Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation

Marina Sokolova*
DIRO, University of Montreal,
Montreal, Canada
sokolovm@iro.umontreal.ca

Nathalie Japkowicz
SITE, University of Ottawa,
Ottawa, Canada
nat@site.uottawa.ca

Stan Szpakowicz
SITE, University of Ottawa
Ottawa, Canada
ICS, Polish Academy of Sciences,
Warsaw, Poland
stan@site.uottawa.ca

Abstract

Different evaluation measures assess different characteristics of machine learning algorithms. The empirical evaluation of algorithms and classifiers is a matter of on-going debate between researchers. Although most measures in use today focus on a classifier's ability to identify classes correctly, we suggest that, in certain cases, other properties, such as failure avoidance or class discrimination may also be useful. We suggest the application of measures which evaluate such properties. These measures – Youden's index, likelihood, Discriminant power – are used in medical diagnosis. We show that these measures are interrelated, and we apply them to a case study from the field of electronic negotiations. We also list other learning problems which may benefit from the application of the proposed measures.

Introduction

Supervised Machine Learning (ML) has several ways of evaluating the performance of learning algorithms and the classifiers they produce. Measures of the quality of classification are built from a confusion matrix which records correctly and incorrectly recognized examples for each class. Table 1 presents a confusion matrix for binary classification, where tp are true positive, fp – false positive, fn – false negative, and tn – true negative counts.

Class \ Recognized	as Positive	as Negative
Positive	tp	fn
Negative	fp	tn

Table 1: A confusion matrix for binary classification

Evaluation of the empirical performance of algorithms and classifiers is a matter of on-going debate among researchers[1]. In the area of the performance of algorithms, Mitchell (1997) suggests paired *t*-test as an appropriate technique of comparing the performance of two classifiers, whereas Dietterich (1998) says that only a special version of *t*-test can evaluate learners. Salzberg (1999) presents an extensive critique of empirical evaluation. He analyzes the methods currently used in ML and their statistical validity. The paper distinguishes two goals of evaluation: a comparison of algorithms, and the feasibility of algorithms on a specific domain. He states that comparison requires more statistical soundness than feasibility.

To understand the current state of the evaluation of ML algorithms, we did an exhaustive search of papers on classification. We wanted to establish a relationship between learning settings and evaluation measures employed to assess the classifiers. We surveyed the proceedings of Neural Information Processing Systems (NIPS)'03–05 (Thrun, Saul, & Schölkopf 2004; Saul, Weiss, & Bottou 2005; Weiss, Schölkopf, & Platt 2006), ICML'03–05 (Fawcett & Mishra 2003; Brodley 2004; Raedt & Wrobel 2005), and other resources. For our survey, we concluded that the clear "leaders" are papers in which evaluation is performed on data from the UCI repository, in biomedical and medical sciences, visual and text classification, and language applications. Such papers typically used accuracy, precision, recall, F-score, and the Receiver Oprating Characteristic (ROC) for their evaluation measures. Some of the papers used sensitivity and specificity measures. Our conclusions concide with those of Demsar (2006) who surveys the comparison of algorithms on multiple data sets. His survey is based on the papers published at the International Conferences on ML (ICML) 2003–2004. The author observes that algorithms are mainly compared on accuracy. McNemar's and *t*-tests are the tools used most often to generalize accuracy results.

This paper argues that the measures in use now do not fully meet the needs of learning problems in which the classes are *equally important* and where *several algorithms are compared. The deficiencies of the evaluation measures used at present lead us to suggest that we borrow measures such as Discriminant Power and likelihoods* from the field of medical diagnosis. Discriminant Power and likelihoods allow a comprehensive comparison of learning algorithms for problems in which (a) the data classes are equally important, (b) one classifier is to be chosen from among two or more classifiers, and (c) data gathering is extremely expensive (for example, when building the data requires time and manual labour). We also propose more intensive use of

*This work was done at the University of Ottawa.

[1]A theoretical assessment of algorithms is beyond the scope of this paper. We will not discuss VC dimension (Vapnik 1995), average margin (Gentile & Warmuth 1999), generalization bounds (Marchand & Sokolova 2005), and so on.

Youden's Index, and we show how it is related to the ROC. Our paper functionally relates the proposed measures to sensitivity and specifi city. This helps make a connection with the standard measures. A case study, in which ML methods are the main tool, supports our claim that new measures are necessary. We conclude by listing a few actively pursued ML applications which benefi t from the proposed measures.

