Confidence Interval for the Difference in Classification Error

William Elazmeh and **Nathalie Japkowicz** and **Stan Matwin** [*]
School of Information Technology and Engineering
University of Ottawa, K1N 6N5 Canada
{welazmeh,nat,stan}@site.uottawa.ca

Abstract

Evaluating classifiers with increased confidence can significantly impact the success of many machine learning applications. However, traditional machine learning evaluation measures fail to provide any levels of confidence in their results. In this paper, we motivate the need for confidence in classifier evaluation at a level suitable for medical studies. We draw a parallel between case-control medical studies and classification in machine learning. We propose the use of Tango's biostatistical test to compute consistent confidence intervals on the difference in classification errors on both classes. Our experiments compare Tango's confidence intervals to accuracy, recall, precision, and the F measure. Our results show that Tango's test provides a statistically sound notion of confidence and is more consistent and reliable than the above measures.

Table 1: A standard confusion matrix.

	Predicted +	Predicted -	total
Label +	a	b	a+b
Label -	c	d	c+d
total	a+c	b+d	n

Table 2: Standard evaluation metrics used in ML.

Accuracy	$\frac{a+d}{n}$
True Positive Rate (Recall)	$\frac{a}{a+b}$
False Positive Rate	$\frac{c}{c+d}$
Precision	$\frac{a}{a+c}$
F score	$\frac{2 \times Recall \times Precision}{Recall + Precision}$

Introduction

In machine learning, a classifier is trained on data examples and is used to predict class labels for unseen examples. The learning is supervised if examples are mapped to the positive or negative negative class. Comparing predictions made by the classifier to class labels of examples produces the confusion matrix shown in table 1. Classifier evaluation involves applying one or more evaluation metrics to the confusion matrix. Commonly used metrics, shown in table 2, produce a single scalar value. Other measures, such as ROC or Cost Curves require further analysis of performance. In general, these measures are perceived, by the machine learning community, to be appropriate and sufficient. Ideally, a suitable evaluation measure is sensitive to changes in particular properties of interest in the problem domain. In more specific communities, such as biostatistics and medicine, researchers use specially designed tests that measure particular properties with confidence, and significance (Motulsky 1995). Evaluating classifier performance with increased confidence significantly impacts its success or failure in many applications. An effective application of learning algorithms in laboratory tests or in clinical trials can reduce financial overhead, human involvement, turn-around time, and can increase productivity. However, traditional machine learning metrics fail to provide confidence in their results at the same level as those used in biostatistics or in medicine. Consequently, the usefulness of some learning algorithms becomes inadequately documented and unconvincingly demonstrated.

In this paper, we show that evaluating classifier performance can be exposed to anomalies that may not be captured by a single metric. Although this situation may be corrected by using a combination of metrics, the choice of such metrics requires expertise in evaluation and does not guarantee confidence in the results. Alternatively, we recommend the use of a biostatistical test, the Tango test, to provide a notion of confidence in the evaluation results. Our experimental results will show that Tango's test is capable of capturing such anomalies and remains consistent in distinguishing between classifiers based on their performance while providing confidence and significance. In the long term, our research aims at developing a novel and reliable evaluation measures capable of evaluating classifiers with improved quality and confidence. Such quality has long been utilized in medical domains. Our approach is to adopt these methods into machine learning. This paper presents an initial step towards a novel notion of confident evaluation. We address the need for confidence in classifier performance by apply-

[*] The author is also affiliated with the Institute of Computer Science, Polish Academy of Sciences, Poland.

ing a confidence test. Experimentally, we show that Tango's test (Tango 1998) is capable, reliable and consistent in comparison to accuracy, recall, precision, and F scores. In future work we plan to compare our evaluation method to more sophisticated evaluation measure, such as ROC curves, AUC, and Cost curves. After the introduction, we review work related to classifier evaluation and draw a parallel between case-control studies and classification. In subsequent sections, we present the experimental results and a discussion followed by conclusions and future work.

