Evaluating Misclassifications in Imbalanced Data

William Elazmeh

Joint work with Nathalie Japkowicz and Stan Matwin

University of Ottawa, Canada

September 14^{th}, 2006

ONTARIO CENTRES of EXCELLENCE

NSERC CRSNG

Who we are and what we do

The TAMALE (Text Analysis and Machine Learning Group), founded by Prof. Stan Matwin in 1988, primary research focuses on knowledge management. Knowledge management is considered here as a research field that combines Data Mining, Text Mining and Language Engineering, and builds on the technologies of Databases, Data Warehousing and Knowledge Bases.

- Stan Matwin
- Diana Inkpen
- Nathalie Japkowicz
- Iluju Kiringa
- Liam Peyton
- Stan Szpakowicz
- Marcel Turcotte
- Herna Viktor
- 1 PostDoc. 12 Ph.D. Students 25 M.Sc. Students.

The task of classification

- data contains examples of observed values for attributes
- each example is mapped to $+$ or $-$ class label
- data is split into training and testing portions
- a classifier is trained on the training examples
- the classifier predicts class label for unseen examples
- sample data can be obtained from the UCI Machine Learning repository [12]

Testing the classifier

- we use examples from the testing portion of data
- for which, the classifier makes Y or N predictions of their class labels
- performance is determined by comparing classifier predictions to class labels
- the comparison produces the confusion matrix

	Y	N
+	T+	F-
-	F+	T-

- performance evaluation applies a performance metric of choice to the above confusion matrix

Commonly used (simple) metrics

	Y	N
+	T+	F-
-	F+	T-

$$\text{F+ Rate} = \frac{F+}{-}$$

$$\text{T+ Rate (Recall)} = \frac{T+}{+}$$

$$\text{Precision} = \frac{T+}{Y}$$

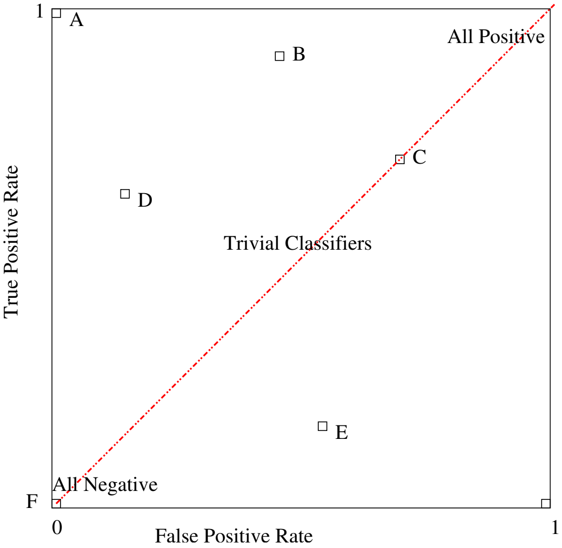
$$\text{Accuracy} = \frac{(T+)+(T-)}{(+)+(-)}$$

$$\text{F-Score} = \text{Precision} \times \text{Recall}$$

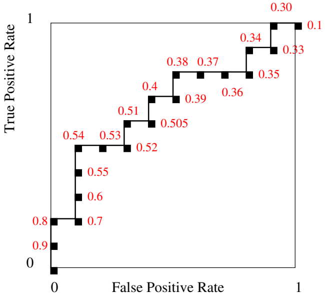
(Not so simple) metrics being used increasingly!

- Receiver Operating Characteristic (ROC) curves [1, 13, 14]
- ROC confidence bands [8, 9]
- Cost curves (slopes of the ROC curve)[3, 4]
- Evaluation is a hard problem [5]
  - parametric methods (assume data distributions)
  - non-parametric methods (empirical, rely on sampling)