Commonly-accepted performance evaluation measures

The vast majority of ML research focus on the settings where the examples are assumed to be identically and independently distributed (IID). This is the case we focus on in this study. Classifi cation performance without focussing on a class is the most general way of comparing algorithms. It does not favour any particular application. The introduction of a new learning problem inevitably concentrates on its domain but omits a detailed analysis. Thus, the most used empirical measure, *accuracy*, does not distinguish between the number of correct labels of different classes:

$$accuracy = \frac{tp \ + \ tn}{tp \ + \ fp \ + \ fn \ + tn} \tag{1}$$

On contraty, two measures that separately estimate a classifi er's performance on different classes are

$$sensitivity = \frac{tp}{tp \ + \ fn} \tag{2}$$

and

$$specificity = \frac{tn}{fp \ + tn}, \tag{3}$$

Sensitivity and specifi city are often employed in bio- and medical applications and in studies involved image and visual data.

Focus on one class is mostly taken in text classifi cation, information extraction, natural language processing, and bioinformatics. In these areas of application the number of examples belonging to one class is often substantially lower than the overall number of examples. The experimental setting is as follows: within a set of classes there is a class of special interest (usually *positive*). Other classes are either left as is – multi-class classifi cation – or combined into one – binary classifi cation. The measures of choice calculated on the positive class [2]:

$$precision = \frac{tp}{tp \ + fp} \tag{4}$$

$$recall = \frac{tp}{tp \ + \ fn} = sensitivity \tag{5}$$

$$F - measure = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \tag{6}$$

All three measures distinguish the correct classifi cation of labels within different classes. They concentrate on one class (positive examples). Recall is a function of its correctly classifi ed examples (true positives) and its misclassifi ed examples (false negatives). Precision is a function of true positives and examples misclassifi ed as positives (false positives). The F-score is evenly balanced when $\beta = 1$. It favours precision when $\beta > 1$, and recall otherwise.

A comprehensive evaluation of classifi er performance can be obtained by the ROC:

$$ROC = \frac{P(x|positive)}{P(x|negative)} \tag{7}$$

$P(x|C)$ denotes the conditional probability that a data entry has the class label C. An ROC curve plots the classifi cation results from the mostpositive classifi cation to the most negative classifi cation (Ferry *et al.* 2005). Due to the wide use in cost/benefi t decision analysis, ROC and the Area under the Curve (AUC) have found application in learning with asymmetric cost functions and imbalanced data sets (Chawla, Japkowicz, & Kolcz 2004). To get the full range of true positives and false negatives, we want easy access to data with different class balances. That is why ROC is used in experimental sciences, where it is feasible to generate much data. The study of radio signals, biomedical and medical science are a steady source of learning problems. Another possibility of building the ROC is to change the decision threshold of an algorithm. The AUC defi ned by one run is widely known as *balanced accuracy*. In this case

$$AUC_b = (sensitivity \ + specificity)/2. \tag{8}$$

The case where the IID assumption does not hold is discussed in (Rüping & Scheffer 2005). Many of such problems arise in image recognition and require a form of semi-supervised learning. Problems which require fully supervised learning have only recently begun to receive recognition under the umbrella of multiple views on the data. A problem usually postulates one question to which answers are sought from different representations of the data by several sets. The sets may belong to different populations, so the IID assumption does not hold. The fi rst practical applications came in bioinformatics. The data are not IID, and the combination of methods is required. This is done by an ensemble of classifi ers or structured learning. So far, no distinct performance measures have been introduced. Accuracy is generally used. We leave this avenue as an option for future work.

Critique of the traditional ML measures

We argue in this paper that performance measures other than accuracy, F-score, precision, recall or ROC do apply and can be benefi cial. As a preamble to our arguments, we remind the reader that ML borrowed those measures from the assessment of medical trials (Isselbacher & Braunwald 1994) and from behavioural research (Cohen 1988; Cozby 2006), where they are intensively used. Our argument focusses on the fact that the last four measures are suitable for applications where one data class is of more interest than others, for example, search engines, information extraction, medical diagnoses. They may be not suitable if

[2]Note that the same measures can be calculated for a negative class; however, they are not reported.

all classes are of interest and yet must be distinguished. For example, consider negotiations (success and failure of a negotiation are equally important) or opinion/sentiment identification (markets need to know what exactly triggers positive and negative opinions).

