Quantitative Methods for Evaluating Learning Classifier System Performance in Forced Two-Choice Decision Tasks

John H. Holmes
Center for Clinical Epidemiology and Biostatistics
University of Pennsylvania School of Medicine
Philadelphia, PA 19104
jholmes@cceb.med.upenn.edu

Abstract

Applying a learning classifier system to two-class decision problems requires a special approach to performance evaluation. This paper presents a suite of quantitative tools that addresses the evaluation requirements of two-class problems. These metrics, borrowed from the domain of medical decision making, are proposed as adjuncts to commonly used evaluation methods such as crude accuracy ("percent correct"). They include sensitivity, specificity, area under the receiver operating characteristic curve, and predictive value. These metrics are shown to be superior to crude accuracy in evaluating learning classifier system performance, especially when applied to data with unequal numbers of positive and negative cases. In addition, these metrics provide information to the researcher that is not available from crude accuracy. When used appropriately, these metrics provide accurate depictions of learning classifier system performance during training and testing in supervised learning environments.

1.0 INTRODUCTION

This work examines several tools for evaluating learning classifier system (LCS) performance on two-class decision problems in supervised learning environments. These tools are commonly used in the domain of medical decision making, where evaluation activity focuses on the ability of clinical tests to discriminate between two classes, usually disease or non-disease states. While these tools are commonplace in medical decision making, they have rarely been used in evaluating the performance of classification systems based on evolutionary computational approaches. This paper presents and discusses these tools, and proposes their use as a comprehensive performance evaluation toolkit by those working in evolutionary computation, particularly in classification domains, and especially by those working with LCS. Although this work focuses on two-choice decision tasks, it can be extended to classification tasks using data with multi-categorical or continuous classes.

Traditionally, the tools of LCS performance evaluation have been restricted to metrics such as "percent correct," "error rate," or the number of iterations required to reach a goal. The first two are equivalent to *crude accuracy*, in that they reflect the proportion of decisions that are correct (or incorrect). These metrics are certainly useful and valid in many environments. However, when there are unequal numbers of cases in the representative classes in training and/or testing data, crude accuracy is a poor estimator of system performance. The third metric, being an indicator of LCS *efficiency*, is useful in terms of local LCS performance, particularly in unsupervised learning environments. However, it provides no information about the *accuracy* of the decisions made by the LCS, either in training or testing.

None of these traditional metrics provide the researcher with *a priori* knowledge of whether or not a given LCS should be used in a particular environment. Furthermore, they do not provide a means whereby one may evaluate, *a posteriori*, whether or not a decision that has been made by an LCS was, in fact accurate. A number of alternative tools exist for evaluating LCS performance. All of these tools address these two shortcomings of traditional LCS metrics. Some of these tools may or may not be useful in all settings, while others are useful only in that they provide the foundation for other, more robust metrics. Each of them will be discussed in turn, as components in the

LCS Metric Toolkit. As the components in the Toolkit are borrowed from the domain of medical decision science (Hennekens and Buring, 1987; Fletcher et al, 1988), they will be introduced in the context of a simple clinical decision making example.

3.0 THE TOOLKIT

3.1 INTRODUCTION

In a two-choice decision problem, a *2x2 contingency* table may be used to describe the characteristics of a test. The contingency table is a confusion matrix in which the rows contain the counts of those with a positive or negative test result; these counts represent the *decisions* of the test. The columns contain the counts of those with a positive or negative result using a *gold standard* test that is accepted as a true indicator of disease. An example of a gold standard result would be one obtained through autopsy.

Test	Gold standard	
	Positive	Negative
Positive	*True positive*	*False positive*
Negative	*False negative*	*True negative*

Figure 1. A generic 2x2 contingency table.

Figure 1 demonstrates a generic 2x2 table, with the individual cells labeled as to the type of result, or decision that can be made, in a two-choice problem. A *true positive* (or *true negative*) result is one in which the test gives the same result as the gold standard. *False positive* (or *false negative*) results represent a discordance between the test and the gold standard; these are also called *Type I* and *Type II errors*, respectively. In practice, each cell contains the counts of each type of result. This is shown below in Figure 2, which is an example of a 2x2 table of diagnosis of infection, determined by a new test, and specimen culture, a widely-accepted gold standard for detecting infection.

