A Framework for Comparative Evaluation of Classifiers in the Presence of Class Imbalance

William Elazmeh[†] WELAZMEH@SITE.UOTTAWA.CA
Nathalie Japkowicz[†] NAT@SITE.UOTTAWA.CA
Stan Matwin[†‡] STAN@SITE.UOTTAWA.CA

[†] School of Information Technology and Engineering, University of Ottawa
[‡] 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, ROC bands become unreliable. We propose a generic framework for classifier evaluation to identify the confident segment of an ROC curve. Confidence is measured by Tango's 95%-confidence interval for the difference in classification errors in both classes. We test our method with severe class imbalance in a two-class problem. Our evaluation favors classifiers with low numbers of classification errors in both classes. We show that our evaluation method is more confident than ROC bands when faced with severe class imbalance.

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 (Ling et al., 2003; Provost & Fawcett, 1997). Furthermore, the usefulness of advanced evaluation measures, like ROC curves (Cohen et al., 1999; Provost & Fawcett, 1997; Swets, 1988) and cost curves (Drummond & Holte, 2000; Drummond & Holte, 2004), 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 (Macskassy et al., 2005; Macskassy & Provost, 2004) 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, sampling methods become unreliable, especially when the data distribution is unknown (Macskassy & Provost, 2004). In fact, with severe imbalance, the entire issue of evaluation becomes a serious challenge even when making assumptions of data distributions (Drummond & Holte, 2005). In contrast, biostatistical and medical domains impose strong emphasis on error estimates, interpretability of prediction schemes, scientific significance, and confidence (Motulsky, 1995) 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 skeptical of their evaluation methods and continue to develop customized statistical tests to measure characteristics of interest. Our work adopts Tango's test (Tango, 1998) 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. Computing the confidence based on the positive or negative rates (using a or d of the confusion matrix in

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

Appearing in Proceedings of the third Workshop on ROC Analysis in Machine Learning, Pittsburgh, USA, 2006.

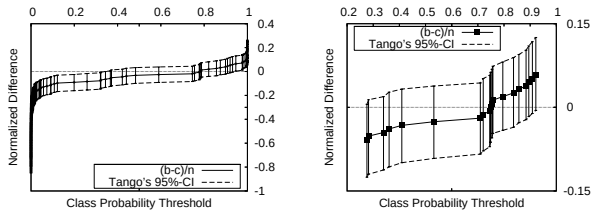

Figure 1. $\frac{b-c}{n}$ and Tango's 95%-confidence intervals for ROC points. Left: all the ROC points. Right: only confident ROC points whose Tango's intervals contain 0.

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, we favor classifiers with similar normalized number of errors in both classes, rather than similar error rates to avoid the imbalance.

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 confident ROC segment 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 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 classification errors and the Area Under the Confident ROC Segment. We present experimental results that show the effectiveness of our approach in severe imbalanced situations compared to ROC bands, ROC curves, and AUC. Having motivated this work, subsequent sections present discussions of classification error proportions in both classes (in section 2), our evaluation framework (in section 3), and our experimental results (in section 4) followed by conclusions and future work (in section 5). We review Tango's statistical test of confidence in appendix A (section 6).

2. 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 tradeoff between accuracy and error rates for a particular class. For the confusion matrix presented in table 1 on page 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 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, 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 classification errors 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 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 confidence test is presented in (Tango, 1998) and is reviewed in appendix A (section 6).

