -A 1000003 -T 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1

Classifier Model
SMO

Classifier for classes: benign, malignant

BinarySMO

Machine linear: showing attribute weights, not support vectors.

         2.003  * (normalized) Clump_Thickness
 +       0.6644 * (normalized) Cell_Size_Uniformity
 +       1.0539 * (normalized) Cell_Shape_Uniformity
 +       1.305  * (normalized) Marginal_Adhesion
 +       0.6501 * (normalized) Single_Epi_Cell_Size
 +       1.7434 * (normalized) Bare_Nuclei
 +       0.9517 * (normalized) Bland_Chromatin
 +       1.1355 * (normalized) Normal_Nucleoli
 +       0.7273 * (normalized) Mitoses
 -       2.3152

Number of kernel evaluations: 6359

Cost Matrix
 1    1.2
 2.4  1


Time taken to build model: 0.39 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         678               96.9957 %
Incorrectly Classified Instances        21                3.0043 %
Kappa statistic                          0.9346
Mean absolute error                      0.03
Root mean squared error                  0.1733
Relative absolute error                  6.6471 %
Root relative squared error             36.4672 %
Total Number of Instances              699

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.959     0.008      0.995     0.959     0.977    benign
  0.992     0.041      0.926     0.992     0.958    malignant
```

```
=== Confusion Matrix ===

   a   b   <-- classified as
 439  19 |   a = benign
   2 239 |   b = malignant
---------------------------------------
```

Discussion

Experiment 1

The first experiment shows that support vector machine (via Weka's SMO algorithm) classification generally performs better than C4.5 (via Weka's J48 algorithm), k-nearest neighbor (via Weka's IBk algorithm), and Naïve Bayes for both data sets. Secondly, and perhaps most importantly, classification techniques are shown to be useful for providing decision support for one of the two data sets, the Wisconsin Breast Cancer data.

The difference in outcomes between the two data sets seems to make a great deal of intuitive sense[∝]. The tests use different types of measures, for one. All of the measures in the WBC set result from direct observation of that which is either benign or malignant while the measures used in the CHD set seem to only share some association with the pathology.

Manually detecting a difference between a benign tumor and a malignant one using fine needle aspiration has a traditionally high rate of success. A 37-series study showed manual accuracy at 94.3% over thousands of fine needle aspirate tests on breast tumor (Frable, 1983). The data is numeric in every case, and a very small set of features tend to rather accurately portend the right diagnosis.

While I could find no rate-of-success statistics with respect to manually diagnosing heart disease using the attributes in the CHD data set, I do know that prediction of occlusion is at best a loosely informed guess based on the attributes contained in the CHD data set. Physicians are likely to recommend a more invasive procedure (*e.g.,* cardiac catheterization) before they make a complete diagnosis on the degree of coronary occlusion. Further, whether someone has less than or more than 50% occlusion is not the same as saying whether or not someone has heart disease; it may be reasonable to consider both the person with 40% occlusion and the person with 60% occlusion as sufferers of heart disease. In contrast, someone with only a benign tumor does not have cancer, while someone with a malignancy does. In a sense, then, it seems that the attributes in the CHD are not sufficient for providing a conclusive diagnosis, and so we should expect that the results of an automated classification of the CHD should be considered suggestive instead of conclusive. Ultimately I do not have sufficient data to

[∝] I have three years of cancer research experience studying the relationship between toxicity, genetics, and cancer; and have, in addition, regularly reviewed and edited heart disease-related research papers over the course of five years for a doctoral student studying under the influential pathologist Oliver Smithies (*e.g.,* see the acknowledgements in Knouff, et al, Apo E structure determines VLDL clearance and atherosclerosis risk in mice, *Journal of Clinical Investigations*, June 1999, Volume 103, Number 11, 1579-1586). I also spent two years helping to set standards for test validations. I do not claim to be a domain expert in oncology, pathology, or genetics or heart disease, but I am at least passable in evaluating test methods in both domains.

decide whether our classification techniques improve the state of affairs in heart disease diagnosis; previous efforts have not proven improvement over clinicians with respect to negative diagnosis (Kukar, 1999).

Experiment 1 underscores an important caveat about machine learning: it is no silver bullet. Each data set must, and transitively, each domain represented by its data set, must be well-understood before the merits of any machine learning approach to the data can be properly assessed.