New test for infection	Culture	
	Positive	Negative
Positive	38	14
Negative	7	41

Figure 2. A 2x2 contingency table for a hypothetical study of a new diagnostic test for infection.

In order to determine the classification performance of the new test, several proportional metrics can be calculated from the 2x2 table. Each of these provides a different perspective on the ability of the new test to discriminate accurately between those with and without infection. These metrics are crude accuracy, sensitivity, specificity, positive and negative predictive values, and area under the receiver operating characteristic curve.

3.2 CRUDE ACCURACY

Crude accuracy (CA) represents the proportion of correct decisions over all decisions made. It is calculated from the 2x2 table as:

$$\text{Crude accuracy} = \frac{\text{True positive decisions} + \text{True negative decisions}}{\text{All decisions}}$$

Thus, from the example in Figure 2:

$$\text{Crude accuracy} = \frac{38 + 41}{100} = 0.79$$

For many classification problems, CA is an appropriate and useful performance measure. However, in the example shown in Figure 2, CA will overestimate the performance of the test. This is due to the imbalance between the number of gold standard (culture)-positive (n=45) and gold standard-negative cases (n=55). CA is sensitive to the base rate of the classes; in order for CA to be useful, it can be applied only to problems where the number of gold standard-positive and negative cases are equal. Figure 3 demonstrates the relationship between CA and base rate. At a base rate of 50% (where the number of gold standard-positive cases equals the number of gold standard-negative cases), CA is 0.70. However, CA increases linearly with decreasing positive base rate (smaller numbers of gold standard-positive cases), and decreases linearly with increasing positive base rate. Thus, CA is an inversely proportional function of base rate, and can substantially over- or underestimate classification

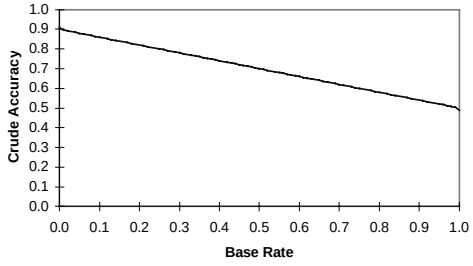

Figure 3. Relationship between crude accuracy and base rate.

performance of a test in environments where the base rate deviates from 50%.

3.3 SENSITIVITY

Sensitivity indicates a test's ability to classify correctly gold standard-positive cases; as a result, sensitivity is often referred to as the *true positive rate*:

$$\text{Sensitivity} = \frac{\text{True positive decisions}}{\text{All gold standard positives}}$$

From the example in Figure 2,

$$\text{Sensitivity} = \frac{38}{45} = 0.84$$

This value indicates that the new test for infection will detect 84% of those with infection. The remaining 16% will test negative but actually have the infection (false negatives). The acceptability of this value is dependent on context; in many cases, a test with a sensitivity of 0.84 would be considered to be very sensitive. However, in a disease such as cancer, where a missed case is potentially a fatal one due to lack of treatment, one would want the sensitivity to be closer to 1.0. Thus, in testing patients for cancer, one would want a test that is highly sensitive.

Like CA, sensitivity is influenced by base rate, with a tendency toward increased false negatives in data where the base rate of positive examples is low. In these data, sensitivity will correspondingly be lower than at higher positive base rates, simply because the proportion is based on a smaller denominator. For example, if there are 10 gold-standard-positives, and five true positive decisions, the sensitivity is 0.50. However, if there are 100 gold-standard positives and the same number (five) of true positives, the sensitivity is 0.95.

3.4 SPECIFICITY

Specificity, or *true negative rate*, measures the ability of a test to classify correctly those *without* disease:

$$\text{Specificity} = \frac{\text{True negative decisions}}{\text{All gold standard negatives}}$$

The specificity calculated from the example in Figure 2 is :

$$\text{Specificity} = \frac{41}{55} = 0.75$$

As is the case with sensitivity, an acceptable value for specificity is dependent on the problem domain. In the example, 75% of people testing negative with the new test will actually be negative. The remaining 25% (false positives) will be those who tested positive but are actually negative. In clinical situations where treating someone for a disease is potentially dangerous (such as the case with certain antibiotic therapies, which can cause severe reactions), a test which is highly specific is very desirable.

Specificity is also influenced by base rate, but in the opposite direction. The proportion of negative examples in a given data set will determine the effect of false positive decisions.