$$1.\ ROC = \left\{ \begin{array}{|c|c|} a_i & b_i \\ \hline c_i & d_i \end{array} \;\middle|\; \begin{array}{l} t_i \in T, i = 1, \cdots, |T|, \\ \begin{array}{|c|c|} a_i & b_i \\ \hline c_i & d_i \end{array} = K(D, t_i), \\ 0 \leq T_i \leq 1 \end{array} \right\}$$

$$2.\ S = \left\{ \begin{array}{|c|c|} a_i & b_i \\ \hline c_i & d_i \end{array} \;\middle|\; \begin{array}{l} (u_i, l_i) = Tango_\alpha(b_i, c_i, n), \\ \begin{array}{|c|c|} a_i & b_i \\ \hline c_i & d_i \end{array} \in ROC, \\ 0 \in [u_i, l_i], \\ i = 1, \cdots, |ROC|, n = |D| \end{array} \right\}$$

$$3.\ CAUC = \begin{cases} 0 & \text{if } S = \text{empty} \\ AUC(S) & \text{if } S \neq \text{empty} \end{cases}$$

$$4.\ AveD = \frac{1}{m} \sum_{i=1}^{m} \frac{b_i - c_i}{n} \quad \forall \begin{array}{|c|c|} a_i & b_i \\ \hline c_i & d_i \end{array} \in S, \; m = |S|$$

Figure 2. Evaluating classifier K (on data D with T class probability thresholds) by Tango at confidence level (1-α). S contains confident ROC points, $CAUC$ is the area under S, and $AveD$ is the average error difference.

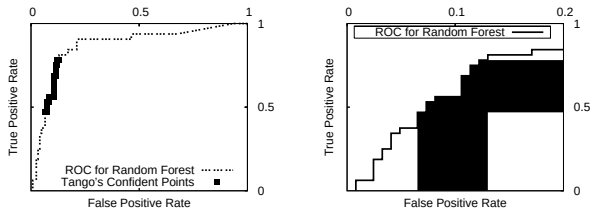

Figure 3. Sample Confident ROC segment (left). Area under ROC segment (right).

3. The Proposed Method of Evaluation

Presented in figure 2, our evaluation 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 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 confident points S. Points in S form the confident ROC segment illustrated in the left plot of figure 3. Our framework is generic and accommodates a test of choice provided that it produces a meaningful interpretation of results. **(3)** Compute $CAUC$ the area under the confident ROC segment S, shown in the right plot of figure 3. **(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 confident ROC segment $CAUC$ against the average observed classification difference $AveD$. Lower values for $AveD$ suggests low classification difference and higher values for $CAUC$ indicate larger confident ROC segment. An effective classifier shows low $AveD$ and high $CAUC$.

Table 2. UCI data sets (Newman et al., 1998)

Data Set	Training	Testing
`dis`	45(+)/(-)2755	13(+)/(-)959
`hypothyroid`	151(+)/(-)3012	–
`sick`	171(+)/(-)2755	13(+)/(-)959
`sick-euthyroid`	293(+)/(-)2870	–
`SPECT`	40(+)/(-)40	15(+)/(-)172
`SPECTF`	40(+)/(-)40	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 (Newman et al., 1998) 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` (Witten & Frank, 2005), 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 particularly 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 4 shows the ROC confidence bands for our four classifiers on the most imbalanced `dis` data set. These bands are gen-

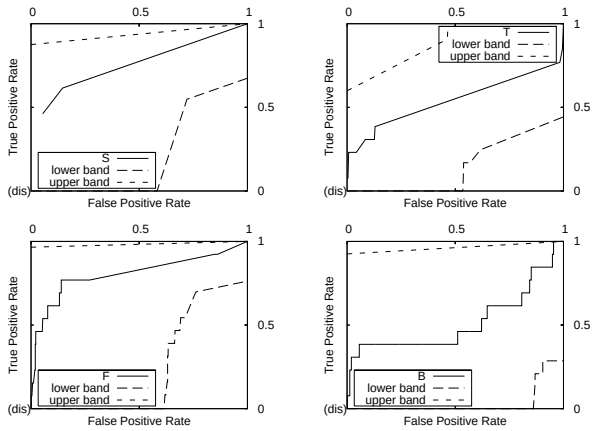

Figure 4. ROC confidence bands for decision stump (S), decision tree (T), random forest (F), and naive Bayes (B) on (`dis`) data set. The bands are wide and are not very useful.

