Evaluating Misclassifications in Imbalanced Data

William Elazmeh[1], Nathalie Japkowicz[1], and Stan Matwin[1,2]

[1] School of Information Technology and Engineering
University of Ottawa, K1N 6N5 Canada
{welazmeh, nat, stan}@site.uottawa.ca
[2] The Institute of Computer Science, Polish Academy of Sciences, Poland

Abstract. Evaluating classifier performance with ROC curves is popular in the machine learning community. To date, the only method to assess confidence of ROC curves is to construct ROC bands. In the case of severe class imbalance with few instances of the minority class, ROC bands become unreliable. We propose a generic framework for classifier evaluation to identify a segment of an ROC curve in which misclassifications are balanced. Confidence is measured by Tango's 95%-confidence interval for the difference in misclassification in both classes. We test our method with severe class imbalance in a two-class problem. Our evaluation favors classifiers with low numbers of misclassifications in both classes. Our results show that the proposed evaluation method is more confident than ROC bands.

1 Motivation

Recently, the machine learning community has increased the focus on classifier evaluation. Evaluation schemes that compute accuracy, precision, recall, or F-score have been shown to be insufficient or inappropriate [1,2]. Furthermore, the usefulness of advanced evaluation measures, like ROC curves [3,2,4] and cost curves [5,6], deteriorates in the presence of a limited number of positive examples. The need for confidence in classifier evaluation in machine learning has lead to the construction of ROC confidence bands. Methods in [7,8] construct ROC bands by computing confidence intervals for points along the ROC curve. These methods are either parametric (making assumptions of data distributions), or non-parametric and rely on carefully crafted sampling methods. When faced with severe class imbalance and with a limited number of instances in the minority class, sampling methods become unreliable, especially when the data distribution is unknown [8]. In fact, with severe imbalance, the entire issue of evaluation becomes a serious challenge even when making assumptions of data distributions [9]. In contrast, biostatistical and medical domains impose strong emphasis on error estimates, interpretability of prediction schemes, scientific significance, and confidence [10] whilst machine learning evaluation measures fail to provide such guarantees. Consequently, the usefulness of some machine learning algorithms remains inadequately documented and unconvincingly demonstrated. Thus, despite their interest in using learning algorithms, biostatisticians remain

J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): ECML 2006, LNAI 4212, pp. 126–137, 2006.
© Springer-Verlag Berlin Heidelberg 2006

Table 1. The statistical proportions in a confusion matrix

	Predicted +	Predicted -	total
Class +	a (q_{11})	b (q_{12})	a+b
Class -	c (q_{21})	d (q_{22})	c+d
total	a+c	b+d	n

skeptical of their evaluation methods and continue to develop customized statistical tests to measure characteristics of interest. Our work adopts Tango's test [11] from biostatistics in an attempt to provide confidence in classifier evaluation. Tango's test is a non-parametric confidence test designed to measure the difference in binomial proportions in paired data. This test is shown in [12] to be reliable and robust with power and coverage probability to produce confidence and significance.

Computing the confidence based on the positive or negative rates (using a or d of table 1) can be influenced by class imbalance in favor of the majority class. Alternatively, applying a statistical significance test to those entries (b or c) that resist such influence may provide a solution. Hence, to counter the class imbalance, particularly when the number of instances in the minority class is very small, we use Tango to favor classifiers with similar normalized number of errors in both classes, rather than similar error rates. This solution assumes that the classifier performs reasonably well, in the sense that, it can at least classify the majority class with high accuracy. Therefore, a large portion of instances in the majority class is correctly classified, and the imbalance has no influence on the error values (b or c). Consequently, any evaluation measure that employs rates (false positive or false negative), such as ROC curves, is influenced by data imbalance, while the error analysis we propose is not. Our approach is based on measuring just the error of classification, and therefore, to capture a fuller evaluation of the classifier, we need to combine Tango's analysis together with another evaluation measure that measures how well the classifier performs. As we measure negatively, we need to use our approach along with another measure (eg. ROC) to evaluate positively.