Related Work

Classifier performance is measured by estimating the accuracy of the learned hypothesis. Estimating hypothesis accuracy with the presence of plentiful data is simple, however, evaluating a hypothesis for future data with the presence of limited data is faced with two challenges defined as *Bias* and *Variance* (Mitchell 1997). The objective is to estimate the accuracy with which a classifier will classify future instances while computing the probable error of this accuracy estimate. In theory, Mitchell (Mitchell 1997) presents methods to evaluate a learned hypothesis, to compare the accuracy of two hypothesis, and to compare the accuracy of two learning algorithms in the presence of limited data. The model is based on basic principles from statistics and sampling theory. In practice, the predictive ability of a classifier is measured by its predictive accuracy (or the error rate) computed on the testing examples (Ling, Huang, & Zang 2003) which, in many cases, has been shown to be insufficient (Ling, Huang, & Zang 2003) or inappropriate (Provost & Fawcett 1997). Performance metrics, in table 2, are commonly used to produce a ranking of classifiers. Alternatively, the ROC (Receiver Operating Characteristics) analysis (Cohen, Schapire, & Singer 1999; Provost & Fawcett 1997; Swets 1988) are used to visualize relative classifier performance to determine the "best" classifier based on comparing the rate of correctly classified positive examples to the rate of incorrectly classified positive examples for a particular class. The AUC (Area under the ROC Curve) (Ling, Huang, & Zang 2003; Caruana & Niculescu-Mizil 2004) produces a single scalar measure to rank classifiers based on how they dominate each other. The ROC curves are shown to be insensitive to the cost of classification (the penalty of miss-classification), therefore, Cost Curves were proposed in (Drummond & Holte 2000; 2004) to introduce costs as a factor in comparing the performance of classifiers.

The above measures fail to answer the question, addressed in (Goutte & Gaussier 2005): given a classifier and its results on a particular collection of data, how confident are we on the computed precision, recall, or F-score? The work in (Goutte & Gaussier 2005) presents a probabilistic interpretation of precision, recall, and F-score to compare performance scores produced by two information retrieval systems. In this case, such work remains probabilistic and makes several assumptions of probability distributions. (Yeh 2000) reports; when comparing differences in values of metrics like Recall, Precision, or balanced F-score, many commonly used statistical significance tests underestimate the differences of their results. (Drummond & Holte 2004) derived confidence intervals to show statistical significance on cost curves by means of sampling and re-sampling. The method is data driven without making any parametric assumptions, e.g. probability distribution assumptions. The sampling method in (Margineantu & Dietterich 2000) use bootstrap methods (Efron & Tibshirani 1993) to generate confidence intervals for classification cost values. These methods may produce confidence level for the evaluation, however, computing confidence intervals that are narrow, reliable and robust is not a simple task (Motulsky 1995). (Newcombe 1998a) shows that the underlying statistical method of Tango's produces more reliable and consistent confidence intervals with good coverage probability. Thus, we expect Tango to out perform any bootstrapping methods (Newcombe 1998a). Comparing these with Tango remains a future work item. In this work, we use the Tango's test to detect if a classifier that produces confident results or not, rather than compute confidence intervals on the results.

Table 3: Marijuana users/non-users and matched controls with (+) or without (-) sleeping difficulties.

	user (+)	user (-)	total
non-user (+)	4	9	13
non-user (-)	3	16	19
total	7	25	32

Case-control Studies and Classification

A case-control study is a medical study to measure the relationship between the exposure to a specific risk factor and the development of a particular disease. The approach is to compare the distributions of patient groups who do develop the disease to those who do not. Such studies can be conducted by clinical trials where groups of patients are exposed to the risk factor and their development of the disease is monitored. However, clinical trials are costly, they are difficult to design, and their resulting distributions of disease development are based on exposure. Alternatively, a case-control study examines exposure history of patients who do or do not develop the disease. Supposedly, patients in the control group are similar to those in the cases group with the absence of the diseases, i.e. individual cases of the disease are matched with individual controls based on a set of attributes, e.g. age, gender, location, etc. (Motulsky 1995). The study becomes stronger when the analysis takes into consideration this paired matching. For example, consider the control-case study in (Tango 1998; Karacan, Fernandez, & Coggins 1976; Altman 1991) of 32 marijuana users matched with 32 controls with respect to their sleeping difficulties shown in table 3. The subjects are matched based on a set of relevant variables (age, location, etc.) Each entry in table 3 is a number of pairs (control and case pairs) where 4 matched pairs of marijuana users and controls experienced sleeping difficulties, 9 controls experienced sleeping difficulties while their matched marijuana users did not experience sleeping difficulties, 3 controls did not experience sleeping difficulties while their matching marijuana users did, and 16 pairs of marijuana users and control subjects (non-users) did not experience sleeping difficulties.