3.5 PREDICTIVE VALUE

The *positive and negative predictive values* correspond to *posterior probabilities*; predictive values provide a sense for the classification performance of a test once the results are in hand. For example, a patient classified as disease-positive by a test with a high positive predictive value will likely actually have the disease.

The positive predictive value measures the probability of the presence of disease in a patient who tests positive:

$$\text{Positive predictive value} = \frac{\text{True positive decisions}}{\text{All positive decisions}}$$

and from the example in Figure 2:

$$\text{Positive predictive value} = \frac{38}{52} = 0.73$$

The negative predictive value provides a complement to the positive:

$$\text{Negative predictive value} = \frac{\text{True negative decisions}}{\text{All negative decisions}}$$

So that, from Figure 2:

$$\text{Negative predictive value} = \frac{41}{48} = 0.85$$

In the example, the new test is shown to be more useful when negative test results are in hand; that is, 85% of those testing negative actually do *not* have an infection, while only 73% of those with a positive result *do* have an infection.

Like sensitivity and specificity, the predictive values do not reflect the simple classification performance of the test. And, like sensitivity and

specificity, predictive values are influenced by the base rate. However, predictive values are also influenced by sensitivity and specificity. Figure 4 shows the relationship between sensitivity, specificity, and base rate. One can see from this figure that at a given base rate, the positive predictive value depends on the sensitivity and specificity of a test. A highly sensitive and specific test will have a higher positive predictive value that one which is less sensitive and specific. The inverse relationship holds true for negative predictive values.

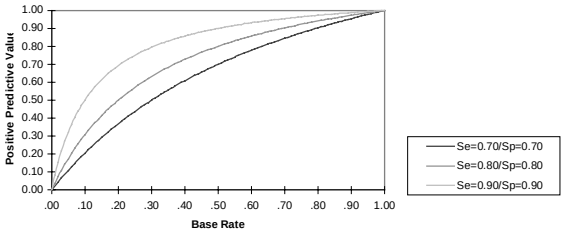

Figure 4. Relationship between positive predictive value, base rate, and sensitivity/specificity.

3.6 THE RECEIVER-OPERATING CHARACTERISTIC CURVE

3.6.1 Introduction

The ROC curve is created by plotting the true positive rate (sensitivity) on the vertical axis against the false positive rate (1-specificity) on the horizontal axis. For a 2x2 table, only one point will be plotted. Figure 5 shows the ROC curve for the example in Figure 2. The significance of the straight line through the origin will be discussed below.

The ROC curve is also employed to determine

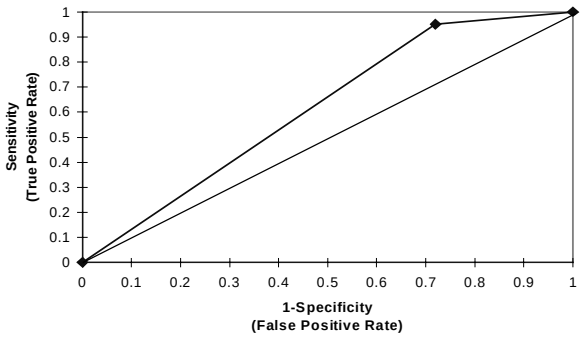

Figure 5. Single-point ROC curve for example data

the overall usefulness of a diagnostic test. In order to determine whether a test is useful, it must be evaluated for its discrimination accuracy, or its ability to classify normal and abnormal patients. The measure used for this purpose is the area under the ROC curve.

3.6.2 The area under the ROC curve (θ)

While the graphical representation of the ROC curve is of interest in determining appropriate diagnostic cutoffs for a given test, the *area under the ROC curve* (θ) is important for demonstrating the ability of the test to classify both true positives and true negatives, simultaneously, as a *single measure*. The area under the ROC curve has been used extensively in medical decision making as a standard method for evaluating diagnostic test performance (Good et al 1990; Somoza et al 1990; and many others). In addition, it has been proposed for use in knowledge discovery and data mining domains (Provost and Fawcett, 1997).