In this paper; (1) we propose a framework for classifier evaluation that identifies confident points along an ROC curve using a statistical confidence test. These points form a balanced misclassification segment on the ROC curve to which we recommend restricting the evaluation. (2) Although our framework can be applied to any data, this work focuses on the presence of severe imbalance (with a very small number of instances in the minority class) where ROC bands, ROC curves and AUC struggle to produce meaningful assessments. (3) We produce a representation of classifier performance based on the average difference in misclassifications and the area under the balanced misclassification segment of the ROC curve. We present experimental results that show the effectiveness of our approach compared to ROC bands, ROC curves, and AUC.

Having motivated this work, subsequent sections present discussions of classification error proportions in both classes, our evaluation framework, and our

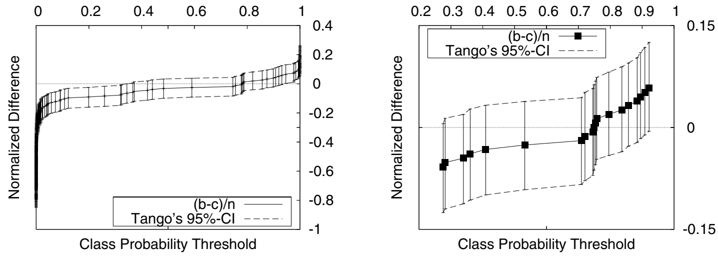

Fig. 1. $\frac{b-c}{n}$ and Tango's 95%-confidence intervals for classification points. Left: all classification points. Right: only points whose Tango's intervals contain 0 difference.

experimental results followed by conclusions and future work. In the appendix, we briefly describe Tango's statistical test of confidence.

2 The Difference in Classification Errors

Common classifier performance measures in machine learning estimate classification accuracy and/or errors. ROC curves provide a visualization of a possible trade-off between accuracy and error rates for a particular class. For the confusion matrix presented in table 1, the ROC curve for the class + plots the true positive rate $\frac{a}{a+b}$ against the false positive rate $\frac{c}{c+d}$. When the number of positive examples is small and is significantly lower than the number of negative examples, the row totals $a + b << c + d$. When changing the class probability threshold, the rate of change in the true positive rate climbs faster with each example than that of the false positives (due to using c and d). This inconsistent rate of change gives the majority class $(-)$ a clear advantage in the rates calculated for the ROC curve. Ideally, a classifier classifies both classes proportionally, but due to the severe imbalance along with a small number of instances of the minority class, comparing the rates of accuracy and/or errors on both classes does not evaluate proportionally. We propose to favor the classifier that performs with similar number of errors in both classes to eliminate the use of the number of correctly classified examples (a and d) in the evaluation to avoid a large portion of examples in the majority class. In fact, our approach favors classifiers that have lower difference in misclassifications in both classes, $\frac{b-c}{n}$. Furthermore, we normalize entries in the confusion matrix by dividing by the number of examples n so the difference $\frac{b-c}{n}$ remains within $[-1, +1]$.

ROC curves are generated by classifying examples while increasing class probability threshold T. When $T = 0$, all data examples are classified as +, thus, $a = |+|$ (the number of positives), $b = 0$, $c = |-|$, $d = 0$, and $\frac{b-c}{n} \in [-1, 0]$. Similarly, for $T = 1$, all examples are classified as $-$, then, $a = 0$, $b = |+|$, $c = 0$, $d = |-|$, and $\frac{b-c}{n} \in [0, +1]$. In fact, these two extreme negative and positive values of $\frac{b-c}{n}$ depend on class distributions in the data. Within these two extremes, $\frac{b-c}{n}$ exhibits a monotone behavior as the threshold varies from 0