Receiver Operating Characteristics (ROC) Space

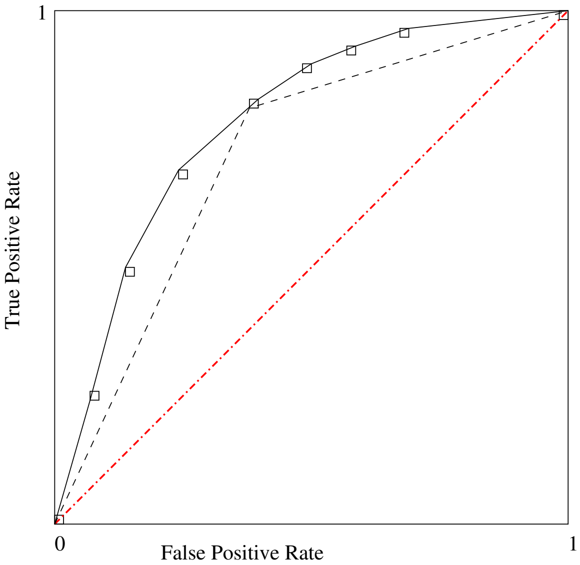


Generating ROC curves

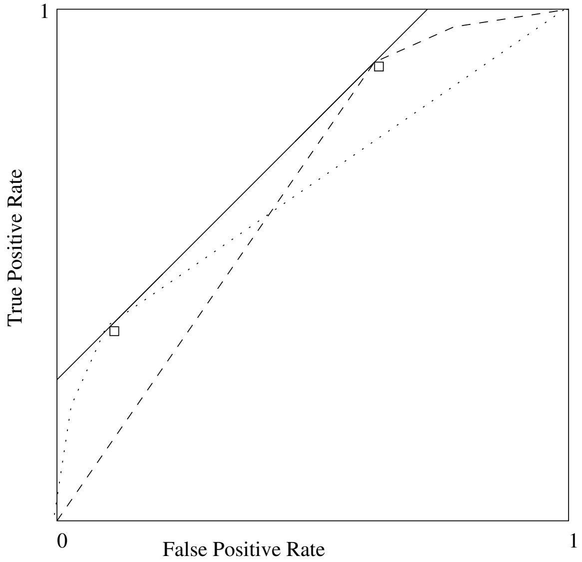


#	Class	Score	#	Class	Score
1	+	0.9	11	+	0.4
2	+	0.8	12	-	0.39
3	-	0.7	13	+	0.38
4	+	0.6	14	-	0.37
5	+	0.55	15	-	0.36
6	+	0.54	16	-	0.35
7	-	0.53	17	+	0.34
8	-	0.52	18	-	0.33
9	+	0.51	19	+	0.30
10	-	0.505	20	-	0.1

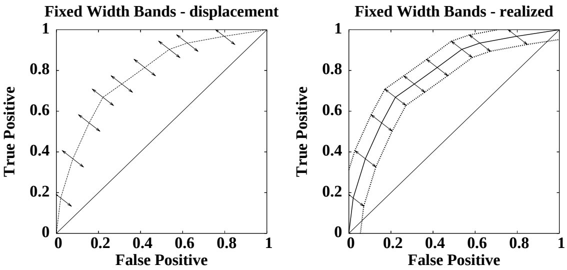
Comparing classifiers' ROC curves



Choosing classifiers in ROC space



Generating ROC confidence bands (FWB) [8]



Challenges in evaluating classifier performance

- variations in data sampling from the domain [2]
- variations in how data represents the concept
- variations in the learning algorithm (bias) [2]
- random classification error (by chance alone)
- domain variability and experimental imprecision (should not affect evaluation)
- sensitivity and limitations of metrics being used, particularly when:
  - data is limited (small in size)
  - classes are severely imbalanced (ratio of $+$ to $-$)
- assumptions may limit our choice of metrics

What's involved in classifier evaluation?

We should:

- understand the domain and the attributes
- decide what "interesting" properties to measure
- choose suitable evaluation methods and metrics
- check preconditions and post-conditions of the above measure and, optionally, select an alternative evaluation method as a benchmark for comparison
- select a classifier "best" suited for the domain
- apply the evaluation method(s) and analyze the results
- develop confidence in our results, i.e. "believe" them!

Do we?

Motivations

- measuring the quality of learning is necessary for the development and deployment of machine learning algorithms
- current performance measures of such algorithms remain primitive with respect to interpretation, significance and confidence
- thus, the usefulness of these algorithms is inadequately documented and unconvincingly demonstrated
- consequently, real-life practitioners abstain from using machine learning methods due to their short comings in real-life applications

This means....

	Y	N
+	T+	F-
-	F+	T-

- Accuracy is insufficient or inappropriate [7, 13]
- most metrics struggle with severe imbalance
- because they use T+ or T- in their calculations
- and they fail to provide confidence in their results

Our intuitions

- recent advances and development in machine learning have reached a mature stage to facilitate more robust evaluation and testing paradigms
- the robust evaluation will encourage practitioners to reconsider Machine Learning algorithms
- the purpose of our work is to survey current statistical methods, then, extract those of interest for machine learning and adapt them to our actual problems
- like biologists, economists, psychologists, etc. who adapted statistical methods to their particular needs (Statistics for Biologists [10], Statistics for Social Scientists, etc), our aim is to design sound evaluation measures adapted to machine learning algorithms

Our intuitions (continued)

- Biostatisticians continue to develop customized statistical tests to measure characteristics of interest
- Our work adopts Tango's test [15] from biostatistics 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 [11] to be reliable and robust with power and coverage probability to produce confidence and significance