The area under the ROC curve represents the probability of a true response in a two-alternative forced-choice (Yes-No) task; thus, the quantity (1-area) is the false alarm rate (Green and Swets 1966). The ROC curve for a test which classifies well will have a "shoulder" closer to the upper left-hand corner of the plot, and farther from the 45-degree diagonal (shown in Figure 5). This test would have a high sensitivity and a high specificity and, as a result, a higher θ. A test which contains no information would plot on the 45-degree diagonal through the origin. The area under such a “curve” would be 0.50, indicating that the test would discriminate only as well as a coin-flip. A nonparametric method based on the Wilcoxson statistic (W) and its standard error (SE_W) is regularly used in approximating θ and its standard error, SE_θ (Hanley and McNeil 1982). The plausibility of calculating θ as derived from a single point has been shown (McNichol 1972).

Software tools exist for constructing ROC curves, calculating areas, and comparing them by various means (Centor 1990; Metz 1993). In addition, the reader is referred to Hanley and McNeil (1982) and McNeil and Hanley (1984) for a thorough description of the algorithm for approximating the area under the ROC curve and its standard error.

3.7 THE INDETERMINANT RATE

When evaluating the performance of any learning system, it is essential to consider the proportion of cases that cannot be matched by the system (and thus cannot be classified); this proportion is referred to as the *indeterminant rate*, or *IR*. The IR is calculated as:

$$\text{Indeterminant Rate} = \text{IR} = \frac{\text{Number of unclassifiable cases}}{\text{Total number of cases to be classified}}$$

The IR should be used to refine the results obtained for the other parameters in the toolkit, by providing an indication of the denominator upon which these metrics were based. For example, it would be incorrect to report a CA or θ without knowing the number of classifiable observations (or *denominator*) on which these metrics are based. While CA and θ could be interpreted in conjunction with IR as a separate measure, it would be cumbersome to do so. As an alternative, single indices of accuracy and θ, corrected for the IR, can be created as follows:

$$\text{Corrected crude accuracy} = \text{CA}_{corr} = \frac{\text{Crude accuracy}}{1 + \text{IR}}$$

and

$$\text{Corrected } \theta = \theta corr = \frac{\theta}{1 + \text{IR}}$$

Thus, a crude accuracy of 0.95 obtained with an IR of 0.30 would yield a CA_{corr} of:

$$\text{CA}corr = \frac{\text{Crude accuracy}}{1 + \text{IR}} = \frac{0.95}{1 + 0.30} = 0.73$$

This example clearly shows the effect of a large IR on an apparently high crude accuracy; applying the correction results in a 23.2% reduction in the CA, but the CA_{corr} is a more valid reflection of accuracy. The IR should be applied to all of the metrics in the toolkit, so that they are comparable when evaluating different LCS, or different trials of the same LCS.

3.8 THE TOOLKIT AND LCS RESEARCH

This work proposes that the metrics described above can serve as a useful collection of tools for use in evaluating LCS performance in two-choice decision problems. Typically, these problems will be single-step (without long payoff chains), and implemented using a stimulus-response LCS such as Wilson's BOOLE (Wilson 1987) or its descendents. The remainder of this paper focuses on the use of this toolkit in evaluating the performance of such a LCS.

4.0 METHODS: AN APPLICATION OF THE LCS METRIC TOOLKIT

4.1 EpiCS: TESTBED LEARNING CLASSIFIER SYSTEM

An object-oriented version of NEWBOOLE (Bonelli et al (1990), called *EpiCS*, was created and used as the classifier system in this investigation. EpiCS departed from NEWBOOLE and its predecessor, BOOLE (Wilson 1987) on several features: population size, algorithms for controlling under- and over-generalization, and a methodology for determining risk as a measure of classification. In addition, EpiCS provides support for the tools described above, and uses many of them in its graphical display. EpiCS has been described in detail elsewhere (Holmes 1996; Holmes 1997).

4.2 TESTBED DATA

In order to provide consistent, manipulable data, four sham datasets consisting of 15 demographic, medical history, and exposure variables, one outcome variable, and 500 observations were created using the random data generator routines supplied with the EpiInfo (Dean et al 1990) epidemiologic analysis software package. These datasets represent data on hepatocellular (liver) carcinoma in a group of individuals. All variables were coded dichotomously, with 0s or 1s used to indicate the absence or presence, respectively, or value categories, of a variable. Each dataset represented a different positive base rate: 0.50, 0.25, 0.15, and 0.10.

Training and testing sets were created by randomly selecting records the dataset at a sampling fraction of 0.50 without replacement; thus, training and testing sets were equal in size, equal in number of positive and negative examples, and mutually exclusive.