In this study, the interest lies in the statistical significance for the difference in proportions experiencing sleeping difficulties. The two diagonal entries in Table 3 (entries a and d in table 1), medically and statistically, provide no information with respect to the difference between sleeping difficulties and exposure to marijuana (Newcombe 1998a). The results of sleeping difficulty in both entries are the same for both users and non-users of marijuana. However, the two off-diagonal entries (exposed controls that are not cases of sleeping difficulties and non-exposed controls that are cases of sleeping difficulties – entries b and c of table 1) are relevant to the association between sleeping difficulties and the exposure to marijuana. (Tango 1998) computes the 1-α confidence intervals for this normalized difference $\frac{b-c}{n}$ in these proportions (n is the total number of cases and controls). For the above example, Tango's produces the 95%-confidence interval of $[-0.02709, 0.38970]$. By definition, the confidence intervals produce the plausible values of the difference in proportions. Since the value 0 is inside the confidence interval, then with 95% confidence it can be stated that the difference $\frac{b-c}{n}$ may have a plausible value of zero and that the two groups (users and non-users of marijuana) exhibit sleeping difficulties with no statistically significant difference. Therefore in the above example, the observed difference is $\frac{9-3}{32} = 18.75\%$ is insignificant. Tango's test is presented in (Tango 1998) and is described in more details in appendix A.

At this point, it is important to draw a parallel to classification in machine learning where a classifier is trained on labeled examples and is tested on unseen labeled examples. Classifier evaluation involves estimating to what extent, if any, the resulting classifier is capable of predicting the class labels of the unseen instances. In the context of the above control-studies, the issue is measuring the difference between distributions of classifier predictions and class labels. On way to assess this ability is to measure the difference in error proportions between instance labels and classifier predictions by using Tango's test. In other words, when taking the case-control approach to classification, testing examples undergo two classifications, labels and predictions. The two classifications may produce identical labels (true positives or true negatives) and/or different labels (false positives and false negatives). The issue of evaluation becomes: at a particular level of confidence (95%), is there a statistically significant difference between the error proportions of classifications. This is parallel to analyzing the error difference of exhibiting sleeping difficulties between users and non-users of marijuana. In this context, this method can evaluate binary classifiers where the positive class is a positive indication of the presence of a particular disease or a particular condition.

Experimental results

This section presents our experimental design with a brief review of our data sets followed by a discussion of our experimental results. The objective is to compare the performance of four classifiers reported by accuracy, recall (true positive rate), false positive rate, precision, and F measures to that reported by Tango. We use Tango's confidence intervals to measure the significance of the difference in error fractions $\frac{b-c}{n}$ between predictions made by the four classifiers and class labels at the 95%-confidence level. Thus, a classifier is considered 95%-confident (✔) when its corresponding Tango's 95%-confidence interval includes the zero value. Otherwise, the observed difference is deemed statistically insignificant. Using the `Weka 3.4.6` software (Witten & Frank 2005) on our datasets, we build four classifiers; a decision stump (S), a decision tree (T), a random forest (F), and a Naive Bayes (B). The rationale is to build classifiers for which we can expect a ranking of their relative performance. A decision stump built without boosting is a decision tree with one test at the root (only 2 leaf nodes) and is expected to perform particularly worse than the decision tree. Relatively, a decision tree is a stronger classifier because its tree is more developed and has more leaf nodes which cover the training examples. In theory, multiple tests on the data should produce a classifier better than performing a single test. The random forest classifier is a reliable classifier and is expected to outperform a single decision tree. Finally, the naive Bayes tends to minimize classification error and is expected to perform well when trained on balanced data.