In such applications, complications arise when a researcher must choose between two or more algorithms (Pang, Lee, & Vaithyanathan 2002; Sokolova *et al.* 2005). It is easy to ask but rather difficult to answer what algorithm we should choose if one performs better on one class and the other – on the other class. Here we present the empirical evidence of a case where such a choice is necessary. We studied the data gathered during *person-to-person* electronic negotiations (e-negotiations); for an overview of machine learning results refer to (Sokolova *et al.* 2005). E-negotiations occur in various domains (for example, labour or business) and involve various users (for example, negotiators or facilitators). As traditional negotiations, e-negotiations are the subject of an intensive research (Nastase & Szpakowicz 2006).

The *Inspire* data (Kersten & others 2006) is the largest collection gathered through e-negotiations (held between people who learn to negotiate and may exchange short free-form messages). Negotiation between a buyer and a seller is successful if the virtual purchase has occurred within the designated time, and is unsuccessful otherwise. The system registers the outcome. The overall task was to find methods better suited to automatic learning of the negotiation outcomes – success and failure. Both classes were equally important for training and research in negotiations: the results on the positive class can reinforce positive traits in new negotiations; the results on the negative class can improve (or prevent) potentially weak negotiations. The amount of data was limited to 2557 entries, each of them a record of one bilateral e-negotiation. Successful negotiations were labelled as positive, unsuccessful – as negative. The data are almost balanced, 55% positive and 45% negative examples. The ML experiments ran Weka's Support Vector Machine (SVM) and Naive Bayes (NB) (Witten & Frank 2005) with tenfold cross-validation. Tables 2 and 3, adapted from Sokolova (2006), show the confusion matrices and the resulting evaluation measures.

Classes	SVM		NB	
	as pos	as neg	as pos	as neg
pos	1242	189	1108	323
neg	390	740	272	858

Table 2: Confusion matrixes for SVM and NB.

Measure	SVM	NB
Accuracy	77.4	76.8
F-score	81.2	78.9
Sensitivity	86.8	77.5
Specifi city	65.4	75.9
AUC	76.1	76.7

Table 3: Traditional measures of classification of e-negotiation outcomes

The overall accuracy, F-score, and the AUC are close, so do not give enough information to choose an algorithm reliably. Sensitivity and specifi city show that SVM is substantially better at identifying positive examples. Conversely, NB outperforms SVM considerably on negative examples.

Search for measures

To address the choice of the learning algorithms where both data classes are equally important, we concentrate on a comparison of measures of the algorithm's performance. Another aspect of evaluation involves the assessment of how representative our results are. This is done through the calculation of measures such as support, confidence and correlation which are often calculated via the *t*-test or the Wilcoxon-Mann-Whitney measure. Refer to (Tan, Kumar, & Srivastava 2004) for an overview of the measures of support, confidence, correlation, interest, and their key properties and to (Salzberg 1999) and (Motulsky 1995) for a discussion of when these measures are applicable.

In this study, we do not discuss this aspect of evaluation. We concentrate on the choice of comparison measures. In particular, we suggest evaluating the performance of classifiers using measures other than accuracy, F-score and ROC. As suggested by Bayes's theory (Cherkassky & Muller 1998; Duda, Hart, & Stork 2000), the measures listed in the survey section have the following effect:

accuracy approximates how effective the algorithm is by showing the probability of the true value of the class label; in other words it assesses the overall effectiveness of the algorithm;

precision estimates the predictive value of a label, either positive or negative, depending on the class for which it is calculated; in other words, it assesses the predictive power of the algorithm;

sensitivity (specificity) approximates the probability of the positive (negative) label being true; in other words, it assesses the effectiveness of the algorithm on a single class;

ROC shows a relation between the sensitivity and the specifi city of the algorithm.

F-score is a composite measure which benefits algorithms with higher sensitivity and challenges algorithms with higher specifi city. See Table 4 for a summary. Based on these considerations, we can conclude that SVM is preferable to NB. But will it always be the case? We will now show that the superiority of an algorithm (such as SVM) with respect to another algorithm largely depends on the applied evaluation measures.

Our main requirement for possible measures is to bring in new characteristics for the algorithm's performance. We also want the measures to be easily comparable. We are interested in two characteristics of an algorithm:

– the confirmation capability with respect to classes, that is, the estimation of the probability of the correct predictions of positive and negative labels;

– the ability to avoid failure, namely, the estimation of the complement of the probability of failure.