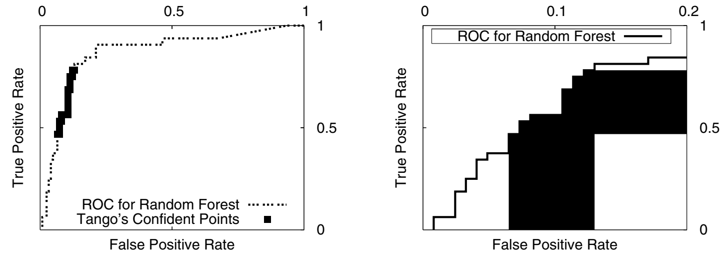

Fig. 2. On the left is a sample balanced misclassification segment and on the right is the area under this segment

to 1. This is illustrated in figure 1. For each threshold value $T := 0$ to 1, the classification produces a confusion matrix a, b, c, d. Initially, a and c are at their maximum values, while b and d are 0. As T increases, examples are classified in any combination of three possibilities; (1) c decreases when false positives become correctly classified, (2) b increases when true positives become misclassified, (3) or, b and c remain unchanged because examples are correctly classified. Since c never increases, b never decreases, and n is constant, then $\frac{b-c}{n}$ exhibits a monotone non-decreasing behavior for a classifier on a set of data. Our evaluation method computes Tango's 95%-confidence intervals for $\frac{b-c}{n}$ for ROC points. Those points whose confidence intervals include the value zero, show no evidence of statistically significant $\frac{b-c}{n}$ and are considered confident. This is explained in more details in the next section. In addition, Tango's test is presented in [11] and is reviewed in the appendix of this paper.

3 The Proposed Method of Evaluation

The proposed evaluation method consists of four steps: **(1)** Generate an ROC curve for a classifier K applied on test examples D with increasing class probability thresholds t_i (0 to 1). **(2)** For each resulting point (a confusion matrix along the ROC curve), apply Tango's test to compute the 95%-confidence interval $[u_i, l_i]$, within which lies the point of the observed normalized error difference $\frac{b_i-c_i}{n}$. If $0 \in [u_i, l_i]$, then this point is identified as a confident point and is added into the set of points S. Points in S form the ROC segment illustrated in the left plot of figure 2: we call it the balanced misclassification segment. This framework is generic and accommodates a test of choice provided that it produces a meaningful interpretation of results. **(3)** Compute $SAUC$ the area under the segment S as shown in the right plot of figure 2. **(4)** Compute $AveD$ the average normalized difference $(\frac{b-c}{n})$ for all points in S. In our experiments, we plot the area under the balanced misclassification segment ($SAUC$) against the average observed misclassification difference ($AveD$). Lower $AveD$ values suggest lower misclassification difference and higher $SAUC$ values indicate larger balanced misclassification segment. An effective classifier shows low $AveD$ and high $SAUC$.

Table 2. UCI data sets [13] and their class distributions $|+|/|-|$

	dis	hypothyroid	sick	sick-euthyroid	SPECT	SPECTF
training	45/2755	151/3012	171/2755	293/2870	40/40	40/40
testing	13/959	–	13/959	–	15/172	55/214

4 Experiments

Having presented our evaluation framework, we now present an overview of our experiments and their data sets followed by an assessment of results to motivate conclusions. The data sets, listed in table 2, are selected from the UCI-Machine Learning repository [13] and consist of examples of two-class problems. They are severely imbalanced with the number of positive examples reaching as low as 1.4% (`dis`) and not exceeding 26% (`spectf`). Only (`spect`) and (`spectf`) data sets have a balanced training set and imbalanced testing set. On these data sets, we train four classifiers and compare their performances as reported by the ROC, by the AUC, and by our method. If testing data sets are unavailable, we use cross-validation of 10 folds. Using `Weka 3.4.6` [14], we build a decision stump classifier without boosting (`S`), a decision tree (`T`), a random forest (`F`), and a Naive Bayes (`B`) classifier. The rationale is to build classifiers for which we can expect a ranking of performance. A decision stump built without boosting is a decision tree with one test at the root (only 2 leave nodes) and is expected to perform significantly worse than a decision tree. Relatively, a decision tree is a stronger classifier since it is more developed and has more leave nodes that cover the training examples. The random forest classifier is a reliable classifier and is expected to outperform a single decision tree. Finally, the naive Bayes classifier tends to minimize classification error and is expected to perform reasonably well when trained on a balanced training set.