Our work

- computing confidence using $F+$ or $T+$ rates can be influenced by class imbalance
- alternatively, we apply a statistical significance test to those entries that resist such influence
- to counter the class imbalance, particularly when the number of instances in the minority class is very small, we use Tango's test to favor classifiers with similar normalized number of errors in both classes
- consequently, any evaluation measure that uses $F+$ and $F-$ rates (ROC) is influenced by data imbalance, while the error analysis we propose is not
- since we measure only the error of classification, we need to combine Tango's analysis together with another evaluation measure (AUC) to measure how well the classifier performs positively

We propose

- a framework for classifier evaluation that identifies confident points along an ROC curve
- these points form a balanced misclassification segment on the ROC curve
- our 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.
- 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

Case-Control Studies

Table: Sleeping Difficulties in Marijuana Users [6, 15]

		Control Y	Control N	total
Case	Y	4	6	13
	N	3	16	19
total		7	25	32

- The relationship between exposure and disease
- Confidence in the evaluation
- Paired Matching (cases and controls are similar)
- Costly clinical trials
- Small number of data points

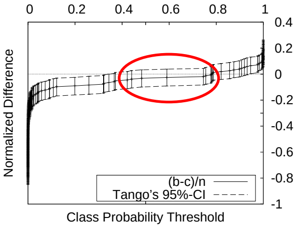
Tango's Statistical test [15]

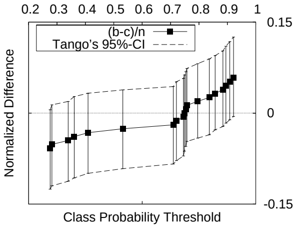
Table: The statistical proportions in a confusion matrix.

	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

- Tango builds $(1-\alpha)$-Confidence Intervals on the difference $\frac{b-c}{n}$
- $H_0 : \delta = q_{12} - q_{21} = 0$ against $H_1 : \delta \neq 0$ ✔
- Tango's CI: $\frac{b-c-n\delta}{\sqrt{n(2\hat{q_{21}}+\delta(1-\delta))}} = \pm Z_{\frac{\alpha}{2}}$ where $Z_{\frac{\alpha}{2}}$ denotes the upper $\frac{\alpha}{2}$-quantile of the normal distribution
- $\hat{q_{21}}$ is estimated by maximum likelihood estimator for q_{21}: $\hat{q_{21}} = \frac{\sqrt{W^2-8n(-c\delta(1-\delta))}-W}{4n}$ where $W = -b - c + (2n - b + c)\delta$.

Misclassification difference



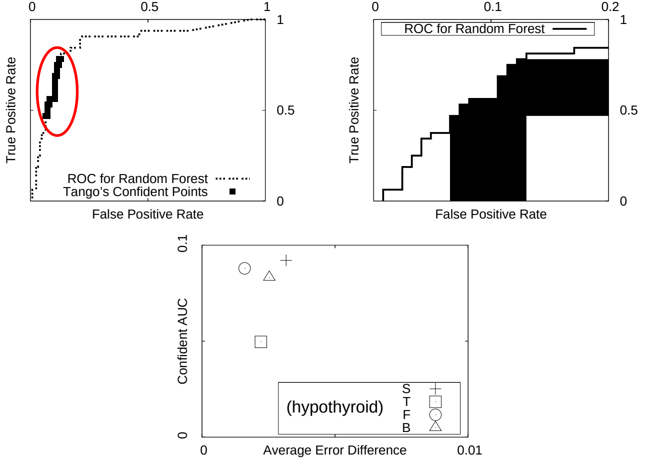


- if $T = 0$, then all are classified positive
- if $T = 1$, then all are classified negative
- if $(T > 0)$ and $(T < 1)$ but increasing, then:
  - c decreases (FP become correctly classified)
  - b increases (TP become incorrectly classified)
  - b and c do not change (correct classification)
- THEN: $\frac{b-c}{n}$ is monotone non-decreasing

The proposed method of evaluation

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.
3. Compute *CAUC* the area under the confident ROC segment S.
4. Compute *AveD* the average normalized difference $(\frac{b-c}{n})$ for all points in S.