Classifier	Overall effectiveness	Predictive power	Class effectiveness
SVM	superior	superior on positive examples	superior on positive examples
NB	inferior	superior on negative examples	superior on negative examples

Table 4: Classifier capabilities shown by traditional measures

Three candidate measures that have attracted our attention have been used in medical diagnosis to analyze tests (Isselbacher & Braunwald 1994). The measures are *Youden's index* (Youden 1950), *likelihood* (Biggerstaff 2000), and *Discriminant power* (Blakeley & Oddone 1995). These measures combine sensitivity and specificity and their complements. A detailed description of the measures follows.

Youden's index The avoidance of failure complements accuracy, or the ability to correctly label examples. Youden's index γ (Youden 1950) evaluates the algorithm's ability to avoid failure. It equally weights the algorithm's performance on positive and negative examples:

$$\gamma = sensitivity - (1 - specificity) \tag{9}$$

Youden's index has been traditionally used to compare diagnostic abilities of two tests (Biggerstaff 2000). It summarizes sensitivity and specificity and has linear correspondence balanced accuracy

$$\gamma = 2AUC_b - 1. \tag{10}$$

A higher value of γ indicates better ability to avoid failure.

Likelihoods If a measure accommodates both sensitivity and specificity, but treats them separately, then we can evaluate the classifier's performance to finer degree with respect to both classes. The following measure combining positive and negative likelihoods allows us to do just that:

$$\rho_+ = \frac{sensitivity}{1 - specificity} \tag{11}$$

and

$$\rho_- = \frac{1 - sensitivity}{specificity} \tag{12}$$

A higher positive likelihood and a lower negative likelihood mean better performance on positive and negative classes respectively. The relation between the likelihood of two algorithms A and B establishes which algorithm is preferable and in which situation (Biggerstaff 2000). Figure 1 lists the relations for algorithms with $\rho_+ \geq 1$. If an algorithm does not satisfy this condition, then "positive" and "negative" likelihood values should be swapped.

Relations depicted in Figure 1 show that the likelihoods are an easy-to-understand measure that gives a comprehensive evaluation of the algorithm's performance.

Discriminant power Another measure that summarizes sensitivity and specificity is discriminant power (DP) (Blakeley & Oddone 1995):

$$DP = \frac{\sqrt{3}}{\pi}(\log X + \log Y), \tag{13}$$

- $\rho_+^A > \rho_+^B$ and $\rho_-^A < \rho_-^B$ implies A is superior overall;
- $\rho_+^A < \rho_+^B$ and $\rho_-^A < \rho_-^B$ implies A is superior for confirmation of negative examples;
- $\rho_+^A > \rho_+^B$ and $\rho_-^A > \rho_-^B$ implies A is superior for confirmation of positive examples;
- $\rho_+^A < \rho_+^B$ and $\rho_-^A > \rho_-^B$ implies A is inferior overall;

Figure 1: Likelihoods and algorithm performance

where

$$X = sensitivity/(1 - sensitivity) \tag{14}$$

$$Y = specificity/(1 - specificity) \tag{15}$$

DP does exactly what its name implies: it evaluates how well an algorithm distinguishes between positive and negative examples. To the best of our knowledge, until now DP has been mostly used in ML for feature selection, for example by Li & Sleep (2004). The algorithm is a poor discriminant if $DP < 1$, limited if $DP < 2$, fair if $DP < 3$, good – in other cases.

Relations between Youden's index, likelihood and discriminant power We want to establish how all these measures correspond to each other. Substitution of Equation 12 into Equation 11 gives:

$$sensitivity = \frac{\rho_+ - \rho_+\rho_-^{-1}}{1 - \rho_+\rho_-^{-1}} \tag{16}$$

The reverse substitution results in:

$$specificity = \frac{\rho_- - \rho_+\rho_-^{-1}}{1 - \rho_+\rho_-^{-1}} \tag{17}$$

Using the results of Equations 16 and 17 we get:

$$\gamma = \frac{(\rho_-^{-1} - 1)(\rho_+ - 1)}{\rho_+\rho_-^{-1} - 1} \tag{18}$$

This expression shows that γ favours algorithms with higher ρ_+ and lower ρ_-. This means that Youden's index corresponds to the conclusions listed in Figure 1.

To have a similar representation for DP, we recall that $\log X + \log Y = \log XY$, so

$$XY = \rho_+\rho_-^{-1} \tag{19}$$

Equation 13 becomes

$$DP = \frac{\sqrt{3}}{\pi}\log \rho_+\rho_-^{-1} \tag{20}$$

The latter expression clarifies that DP, through a positive correlation with the product of sensitivity and specificity, reinforces correctly classified examples in both classes. It diminishes the misclassified examples by negative correlation with the product of their complements.