We first investigate the usefulness of ROC confidence bands on data with imbalance. Figure 3 shows the ROC confidence bands for our four classifiers on the most imbalanced `dis` data set. These bands are generated using the empirical fixed-width method [8] at the 95% level of confidence. We claim that with severe imbalance and a very small minority class, sampling-based techniques do not work. Clearly, the generated bands are very wide and contain more than 50% of the ROC space proving that they are not very useful. This result is also consistent on the other data sets. Given this failure of the ROC bands, we propose to use Tango's test which is designed to accommodate a small size of proportions (the minority class) while resisting the influence of class imbalance. As shown in the Appendix, Tango test is related to McNemar test which has been used to rank performance of classifiers in [15]. In addition, Newcombe in [12] (which compares 10 different statistical methods that Tango is based on) shows that Tango's confidence intervals are robust (they do not collapse in boundary conditions and do not produce tethering points) and reliable with good probability coverage.

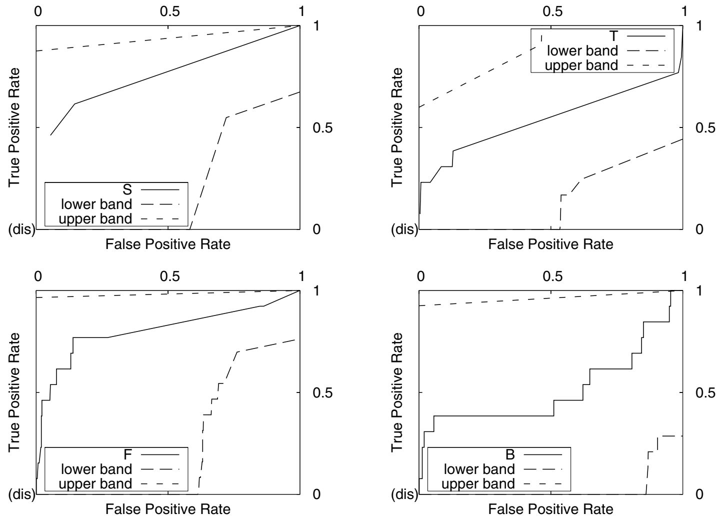

Fig. 3. ROC confidence bands for decision stump (S), decision tree (T), random forest (F), and naive Bayes (B) on (`dis`) data set. The bands are wide and are not very useful.

Next, we consider the ROC curves of our four classifiers on all data sets shown in figure 4. Recall, ROC curves are compared by being more dominantly placed towards the north-west of the plot (higher true positive rate and lower false positive rate). We observe that the decision stump (S) performs the same or better than the decision tree (T) on all data sets. In addition, the random forest (F), consistently, outperforms the naive Bayes (B). In fact, (F) shows the best performance on most data sets. When we consider the AUC values of these classifiers, shown in table 3, (S) has similar or higher AUC values than (T). Furthermore, the AUC of (F) is, clearly, higher than that of the others on most data sets (the bold numbers in table 3). When trained on a balanced data set (`SPECT`), (F) and (B) classifiers perform significantly better than the others.

In contrast, the results obtained by our proposed evaluation measure are presented in figure 5. Each of the six plots in the figure reports our evaluation of the four classifiers on each data set. The x-axis represents the average normalized misclassification difference $\frac{b-c}{n}$ for those points on the balanced misclassification segment of the ROC curve. The y-axis represents the area under this segment. Classifiers placed towards the top-left corner perform better (bigger area under the balanced misclassification segment and less difference in classification error) than those placed closer to the bottom right corner (smaller area and higher difference in misclassifications). Classifiers that fail to produce confident points on their ROC curves are excluded from the plots. The decision stump (S) fails to produce confident points along its ROC, therefore, it does not appear in any of the plots in the top row of figure 5. This is consistent with our expectation of it being less effective. In fact, plots in the bottom row of the same figure show that (S) also performs poorly producing higher misclassification difference.