Table 4: UCI datasets used and their class distributions.

#	Dataset	Attr.	Train +/-	Test +/-
1	WPBC	34	151/47	c. valid.10
2	WBCD	11	458/241	c. valid.10
3	Pima Diabetes	9	500/268	c. valid.10
4	Echocardiogram	13	50/24	c. valid.10
5	Hepatitis	20	32/123	c. valid.10
6	Hypothyroid	26	151/3012	c. valid.10
7	WDBC	31	212/357	c. valid.10
8	Thyroid (euthy.)	26	293/2870	c. valid.10
9	Thyroid (dis)	30	45/2755	13/959
10	Thyroid (sick)	30	171/2629	60/912
11	SPECT	23	40/40	15/172
12	SPECTF	45	40/40	55/214

Our data sets, listed in table 4, are selected from the UCI Machine Learning repository (Newman *et al.* 1998). The data consists of records that describe characteristics of a class in a two class problem (positive and negative). The data sets are grouped into three groups based on their class distributions of positive and negative examples. The first two groups have no testing examples, therefore, we use cross-validation test method of 10 folds. The third group of data sets has test sets that are severely imbalanced. The experimental results are presented in two tables. Table 5 presents the performance of our classifiers on our data sets as evaluated by the Accuracy (Acc.), Recall (Rec.) or true positive rate, Precision (Prec.), and F score. The table also shows which classifiers are found to be confident by Tango (indicated by ✔). Table 6 shows the performance evaluation of the same classifiers on the same data sets as measured by Tango's 95%-confidence intervals. The table also shows the entries to each of the confusion matrices (a, b, c, d as described in table 1) along with the observed $\frac{b-c}{n}$.

Consider the evaluation results in both tables for data set #1. We observe that (S) shows very good performance values in table 5. This is contradicted by Tango in table 6 where (S) shows the widest confidence intervals with a significant observed difference. We examine the corresponding confusion matrix (a, b, c, d entries in table 6) to see that (S) predicts a positive class for all examples ($b = d = 0$). This is a situation that cannot be captured from table 5, unless, one considers the FP rate because it is related to the a poor performance on the negative class. On the same data set, (B) shows a much less FP rate with significantly lower scores in table 5. This is due to its error on both classes (both b and c are higher). Alternatively, Tango, in table 6, indicates that (T) is the confident classifier. When we consider the confusion matrices shown in table 6, we see that Tango's test detects significant increases in classification errors on either one of the two classes (when either b or c is significantly higher). In fact, Tango's test is designed to favor those classifiers that have lower classification errors in both classes. This criteria imposes accuracy constraints that prevents classifiers such as (S) from obtaining high evaluation ranks.

For data sets #5 and #9, consider the confusion matrices (entries a, b, c, d in table 6) of all four classifiers. If we compare their corresponding $\frac{a}{a+b}$ and $\frac{d}{c+d}$, we see that all classifiers perform significantly better on the negative class than on the positive class. In table 5, their accuracies remains high. Their FP rates, recall, precision, and F scores are relatively low. Accuracy, in this case, is inappropriate due to the significantly higher number of negative examples while recall, precision, and F scores are computed for the positive class only. In fact, Tango finds that (B) on set #5 and (T) on set # 9 produce reliable classifiers that have the least observed $\frac{b-c}{n}$ difference. Their corresponding F scores are the highest on each set respectively. Furthermore, accuracy can be inappropriate in other circumstances. For instance, consider data set #9. The performance of classifier (S) is vary poor in table 5, yet its accuracy remains high. When looking at the same row in table 6, we see that (S) classified all instances in the negative class but due to the severe imbalance, the accuracy on the negative class becomes sufficient for (S) to obtain a high accuracy score.