Table 3. AUC values for decision stump (S), decision tree (T), random forest (F), and naive Bayes (B) on data sets.

Data Set	(S)	(T)	(F)	(B)
dis	0.752	0.541	**0.805**	0.516
hypothyroid	0.949	0.936	**0.978**	0.972
sick	0.952	0.956	**0.997**	0.946
sick-euthyroid	0.931	0.930	**0.978**	0.922
spect	0.730	0.745	0.833	**0.835**
spectf	0.674	0.690	**0.893**	0.858

erated using the empirical fixed-width method (Macskassy & Provost, 2004) at the 95% level of confidence (like Tango's test, this method of generating ROC bands does not make assumptions of the underlying distributions of the data). We claim that with severe imbalance, 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.

Next, we consider the ROC curves of our four classifiers on all data sets shown in figure 5. 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

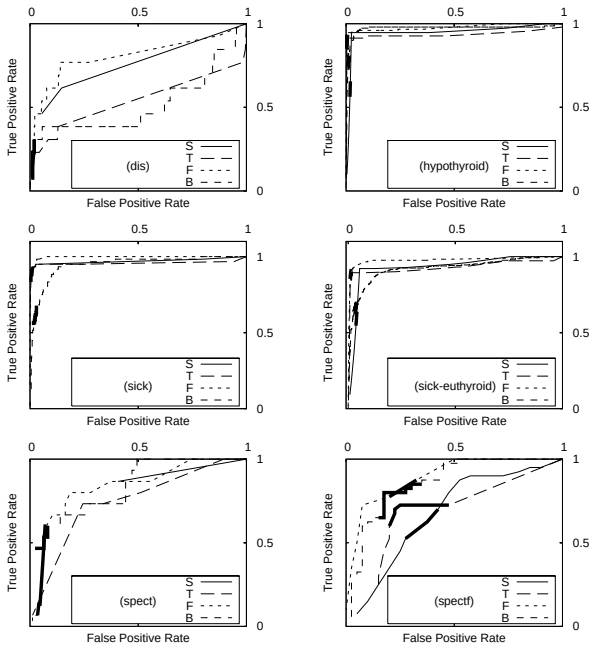

Figure 5. ROC curves for decision stump (S), decision tree (T), random forest (F), and naive Bayes (B) on all data set. The dark segments are Tango's confident points.

higher AUC values than (T). Furthermore, the AUC of (F) is, clearly, higher than that of the others on most the 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 6. Each plot in the figure reports our evaluation of the four classifiers on each data set. The x-axis represents the average normalized classification difference $\frac{b-c}{n}$ for those confident points on the ROC. The y-axis represents the area under the confident segment of the ROC. This area includes the TP area (vertical area) and the FP area (the horizontal area) as illustrated in figure 3 on page 3. Classifiers placed towards the top-left corner perform better (bigger area under the confident ROC segment and less difference in classification error) than those placed closer to the bottom right corner (smaller confident area and higher difference in classification error). 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 left column of figure 6. This is consistent with our expectation of it being less effective. In fact, plots in the right column of the same figure show that (S) also performs poorly producing higher classi-