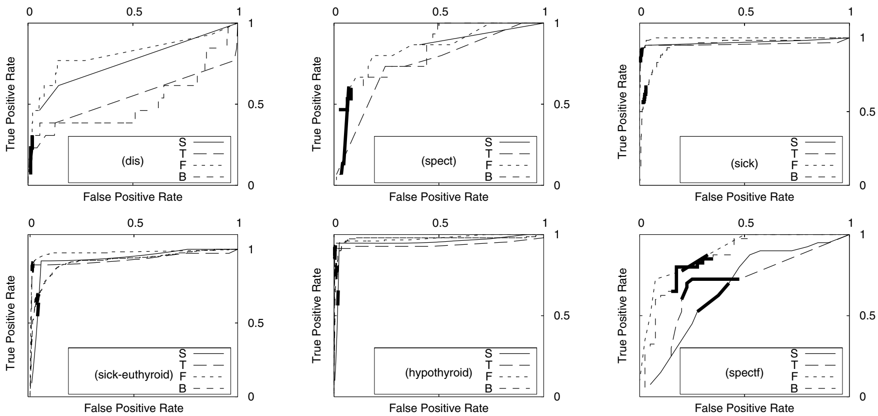

Fig. 4. ROC curves for decision stump (S), decision tree (T), random forest (F), and naive Bayes (B) on all data set. The dark segments are Tango's confident points.

Table 3. AUC values for classifiers (S), (T), (F), and (B) on our data sets

Classifier	dis	hypothyroid	sick	sick-euthyroid	spect	spectf
S	0.7517	0.9491	0.9523	0.9312	0.7298	0.6744
T	0.5408	0.9360	0.9559	0.9296	0.7453	0.6900
F	**0.8051**	**0.9784**	**0.9966**	**0.9777**	0.8326	**0.8925**
B	0.5164	0.9720	0.9460	0.9215	**0.8347**	0.8575

In fact, even when (S) has slightly higher SAUC than (T), in the bottom left and middle plots of figure 5, (S) still shows a significantly higher difference in misclassification than that of (T). The tree (T), on the other hand, performs well in most cases particularly in the top and bottom right plots of figure 5. (T) certainly outperforms the (S) which contradicts observations based on the ROCs and AUCs. Furthermore, (T) fails to produce confident points on the (`spect`) data set (top middle plot of the same figure). Perhaps, since (`spect`) is a binary data set extracted from the continuous (`spectf`) set, this may suggest that the extraction process hinders the decision tree learning. (F) and (B) classifiers appear reasonably consistent on all data sets with (B) being particularly strong on the (`dis`) data set. However, the surprise is (B) showing significantly higher SAUC than (F) in the top and bottom right plots of figure 5.

Our results, clearly, contradict conclusions based on the ROC and AUC evaluations. Therefore, we investigate those points along the balanced misclassification segment for two situations. First, when the four classifiers are trained and tested on imbalanced `dis` data sets, and second, when the four classifiers are trained on a balanced training set and are tested on an imbalanced testing `SPECTF` data set. For the first situation (`dis` data sets), the ROC curves reveal that three

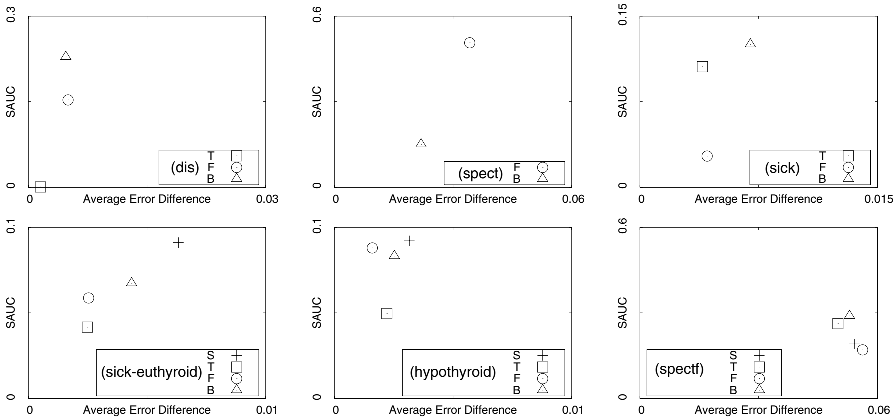

Fig. 5. Our evaluation for decision stump (S), decision tree (T), random forest (F), and naive Bayes (B) on our data sets. The y-axis shows the area under the balanced misclassification segment (SAUC) and the x-axis shows the average observed normalized misclassification difference $\frac{b-c}{n}$.