On data sets # 2, 4, 6, 7, 8, and 10, we see that most of the scores in table 5 are relatively high, in particular, the F scores. In addition, the highest F scores of classifiers, in each set in this group, are found to be confident by Tango. If we consider the confusion matrices for these classifiers on these sets in table 6, we see that most of them perform well on both classes and their errors (b and c) are low. We can see that Tango favors those classifiers that produces lower values of error on both classes. Furthermore, Tango's loses confidence in those classifiers that produce a significant error on only one of the two classes. This loss of confidence appears to correspond with lower F scores in table 5. However, the values of those confident F scores appears to vary for different data sets. For example, an F score of 80.2% for (T) on set # 3 is confident but 86.2% F score is not confident for (F) on data set # 10. In addition, on data set # 2, (T) has a confident F score of 95.7% while (B) has 98% F score with no confidence. This is explained by comparing their corresponding $b - c$ values shown in table 6. Tango produces more confidence for classifier with lower $\frac{b-c}{n}$ values.

Conclusions and Future work

A conclusive observation is that Tango shows a consistent confidence with lower values of $\frac{b-c}{n}$ (in the absolute value) which favors lower errors on both classes. The accuracy can possibly be misleading or inappropriate. Recall alone is insufficient and must take into account precision. The FP rate is helpful when combined with some other metric (e.g. recall). However, the question remains: for evaluation, which metric should one consider? The answer is not necessarily obvious and requires an extensive understanding of the metric and its shortcomings. The F score alone is not sufficient (data set #1). In this work, we propose using Tango's test to assess classification errors on both classes. Our experiments show that standard machine learning evaluation metrics, namely, accuracy, recall, false positive rate, precision, and F score are unable to single-handedly provide confidence in their evaluation results. Tango provides a consistent notion of confidence in the form of statistically sound confidence intervals on the difference in classification error. Our subsequent work will extend our comparisons to include ROC curves and cost curves in search for an improved and confident evaluation measure for supervised, and possibly for unsupervised, 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

Altman, D. G. 1991. *Practical Statistics for Medical Research*. Chapman-Hall.

Caruana, R., and Niculescu-Mizil, A. 2004. An empirical evaluation of supervised learning for roc area. *The 1st Workshop on ROC Analysis in AI, ECI 2004* 1–8.

Cohen, W. W.; Schapire, R. E.; and Singer, Y. 1999. Learning to order things. *Journal of Artificial Intelligence Research* (10):243–270.

Dietterich, T. G. 1998. Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computations* 10(7):1895–1923.

Drummond, C., and Holte, R. C. 2000. Explicitly representing expected cost: An alternative to roc representation. *The Sixth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* 198–207.

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

Table 5: Accuracy (Acc.), Recall (Rec.), FP rate, Precision (Prec.), and F-score for C = decision stump (S), decision tree (T), random forest (F), or naive Bayes (B) on data sets # (table 4). Confident classifiers by Tango are ✔.

#	C	Acc.	Rec.	FP Rate	Prec.	F	Tango
1	S	76.3	100	100	76.3	86.5	
	T	75.8	85.4	55.3	83.2	84.3	✔
	F	79.3	96.7	76.6	80.2	87.7	
	B	67.2	71.5	46.8	83.1	76.9	
2	S	88.3	83.0	1.7	99.0	90.3	
	T	94.4	95.4	7.5	96.0	95.7	✔
	F	96.0	96.5	5.0	97.4	96.9	✔
	B	97.4	96.7	1.2	99.3	98.0	
3	S	71.9	79.6	42.5	77.7	78.7	✔
	T	73.8	81.4	40.3	79	80.2	✔
	F	72.5	83.6	48.1	76.4	79.8	
	B	76.3	84.4	38.8	80.2	82.3	✔
4	S	94.6	96.0	8.3	96.0	96.0	✔
	T	96.0	98.0	8.3	96.1	97.0	✔
	F	100	100	0.0	100	100	✔
	B	94.6	98.0	12.5	94.2	96.1	✔
5	S	77.4	6.3	4.1	28.6	10.3	
	T	83.9	43.8	5.7	66.7	52.8	
	F	83.2	37.5	4.9	66.7	48.0	
	B	84.5	68.8	11.4	61.1	64.7	✔
6	S	97.4	94.7	2.5	65.6	77.5	
	T	99.2	91.4	0.4	92.6	92.0	✔
	F	99.0	88.7	0.5	90.5	89.6	✔
	B	97.9	77.5	1.1	78.5	78.0	✔
7	S	88.9	77.8	4.5	91.2	84.0	
	T	93.2	92.5	6.4	89.5	91.0	✔
	F	95.1	93.4	3.9	93.4	93.4	✔
	B	93.0	89.6	5.0	91.3	90.5	✔
8	S	94.4	92.2	5.3	63.8	75.4	
	T	97.9	87.0	1.0	89.8	88.4	✔
	F	97.9	87.7	1.1	89.2	88.5	✔
	B	84.4	89.8	16.2	36.1	51.5	
9	S	98.7	0.0	0.0	0.0	0.0	
	S	98.3	23.1	0.7	30.0	26.1	✔
	F	98.6	7.7	0.2	33.3	12.5	
	B	94.3	30.8	4.8	8.0	12.7	
10	S	96.0	95.0	3.9	61.3	74.5	
	T	98.8	86.7	0.4	92.9	89.7	✔
	F	98.5	78.3	0.2	95.9	86.2	
	B	93.5	73.3	5.2	48.4	58.3	
11	S	61.5	86.7	40.7	15.7	26.5	
	T	75.4	73.3	24.4	20.8	32.4	
	F	76.5	80.0	23.8	22.6	35.3	
	B	74.9	66.7	24.4	19.2	29.9	
12	S	66.9	85.5	37.9	36.7	51.4	
	T	77.3	85.5	24.8	47.0	60.6	
	F	78.8	92.7	24.8	49.0	64.2	
	B	74.0	92.7	30.8	43.6	59.3	