4.3 TRAINING-TESTING SEQUENCE

A total of 20 trials, each consisting of a training epoch and a testing epoch, were performed. During the training epoch, cases were selected in random order from the training set and presented to the system over a total of 30,000 iterations; a single case presentation comprised one iteration. The system was evaluated by calculating the metrics in the toolkit during training, at every 100th iteration, to monitor learning performance. This was achieved by testing the system with every case in the *training* set. At the conclusion of the training epoch, the system was tested

with each case in the *testing* set, and its classification performance was evaluated by the metrics provided in the toolkit. The results reported here focus on both the training and testing epochs of the trials.

5.0 RESULTS AND DISCUSSION

5.1 TRAINING

EpiCS supports a graphical display of the training epoch, including textual output of the parameters in the toolkit, as well as graphical display of the CA_{corr} and θ_{corr}. These displays are not ROC curves; rather, each data point on the plot represents the calculation performed at each 100th iteration (as described above) to obtain these metrics. Plots of the CA_{corr} and θ_{corr} obtained over the training epochs at the four positive base rates are shown in Figures 6-9.

The most noteworthy feature of these four figures is the progressive separation between CA_{corr} and θ_{corr} in moving to smaller positive base rates. This clearly demonstrates the problem with using CA (corrected or not) as a performance measure at positive base rates less than or greater than 0.50. If one were using a traditional measure of convergence, such as appearance of curve shoulder, to determine the end of a training epoch, the epoch would be brought to a premature end at lower positive base rates. In addition, CA_{corr} appears to provide a stable estimate of performance; the overall shape of the plot of this measure changes only slightly in moving from a high to a low positive base rate. However, CA_{corr} is a deceptive measure, in that its computation is based on total number "correct" and does not account for the two types of error that occur in classification problems. At lower positive base rates, the effect of erroneous (false negative) decisions is diluted by the large number of gold-standard negatives which are incorporated into the denominator in calculating CA_{corr}.

However, θ_{corr} is shown to change over all base rates; in fact, it appears to be progressively unstable with decreasing positive base rate. This also is deceptive; θ_{corr} is actually the more accurate measure of classification performance, in that it does account for the two types of error. Furthermore, the components (sensitivity and specificity) used to construct the ROC curve, and consequently, θ_{corr}, are based on denominators that are similarly affected by changes in base rate. Because sensitivity and specificity are equally affected by base rate, θ_{corr} is not itself so influenced, as it is essentially a proportion of these two metrics.

In summary, the application of the metrics in the toolkit during training provides a sound indication of LCS performance during this epoch. As a result, it is possible to identify traditional measures of convergence, such as curve shoulder; these measures will be more valid, as they are based on θ_{corr} which is more robust than $CA_{corr.}$

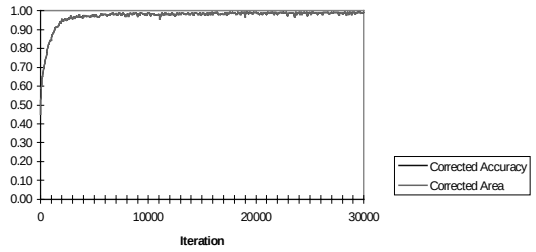

Figure 6. Training epoch at positive base rate=0.50

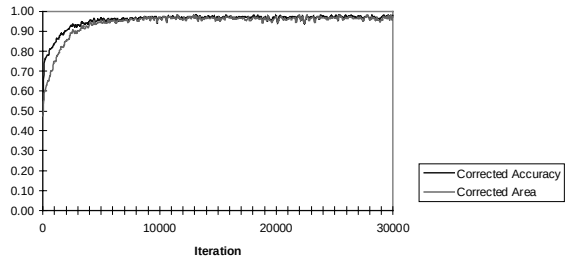

Figure 7. Training epoch at positive base rate=0.25

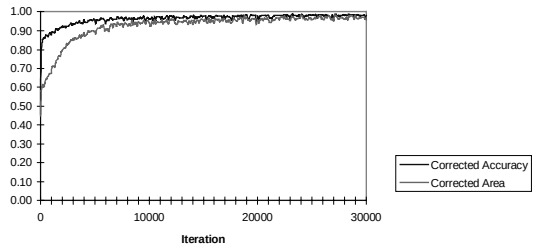

Figure 8. Training epoch at positive base rate=0.15

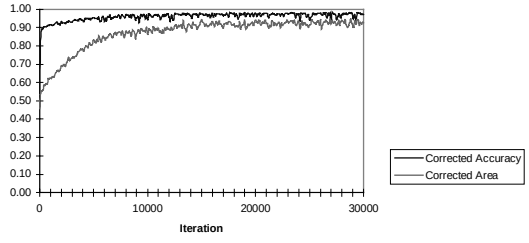

Figure 9. Training epoch at positive base rate=0.10

Table 1. Results at testing for each base rate.