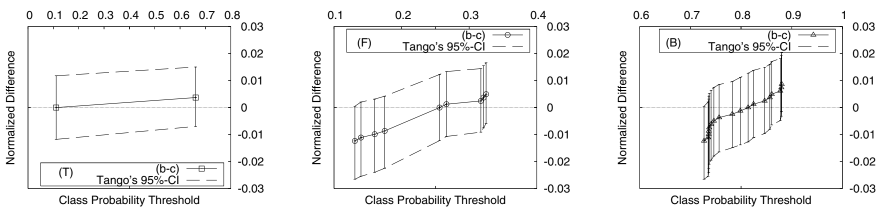

Fig. 6. Tango's 95%-confidence intervals for classification points for decision tree (T), random forest (F), and naive Bayes (B) on (`dis`) set. The center points are ($\frac{b-c}{n}$). (B) produces the most points that satisfy the Tango test.

of the classifiers produce balanced misclassification points in the bottom left section of the ROC space (see the bold segments in the top left plot of figure 4). These points are detected by our method at the 95% level of confidence and are consistent with having severely imbalanced data sets with very few positive examples. When we consider the corresponding Tangos 95%-confidence intervals for these classifier (see figure 6), we see that (T) produces few points (only two) which cover a wider range of probability threshold (0.1 to 0.65 on the x-axis of the left plot). (T) produces only two points which may be due to the very low number of positive examples. Alternatively, despite generating many more confident points, (F) and (B) classifiers show higher variations of misclassification difference for a much narrower range of thresholds values. The top left plot of figure 5 shows that (B) and (F) have a higher SAUC values than (T) which has a significantly lower misclassification difference. At the least, this indicates a distinction between these classifiers.

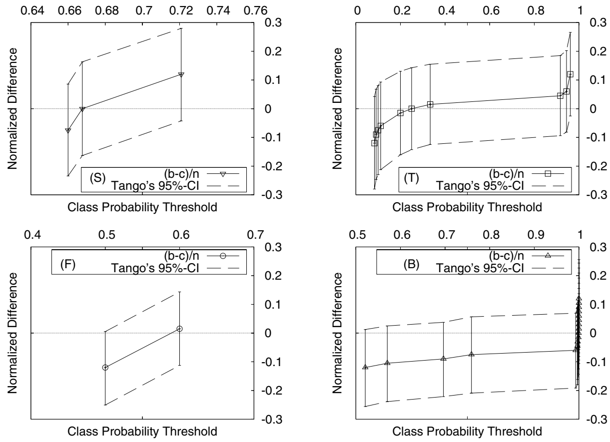

Fig. 7. Tango's 95%-confidence intervals for classification points for decision stump (S), decision tree (T), random forest (F), and naive Bayes (B) on (`spectf`) set. The center points are ($\frac{b-c}{n}$). (T) and (B) produce more points that satisfy Tango's test.

For the second situation (`SPECTF` data sets), the ROC curves in the bottom right plot of figure 4 show that both (F) and (B) dominate (T) and (S). Our method in the bottom right plot of figure 5 suggest that both (T) and (B) outperform (F) and (S). Tango's 95%-confidence intervals of their classification points (shown in figure 7) show that (T) and (B) produce the most number of points on their balanced misclassification segment with low misclassification difference. Also in the same figure, (T) and (B) produce classification points that have exactly zero misclassification difference while the other two come close to the zero misclassification difference.