Table 6: Tango's 95%-Confidence Intervals and $\frac{b-c}{n}$ for C = decision stump (S), decision tree (T), random forest (F), or naive Bayes (B) on data sets # (table 4). (a,b,c,d) is a confusion matrix. Confident classifiers by Tango are ✔.

#	C	(a,b,c,d)	Tango's CI	$\frac{b-c}{n}$	
1	S	(151,0,47,0)	[-30.1,-18.4]	-23.7	
	T	(129,22,26,21)	**[-9.0,5.0]**	-2.0	✔
	F	(146,5,36,11)	[-22.0,-9.9]	-15.7	
	B	(108,43,22,25)	[2.7,18.5]	10.6	
2	S	(380,78,4,237)	[8.3,13.2]	10.6	
	T	(437,21,18,223)	**[-1.4, 2.3]**	0.4	✔
	F	(442,16,12,229)	**[-1.0, 2.2]**	0.6	✔
	B	(443,15,3,238)	[0.6, 3.1]	1.7	
3	S	(398,102,114,154)	**[-5.3, 2.2]**	-1.6	✔
	T	(407,93,108,160)	**[-5.6, 1.7]**	-2.0	✔
	F	(418,82,129,139)	[-9.8,-2.4]	-6.1	
	B	(422,78,104,164)	**[-6.9, 0.1]**	-3.4	✔
4	S	(48,2,2,22)	**[-7.1,7.1]**	0.0	✔
	T	(49,1,2,22)	**[-8.2,4.9]**	-1.4	✔
	F	(50,0,0,24)	**[-4.9,4.9]**	0.0	✔
	B	(49,1,3,21)	**[-10.1,3.8]**	-2.7	✔
5	S	(2,30,5,118)	[9.3,23.6]	16.1	
	T	(14,18,7,116)	[0.8,13.8]	7.1	
	F	(12,20,6,117)	[2.8,15.9]	9.0	
	B	(22,10,14,109)	**[-9.2, 3.8]**	-2.6	✔
6	S	(143,8,75,2937)	[-2.7,-1.6]	-2.1	
	T	(138,13,11,3001)	**[-0.3, 0.4]**	0.1	✔
	F	(134,17,14,2998)	**[-0.3, 0.5]**	0.1	✔
	B	(117,34,32,2980)	**[-0.5, 0.6]**	0.1	✔
7	S	(165,47,16,341)	[2.8,8.3]	5.5	
	T	(196,16,23,334)	**[-3.5,1.0]**	-1.2	✔
	F	(198,14,14,343)	**[-1.9,1.9]**	0.0	✔
	B	(190,22,18,339)	**[-1.5,3.0]**	0.7	✔
8	S	(270,23,153,2717)	[-5.0, -3.3]	-4.1	
	T	(255,38,29,2841)	**[-0.2, 0.8]**	0.3	✔
	F	(257,36,31,2839)	**[-0.4, 0.7]**	0.2	✔
	B	(263,30,465,2405)	[-15.1,-12.5]	-13.8	
9	S	(0,13,0,959)	[0.8, 2.3]	1.3	
	S	(3,10,7,952)	**[-0.6, 1.3]**	0.3	✔
	F	(1,12,2,957)	[0.3, 2.0]	1.0	
	B	(4,9,46,913)	[-5.4,-2.4]	-3.8	
10	S	(57,3,36,876)	[-4.8,-2.3]	-3.4	
	T	(52,8,4,908)	**[-0.3, 1.3]**	0.4	✔
	F	(47,13,2,910)	[0.4, 2.1]	1.1	
	B	(44,16,47,865)	[-4.9,-1.6]	-3.2	
11	S	(13,2,70,102)	[-43.7,-29.3]	-36.4	
	T	(11,4,42,130)	[-27.2,-14.1]	-20.3	
	F	(12,3,41,131)	[-27.1,-14.3]	-20.3	
	B	(10,5,42,130)	[-26.7,-13.4]	-19.8	
12	S	(47,8,81,133)	[-33.3,-21.2]	-27.1	
	T	(47,8,53,161)	[-22.3,-11.6]	-16.7	
	F	(51,4,53,161)	[-23.6,-13.4]	-18.2	
	B	(51,4,66,148)	[-28.8,-17.8]	-23.1	