Positive base rate	CA_{corr}	θ_{corr}	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
0.50	0.95 (0.02)	0.95 (0.02)	0.96 (0.03)	0.93 (0.02)	0.93 (0.02)	0.96 (0.03)
0.25	0.92 (0.03)	0.89 (0.05)	0.84 (0.10)	0.95 (0.02)	0.85 (0.05)	0.95 (0.03)
0.15	0.90 (0.02)	0.84 (0.09)	0.76 (0.18)	0.92 (0.02)	0.63 (0.07)	0.96 (0.03)
0.10	0.92 (0.04)	0.78 (0.06)	0.61 (0.14)	0.95 (0.05)	0.64 (0.15)	0.96 (0.01)

5.2 TESTING

The results obtained at testing for each base rate are shown in Table 1. The results are averaged over 20 trials; numbers in parentheses represent one standard deviation. The progressive divergence between CA_{corr} and θ_{corr} observed during the training epoch with smaller positive base rates is maintained at testing. In addition, a similar pattern is seen with sensitivity and positive predictive value, as expected. These patterns would be reversed, were the base rate scheme reversed, i.e., using a gradient from 0.50. to 0.90. The data in Table 1 clearly demonstrate the relationship between base rate and crude accuracy, as illustrated in Figure 3, and predictive values, as illustrated in Figure 4.

The values of the metrics in Table 1 are useful for interpreting the performance of EpiCS. For example, at positive base rate=0.50, it is clear that the system performs quite well. One may infer that at this base rate, EpiCS will accurately detect both positives and negatives; this is evident from the sensitivity and specificity, respectively. In addition, given a positive (or negative) classification, one can have confidence that EpiCS has made an accurate decision, and this is borne out by the positive (or negative) predictive values at this base rate.

However, at lower positive base rates, one should lose confidence in positive classifications made by EpiCS. At base rates less than 0.25, the probability that a positive decision made by EpiCS is incorrect is as high as 0.37 (1-positive predictive value). In addition, the prior probability (sensitivity) at these base rates indicates that if EpiCS were to be considered for use on a data set with a small number of positive examples, it would misclassify as many as 39% of cases presented at testing. EpiCS is very good at accurately classifying both positive and negative cases after training, as long as the positive base rate is at least 0.25 and no greater than 0.75.

While the toolkit is very useful during the training epoch, it is its application to the testing epoch that provides the most valuable information to a user of a LCS. For example, a LCS may be used as rule-base system for detecting a disease in a patient. This assumes that the LCS is fully trained and tested, and that the patient represents a novel case. The metrics obtained during the testing epoch will provide the user with the information needed to evaluate whether the LCS will accurately detect positive and negative cases (sensitivity and specificity, respectively. Second, the positive and negative predictive values will indicate if the results obtained from using the LCS on the patient are accurate for both types of decision, positive and negative. Finally, the θ_{corr} provides a well-known and recognized indicator of overall classification performance.

6.0 CONCLUSIONS AND RECOMMENDATIONS

This investigation examined the application of tools commonly used in medical decision making to evaluating LCS performance during training and testing in supervised learning of two-choice problems. Previously, LCS performance was commonly reported as “percent correct” or “number of steps required to reach a goal.” The toolkit proposed in this paper adds substantially to the metrics LCS researchers can use in evaluating classification performance. The tools are easy to implement and interpret, while providing robust indicators of a variety of performance dimensions.

Although the toolkit is useful for the accurate determination of LCS performance in two-class problems, it will not solve the “base rate problem.” Holmes (1998) showed that employing a differential penalty to false positives and false negatives could enhance learning rate in a LCS, particularly EpiCS, there was no evidence that this improved classification performance on testing. However, it is possible that “boosting,” a method for enhancing predictive accuracy in data mining, can be employed with some benefit. By weighting the training cases, such that cases with lower class prevalence would be presented to the system more proportionally more frequently than those with higher prevalence, an improvement in classification performance might be realized. Clearly, this is an area for future investigation.