5 Conclusions and Future Work

We propose a method to address classifier evaluation in the presence of severe class imbalance with significantly fewer positive examples. In this case, our experiments show that ROC confidence bands fail to provide meaningful results. We propose a notion of statistical confidence by using a statistical tests, borrowed from biostatistics, to compute the 95%-confidence intervals on the difference in misclassification. This work presents error-based analysis (using Tango's test) which aims to balance misclassifications. To capture a fuller evaluation of the classifier, we need to combine Tango's analysis together with another evaluation measure (eg. ROC) that measures how well the classifier performs. Our method plots of the trade-off between misclassification difference and area under the balanced misclassification segment of the ROC curve. Our experiments show that our method is more reliable than general ROC and AUC measures.

In the future, it can be useful to compute confidence bands or intervals for these proposed confident ROC segments. This remains a difficult task because the confidence in our method is computed on the misclassification difference which may not map easily to the ROC space. We plan to investigate the feasibility of mapping the confidence intervals from this work into the ROC space. This may be interesting particularly when there is no danger of imbalance. Although this work addresses the case of severe imbalance in the data, Tango's test of confidence can still be applied to balanced data sets. We plan to explore our framework in balanced situations with the aim to drive useful and meaningful evaluation metrics to provide confidence and reliability.

Acknowledgments

The authors thank James D. Malley at the National Institute of Health for his suggestions and communications. In addition, we acknowledge the support of The Natural Sciences and Engineering Research Council of Canada (NSERC) and of the Ontario Centers of Excellence (OCE).

References

1. Ling, C.X., Huang, J., Zang, H.: Auc: a better measure than accuracy in comparing learning algorithms. Canadian Conference on AI (2003) 329–341
2. Provost, F., Fawcett, T.: Analysis and visualization f classifier performance: Comparison under imprecise class and cost distributions. the Third International Conference on Knowledge Discovery and Data Mining (1997) 34–48
3. Cohen, W.W., Schapire, R.E., Singer, Y.: Learning to order things. Journal of Artificial Intelligence Research (10) (1999) 243–270
4. Swets, J.: Measuring the accuracy of diagnostic systems. Science (240) (1988) 1285–1293
5. Drummond, C., Holte, R.C.: Explicitly representing expected cost: An alternative to roc representation. the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2000) 198–207
6. Drummond, C., Holte, R.C.: What roc curves can't do (and cost curves can). ECAI'2004 Workshop on ROC Analysis in AI (2004)
7. Macskassy, S.A., Provost, F., Rosset, S.: Roc confidence bands: An empirical evaluation. in Proceedings of the 22nd International Conference on Machine Learning (ICML 2005) (2005) 537 – 544
8. Macskassy, S.A., Provost, F.: Confidence bands for roc curves: Methods and empirical study. in Proceedings of the 1st Workshop on ROC Analasis in AI (ROCAI-2004) at ECAI-2004 (2004)
9. Drummond, C., Holte, R.C.: Severe class imbalance: Why better algorithms aren't the answer. Proceedings of the 16th European Conference of Machine Learning (2005) 539–546
10. Motulsky, H.: Intuitive Biostatistics. Oxford University Press, New York (1995)
11. Tango, T.: Equivalence test and confidence interval for the difference in proportions for the paired-sample design. Statistics in Medicine **17** (1998) 891–908
12. Newcombe, R.G.: Improved confidence intervals for the difference between binomial proportions based on paired data. Statistics in Medicine **17** (1998) 2635–2650

13. Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J.: UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html (1998) University of California, Irvine, Dept. of Information and Computer Sciences.
14. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann (2005)
15. Dietterich, T.G.: Approximate statistical test for comparing supervised classification learning algorithms. Neural Computation **10**(7) (1998) 1895–1923
16. Newcombe, R.G.: Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in Medicine **17** (1998) 857–872
17. Everitt, B.S.: The analysis of contingency tables. Chapman-Hall (1992)

Appendix A: Tango's Confidence Intervals

Clinical trials, case-control studies, and sensitivity comparisons of two laboratory tests are examples of medical studies that deal with the difference of two proportions in a paired design. Tango's test [11] builds a model to derive a one-sided test for equivalence of two proportions. Medical equivalence is defined as no more than 100Δ percent inferior, where $\Delta(> 0)$ is a pre-specified acceptable difference. Tango's test also derives a score-based confidence interval for the difference of binomial proportions in paired data. Statisticians have long been concerned with the limitations of hypothesis testing used to summarize data [16]. Medical statisticians prefer the use of confidence intervals rather than p-values to present results. Confidence intervals have the advantage of being close to the data and on the same scale of measurement, whereas p-values are a probabilistic abstraction. Confidence intervals are usually interpreted as margin of errors because they provide magnitude and precision. A method deriving confidence intervals must be a priori reasonable (justified derivation and coverage probability) with respect to the data [16].