Efron, B., and Tibshirani, R. 1993. *An Introduction to the Bootstrap*. Chapman-Hall.

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

Goutte, C., and Gaussier, E. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. *The 27th European Conf. on Information Retrieval, ECIR* 345–359.

Karacan, I.; Fernandez, S. A.; and Coggins, W. S. 1976. Sleep electrocephalographic-electrooculographic characteristics of chronic marijuana users: part 1. *The New York Academy of Science* (282):348–374.

Ling, C. X.; Huang, J.; and Zang, H. 2003. Auc: a better measure than accuracy in comparing learning algorithms. *The Canadian Conf. on AI* 329–341.

Margineantu, D. D., and Dietterich, T. G. 2000. Bootstrap methods for the cost-sensitive evaluation of classifiers. *The 17th Int. Conf. on Machine Learning* 582–590.

Mitchell, T. M. 1997. *Machine Learning*. McGraw-Hill.

Motulsky, H. 1995. *Intuitive Biostatistics*. Oxford U. Press.

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.; and Merz, C. J. 1998. UCI machine learning repository. http://www.ics.uci.edu/∼mlearn/MLRepository.html. University of California.

Provost, F., and Fawcett, T. 1997. Analysis and visualization f classifier performance: Comparison under imprecise class and cost distributions. *The 3rd Int. Conf. on Knowledge Discovery and Data Mining* 34–48.

Swets, J. 1988. Measuring the accuracy of diagnostic systems. *The Journal of Science* (240):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., and Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques*. M. Kaufmann.

Yeh, A. 2000. More accurate tests for the statistical significance of result differences. *The 18th conf. on Computational Linguistics* 2:947 – 953.

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 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 must make 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.

Entries a and d (in table 1) 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: $\frac{b-c-n\delta}{\sqrt{n(2\hat{q_{21}}+\delta(1-\delta))}} = \pm Z_{\frac{\alpha}{2}}$ 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}: $\hat{q_{21}} = \frac{\sqrt{W^2-8n(-c\delta(1-\delta))}-W}{4n}$ 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. The null hypothesis states that there is no difference $(\delta = 0)$. By definition, a confidence interval includes plausible values of the null hypothesis. Thus, if zero is not in the interval, then $\delta = 0$ is rejected. If zero is in the interval, then there is no sufficient evidence to reject $\delta = 0$, and the conclusion is that the difference can be of any value within the confidence interval at the specified level of confidence (1-α).