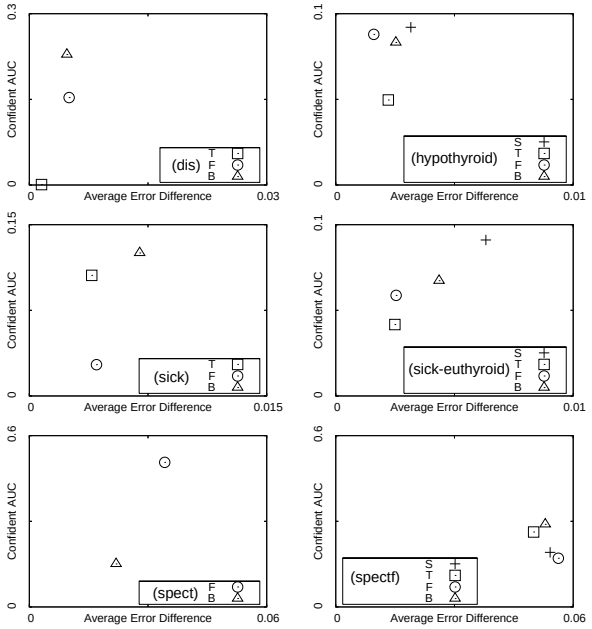

Figure 6. Our evaluation for decision stump (S), decision tree (T), random forest (F), and naive Bayes (B) on our data sets. The y-axis shows the area under the confident ROC segment and the x-axis shows the average observed classification difference $\frac{b-c}{n}$.

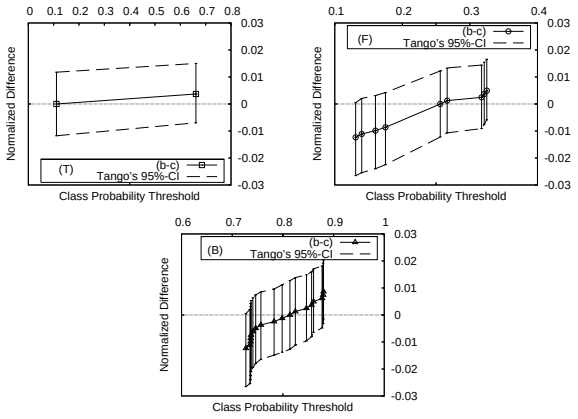

Figure 7. Tango's 95%-confidence intervals for ROC points of decision tree (T), random forest (F), and naive Bayes (B) on (`dis`) set. The center points are ($\frac{b-c}{n}$). (T) has a wide range of thresholds (x-axis).

fication difference and/or covering smaller area under its confident ROC segment. In fact, even when (S) has higher confident AUC than (T), in the right plots of figure 6, (S) still shows a significantly higher difference in classification error than that of (T). The decision tree (T), on the other hand, performs well in most cases and outperforms all other classifiers in the bottom right plot in figure 6. (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 (bottom left 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 confident AUC than (F) on all data sets with the exception of the `spect` data set in the bottom left plot of figure 6. Moreover, (B) shows significantly better performance particularly on the (`dis`) data set.

Our results, clearly, contradict conclusions based on the ROC and AUC evaluations. Therefore, we investigate those confident points along the ROCs for two situations. First, when the four classifiers are trained and tested on the same imbalanced `dis` data set using cross-validation. 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 set), the ROC curves reveal that three of the classifiers produce confident classification points in the bottom left section of the ROC space (see the bold segments in the top left plot of figure 5). These confident points are detected by our method at the 95% level of confidence and are consistent with having severely imbalanced data sets. When we consider the corresponding Tangos 95%-confidence intervals for these classifier (see figure 7), we see that confident points produced by (T) cover a wider range of probability threshold (0.1 to 0.65 on the x-axis of the top left plot) with a low classification difference (y-axis). This indicates added confidence in (T)'s performance. (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 classification difference for a much narrower range of thresholds values. At the least, this indicates a distinction between these classifiers.

For the second situation (`SPECTF` data set), consider the ROC curves in the bottom right plot of figure 5. (T) and (B), clearly, outperform (S) and (F) on this data set. Tango's 95%-confidence intervals of the confident ROC points (shown in figure 8) show that (T) and (B) outperform the other classifiers. When trained on the balanced `spectf` data set, (T) shows the least difference in classification error and has a significantly wider range of threshold values in which it produces