Another area needing attention is the issue of multi-class decisions. As the number of classes increases in a given problem, the possibility of

unevenly distributed classes increases, resulting in accuracy determinations that will be more heavily skewed than those seen in two-decision problems. Although not the focus of this paper, the toolkit will address these problems as well; multi-class domains are very frequent in medical decision making, and these tools have a long history of application in these domains. For example, ROC curves are commonly constructed from several points, each point representing a classification category.

Finally, there is the issue of "continuous decisions." When the output of a LCS is continuous, such as a probability, the toolkit will need to include a component to cut the output at regular intervals. These "cutpoints" serve the same function as the categories in multi-class decisions, as described above. Holmes (1997) described this method as applied to disease risk, which is an example of a "continuous decision."

Acknowledgment

The author gratefully acknowledges Stewart W. Wilson, PhD, of Prediction Dynamics, Concord, MA, for his suggestions that led to this paper, and to the comments of anonymous reviewers that were helpful in refining it.

References

Bonelli, P.; Parodi, A.; Sen, S.; Wilson, S. NEWBOOLE: A fast GBML system. Porter, B.; Mooney, R. Machine Learning: *Proceedings of the Seventh International Conference*; 1990 Jun 21; Austin, Texas. San Mateo, CA: Morgan Kaufmann Publishers, Inc.; 1990: 153-159.

Centor, R.; Keightley, GE. ROC ANALYZER for the IBM PC. Medical College of Virginia; 1990.

Dean, AD; Dean, JA; Burton, JH; Dicker, RC. Epi Info, Version 5: a word processing, database, and statistics program for epidemiology on microcomputers. Centers for Disease Control, Atlanta, Georgia, 1990.

Fletcher, RH; Fletcher, SW; Wagner, EH. *Clinical Epidemiology*. *The Essentials*. Baltimore: Williams and Wilkins, 1988.

Good, WF; Gur, D; Straub, WH.; Feist, JH. Comparing imaging systems by ROC studies. Detection versus interpretation. *Invest Radiol*. 1989 Nov; 24(11): 932-3.

Green, DM; Swets, JA. *Signal Detection Theory and Psychophysic*s. New York: John Wiley Sons; 1966.

Hanley, JA; McNeil, BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982 143:29-36.

Hennekens, CH; Buring, JE; Mayrent, SL (ed.). *Epidemiology in Medicine*. Boston: Little, Brown and Company; 1987.

Holmes JH: Differential negative reinforcement improves classifier system learning rate in two-class problems with unequal base rates. Koza JR, Banzhaf W, Chellapilla K, Deb K, Dorigo M, Fogel DB, Garzon MH, Goldberg DE, Iba H, and Riolo, R, eds., *Genetic Programming 1998: Proceedings of the Third Annual Conference*, Morgan Kaufmann, San Francisco, 1998.

Holmes JH: Discovery of disease risk with a learning classifier system. In T. Baeck, ed., *Proceedings of the Seventh International Conference on Genetic Algorithms (ICGA97)*, Morgan Kaufmann, San Francisco, 1997, pp. 426-433.

McNichol, DA. *Primer of Signal Detection Theory*. London: George Allen and Unwin, Ltd.; 1972.

Metz, CE; Shen, J-H; Kronman, HB. LabROC4: A program for maximum likelihood estimation of a binormal ROC curve and its associated parameters from a set of continuously-distributed data. University of Chicago; 1993.

Provost, F; Fawcett, T. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. Heckerman, D; Mannila, H; Pregibon, D; Uthurusamy, R. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*; 1997 Aug 14; Newport Beach, CA. Menlo Park, CA: AAAI Press, 1997; 43-48.

McNeil, BJ; Hanley, JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Med Decis Making*. 1984; 4:137-150.

Somoza, E.; Soutullo-Esperon, L.; Mossman, D. Evaluation and optimization of diagnostic tests using receiver operating characteristic analysis and information theory. *Int J Biomed Comput*. 1989 Sep; 24(3): 153-89.

Wilson, SW. Classifier systems and the animat problem. *Machine Learning*. 1987; 2: 199-228.