Machine-based classification can work in a clinical setting, but it does not always outperform manual or traditional means. The mere consideration of out-performance may not always be particularly informative: in some domains, such as medicine, a machine learning classification scheme may be best considered as an assistant to diagnosis, as decision *support*, rather than an independent diagnostic machine. It is likely that the marriage of clinicians and machine learning may outperform either one on its own. Any particular use of machine learning-derived diagnosis methods should undergo rigorous clinical testing before widespread clinical use and should only be deployed as a tool to assist a clinician.

Experiment 2

Modifying SMO by selecting best attributes does not enhance or impoverish results in any marked way. However what is noteworthy is that reducing the attribute set down to 7 (from 14 in the CHD and 10 in the WBC) does *not* detract from performance in any way. Such an observation is in line with the early intentions of those who put together the WBC data (Wohlberg, 1995); machine learning was previously used in order to determine which diagnostic measures provided the most crucial information.

Despite the precedences set from earlier dimensionality reduction experiments, I find the results of experiment 2 to be somewhat surprising with respect to the WBC data. For one, diseases such as cancer and heart disease are highly polygenic; they most likely involve a large number of diverse genes and an even more diverse set of conditions operating on them. In other words, organisms are incredibly complex and unformulaic, and, unlike text mining where more attributes may easily introduce more noise, the more information made available to us related to elements of the disease's polygeny, the greater our chances of making good diagnoses. All of the attributes of the WBC data set deal directly with the site of pathology, such as cellular features of the mass. All of the measures in the WBC should provide some information, and their removal should degrade performance somewhat. Not so, as it turns out. I am less surprised with respect to the CHD data. The 14 attributes of the CHD sata set were carefully pared down from an original 77 for performance reasons, and so what attributes remain should provide excellent information. However, the additional dimensionality reduction removes the least direct measures, the more epidemiological measures, those measures less directly related to the pathology (*e.g.,* age, sex, serum cholesterol levels), and so we should expect that they can be thrown away without deleterious effect.

Experiment 3

Adding a cost matrix to SMO in evaluating the WBC data set not only provided the intended result of reducing false negatives—it also improved overall accuracy. We wanted to see if we could, via a cost matrix, reduce the number of false negatives while maintaining the quality of our results, and it worked. The trade-off was in false positives, but we would rather have a false positive than a false negative as discussed earlier. A false negative effectively represents a diagnosis of benign in the case of a malignancy while the false positive represents the very opposite. Diagnosing a patient is done in the context of further testing and or treatment; recommending against further evaluation and/or treatment of a sick person is inherently worse than increasing evaluation and/or treatment of a healthy person.

Conclusions

Of the nine methods tested, plain SMO enhanced with a cost matrix that adds a penalty to false negatives provides the best results overall. Attribute selection may provide us with the means to reduce the number of clinical measures made while still maintaining accuracy and reducing false negative rates.

Data mining techniques are useful in certain cases in order to advise against more aggressive treatments that may be superfluous and/or detrimental. With respect to breast cancer, classification analysis of minimally invasive fine needle aspiration test results reliably provides evidence against further aggressive testing and/or treatment. Further, classification techniques can be used to determine which individual medical tests and are more useful than others in providing diagnostic information. The author recommends that classification data mining techniques for fine needle aspiration tests are worthy of clinical testing. Further, other areas of pathology seem to be well-disposed to data mining; machine learning classification should at least be aggressively tested in other areas of pathology. Analysis of such results may benefit greatly in the light of longitudinal medical treatment outcomes data, data that was not available for the present study.

References

Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Frable, W.J. (1983). Thin-needle aspiration biopsy. *Major Problems in Pathology*. WB Saunders Co., Philadelphia.

Kukar, M.; Kononenko, I.; Groselj, C.; Kralj, K.; Fettich, J. Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artificial Intelligence in Medicine*, 1999; 16:25-50.

Wohlberg, W.; Street, N.; Helsey, D.; Mangasarian, O. Computerized Breast Cancer Diagnosis and Prognosis From Fine Needle Aspirates. *Archives of Surgery*, 1995; 130:511-516.