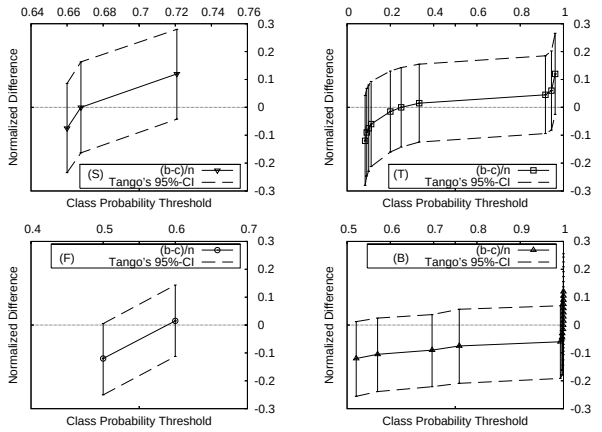

Figure 8. Tango's 95%-confidence intervals for ROC points of decision stump (S), decision tree (T), random forest (F), and naive Bayes (B) on (`spectf`) set. The center points are $(\frac{b-c}{n})$. (T) and (B) have a wider range of thresholds (x-axis) and produce more confident points.

many confident points (0 to 1 along the x-axis of the top right plot in figure 8). Also in this figure, (S) and (T) produce classification points that have exactly zero classification difference while the other two come close to the zero classification 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 classification. Our framework incorporates this evaluation test into the space of the ROC curves to produce confidence oriented evaluation. Our method results in the presentation of the trade-off between classification difference and area under the confident segment of the ROC curve. Our experiments show that our method is more reliable than general ROC and AUC measures.

In the future, we plan to compare our evaluation results to other methods of generating ROC bands to show further usefulness of our framework. Also, 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 classification 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. Furthermore, Tango's test is a clinical equivalence test. This may possibly provide the basis to derive a notion of equivalence on classification.

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).

References

Cohen, W. W., Schapire, R. E., & Singer, Y. (1999). Learning to order things. *Journal of Artificial Intelligence Research*, 243–270.

Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, *10*, 1895–1923.

Drummond, C., & Holte, R. C. (2000). Explicitly representing expected cost: An alternative to roc representation. *the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 198–207.

Drummond, C., & Holte, R. C. (2004). What roc curves can't do (and cost curves can). *ECAI'2004 Workshop on ROC Analysis in AI.*

Drummond, C., & Holte, R. C. (2005). Severe class imbalance: Why better algorithms aren't the answer. *Proceedings of the 16th European Conference of Machine Learning*, 539–546.

Everitt, B. S. (1992). *The analysis of contingency tables.* Chapman-Hall.

Ling, C. X., Huang, J., & Zang, H. (2003). Auc: a better measure than accuracy in comparing learning algorithms. *Canadian Conference on AI*, 329–341.

Macskassy, S. A., & Provost, F. (2004). 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.*

Macskassy, S. A., Provost, F., & Rosset, S. (2005). Roc confidence bands: An empirical evaluation. *in Proceedings of the 22nd International Conference on Machine Learning (ICML 2005)*, 537 – 544.

Motulsky, H. (1995). *Intuitive biostatistics*. Oxford University Press, New York.

Newcombe, R. G. (1998a). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, *17*, 2635–2650.

Newcombe, R. G. (1998b). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, *17*, 857–872.

Newman, D. J., Hettich, S., Blake, C. L., & Merz, C. J. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html. University of California, Irvine, Dept. of Information and Computer Sciences.

Provost, F., & Fawcett, T. (1997). Analysis and visualization f classifier performance: Comparison under imprecise class and cost distributions. *the Third International Conference on Knowledge Discovery and Data Mining*, 34–48.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, 1285–1293.

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine*, *17*, 891–908.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. 2nd Edition, Morgan Kaufmann.

6. 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 (Tango, 1998) 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 (Newcombe, 1998b). 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 (Newcombe, 1998b).

The McNemar test is introduced in (Everitt, 1992) and has been used to rank the performance of classifiers in (Dietterich, 1998). 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 (Dietterich, 1998). 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 (Tango, 1998). In this work, we use Tango's test to compute confidence intervals on the difference in classification errors 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 1, 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 ($q_{12} + q_{21}$) (Newcombe, 1998a). The magnitude of the difference δ in classifications errors 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 (Newcombe, 1998a). 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 (Tango, 1998) 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.