The proposed method illustrated



The experiments

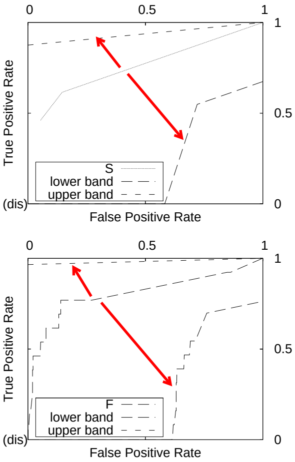
- we have a collection of binary classification data sets from th UCI repository [12]
- using Weka [16], build four classifiers:
  1. a decision stump without boosting (S)
  2. a single decision tree (T)
  3. a random forest (R)
  4. a naive Bayes (B)
- produce the ROC bands to illustrate their struggle
- compare the performance of all four classifiers using:
  1. ROC curves
  2. AUC
  3. our method

The data sets

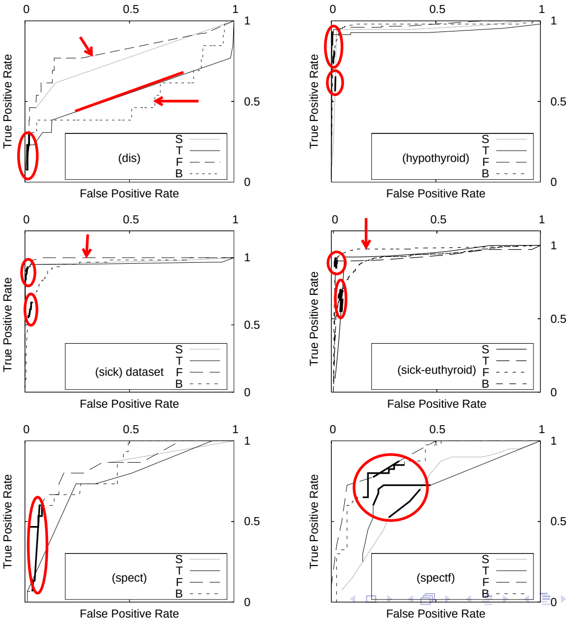
Data Set	Training	Testing
dis	45(+)/(-)2755	13(+)/(-)959
hypothyroid	151(+)/(-)3012	–
sick	171(+)/(-)2755	13(+)/(-)959
sick-euthyroid	293(+)/(-)2870	–
SPECT	40(+)/(-)40	15(+)/(-)172
SPECTF	40(+)/(-)40	55(+)/(-)214

- severe imbalance
- very few + examples
- some have balanced training data
- use cross-validation (10 folds) when there is no test data

ROC Bands for `dis` data set



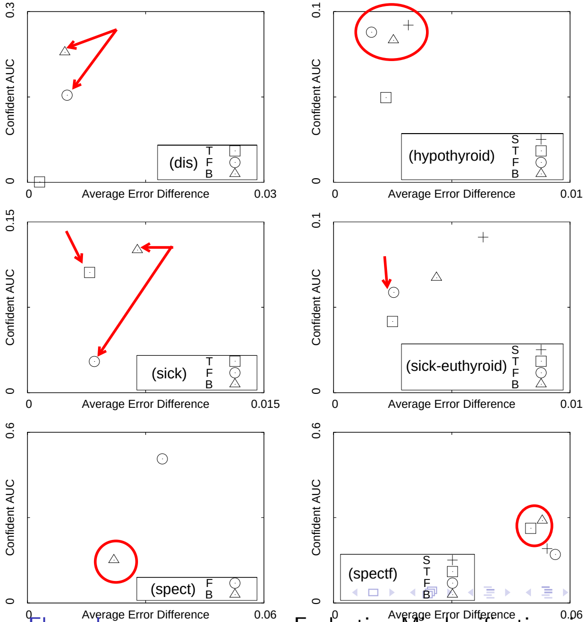
ROC Bands for `dis` data set



AUC of ROC curves

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

Proposed method's results



Conclusions

- ROC curves struggle with imbalance on small data
- AUC not much better
- ROC Bands unreliable
- Tango resists imbalance and handles small data
- Confidence-oriented framework for evaluation
- Focused evaluation on confident ROC segments
- For the future, we aim to derive confidence intervals based on Tango's test
- Apply Tango's test to general classification

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

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

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

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

C. Drummond and Robert. C. Holte.
Severe class imbalance: Why better algorithms aren't the answer.
Proceedings of the 16th European Conference of Machine Learning, pages 539–546, 2005.

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

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

Sofus A. Macskassy and Foster Provost.
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.

Sofus A. Macskassy, Foster Provost, and Saharon Rosset.

Roc confidence bands: An empirical evaluation.
in Proceedings of the 22nd International Conference on Machine Learning (ICML 2005), pages 537 – 544, 2005.

Harvey Motulsky.
Intuitive Biostatistics.
Oxford University Press, New York, 1995.

Robert G. Newcombe.
Improved confidence intervals for the difference between binomial proportions based on paired data.
Statistics in Medicine, 17:2635–2650, 1998.

D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz.
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.

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

J. Swets.
Measuring the accuracy of diagnostic systems.
Science, (240):1285–1293, 1988.

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

Ian H. Witten and Eibe Frank.
Data Mining: Practical machine learning tools and techniques.
2nd Edition, Morgan Kaufmann, 2005.