The McNemar test is introduced in [17] and has been used to rank the performance of classifiers in [15]. Although inconclusive, the study showed that the McNemar test has low Type I error with high power (the ability to detect algorithm differences when they do exist). For algorithms that can be executed only once, the McNemar test is the only test that produced an acceptable Type I error [15]. Despite Tango's test being an equivalence test, setting the minimum acceptable difference Δ to zero produces an identical test to the McNemar test with strong power and coverage probability [11]. In this work, we use Tango's test to compute confidence intervals on the difference in misclassifications in both classes with a minimum acceptable difference $\Delta = 0$ at the (1-α) confidence level. Tango makes few assumptions; (1) the data points are representative of the class. (2) The predictions are reasonably correlated with class labels. This means that the misclassified positives and negatives are relatively smaller than the correctly classified positives and negatives respectively. In other words, the classifier does reasonable well on both classes, rather than performing a random classification. We consider classifier predictions and class labels as paired machines that fit the matched paired design. As shown in table 1 on page 127, entries a and d are the informative or the discordant pairs indicating the agreement portion $(q_{11} + q_{22})$, while b and c are the uninformative or concordant pairs representing the proportion of disagreement $(q_{12} + q_{21})$ [12]. The magnitude of the difference δ in misclassifications can be measured by testing the null hypothesis $H_0 : \delta = q_{12} - q_{21} = 0$. This magnitude is conditional on the observed split of b and c [12]. The null hypothesis H_0 is tested against the alternative $H_1 : \delta \neq 0$. Tango's test derives a simple asymptotic (1-α)-confidence interval for the difference δ and is shown to have good power and coverage probability. Tango's confidence intervals can be computed by:

$$\frac{b - c - n\delta}{\sqrt{n(2\hat{q_{21}} + \delta(1 - \delta))}} = \pm Z_{\frac{\alpha}{2}} \tag{1}$$

where $Z_{\frac{\alpha}{2}}$ denotes the upper $\frac{\alpha}{2}$-quantile of the normal distribution. In addition, $\hat{q_{21}}$ can be estimated by the maximum likelihood estimator for q_{21}:

$$\hat{q_{21}} = \frac{\sqrt{W^2 - 8n(-c\delta(1 - \delta))} - W}{4n} \tag{2}$$

where $W = -b - c + (2n - b + c)\delta$. Statistical hypothesis testing begins with a null hypothesis and searches for sufficient evidence to reject that null hypothesis. In this case, the null hypothesis states that there is no difference, or $\delta = 0$. By definition, a confidence interval includes plausible values for the null hypothesis. Therefore, if the zero is not included in the computed interval, then the null hypothesis $\delta = 0$ is rejected. On the other hand, if the zero value is included in the interval, then we do not have sufficient evidence to reject the difference being zero, and the conclusion is that the difference can be of any value within the confidence interval at the specified level of confidence (1-α).

Tango's test of equivalence can reach its limits in two cases; (1) when the values of b and c are both equal to zero where the Z statistic does not produce a value. This case occurs when we build a perfect classifier and is consistent with the test not using the number of correctly classified examples a and d. (2) The values b and c differ greatly. This is consistent with the assumption that the classifier is somewhat reasonably good, i.e. the classifier is capable of detecting a reasonable portion of the correct classifications in the domain. In both cases of limitations, the confidence intervals are still produced and are reliable [11] but may be wider in range. Tango's confidence intervals are shown not to collapse nor they exceed the boundaries of the normalized difference of $[-1, 1]$ even for small values of b and c.