Experiments We applied Youden's index, positive and negative likelihood and DP to the results of learning from the data of e-negotiations. We calculate the proposed measures to evaluate the algorithms' performance – see Table 5.

Measure	SVM	NB
γ	0.522	0.534
ρ_+	2.51	3.22
ρ_-	0.20	0.30
DP	1.39	1.31

Table 5: Evaluation of the classification of e-negotiation outcomes by new measures

Youden's index and likelihood values favour NB's performance. NB's γ is always higher than SVM's. This differs from the accuracy values (higher for SVM in two experimental settings) and the F-score (higher in all the three settings). The higher values of γ indicate that NB is better at avoiding failure. Further observation shows that the positive and negative likelihood favour classifiers with more balanced performance over classifiers with high achievement on one class and poor results on the other. When a classifier performs poorly, the likelihood values support the class labels of the under-performing classifier. DP' values are rather low[3]. Its results are similar to those of accuracy, of F-score which prefer SVM to NB. We attribute this correspondence to a) positive correlation of all the three measures with sensitivity; b) close values of specificity and sensitivity in our case study. As expected, Youden's index values correspond to those of the AUC. Youden's index is the most complex measure and its results (NB consistently superior to SVM) do not correlate with the standard measures. This confirms that the ability to avoid failure differs from the ability of successful identification of the classification labels. Based on these results and relations given by the scheme from Figure 1, we summarize SVM's and NB's abilities in Table 6.

Classifier	Avoidance of failure	Confirmation of classes	Discrimination of classes
SVM	inferior	superior for negatives	limited
NB	superior	superior for positives	limited

Table 6: Comparison of the classifiers' abilities by new measures

These results show that NB is marginally superior to SVM. Thus, we confirm our hypothesis that the superiority of an algorithm is related to the way in which evaluation is measured.

The results reported in Tables 3, 4, 5 and 6 show that higher accuracy does not guarantee overall better performance of an algorithm. The same conclusion applies to every performance measure if it is considered separately from others. On the other hand, a combination of measures gives a balanced evaluation of the algorithm's performance.

Conclusion

We have proposed a new approach to the evaluation of learning algorithms. It is based on measuring the algorithm's ability to distinguish classes and, consequently, to avoid failure in classification. We have argued that such approach has not been previously used in ML. The measures which originate in medical diagnosis are Youden's index γ, the likelihood values ρ_-, ρ_+ , and Discriminant Power DP. Our case study of the classification of electronic negotiations has shown that there exist ML applications which benefit from the use of these measures. We also gave a general description of the learning problems which may employ γ, ρ_-, ρ_+ and DP. These problems are characterized by a restricted access to data, the need to compare several classifiers, and equally-weighted classes.

Such learning problems arise when researchers work with data gathered during social activities of certain group. We have presented some results at conferences with a focus more general than ML. We note that the particularly apt problems include knowledge-based non-topic text classification (mood classification (Mishne 2005), classification of the outcomes of person-to-person e-negotiations (Sokolova *et al.* 2005), opinion and sentiment analysis (Hu & Liu 2004), and so on) and classification of email conversations (Boparai & Kay 2002). All these studies involve data sets gathered with specific purposes in a well-defined environment: researchers discussing the time and venue of a meeting (Boparai & Kay 2002), bloggers labelling with their moods blogs posted on a Web site (Mishne 2005), participants of e-negotiations held by a negotiation support system (Sokolova *et al.* 2005). The classification tools employed include Support Vector Machine (SVM), Naïve Bayes (NB) and Decision Trees (DT). All cited papers employ commonly used ML measures. For example, Sokolova *et al.* (2005) and Hu & Liu (2004) report accuracy, precision, recall and F-score; many other papers, especially on sentiment analysis, report only accuracy, for example (Pang, Lee, & Vaithyanathan 2002).

Our future work will follow several interconnected avenues: find new characteristics of the algorithms which must be evaluated, consider new measures of algorithm performance, and search for ML applications which require measures other than standard accuracy, F-score and ROC.

Acknowledgment

This work has been supported by the Natural Science and Engineering Research Council of Canada.

[3]Discriminant power is strong when it is close to 3.

References

Biggerstaff, B. 2000. Comparing diagnostic tests: a simple graphic using likelihood ratios. *Statistics in Medicine* 19(5):649–663.

Blakeley, D., and Oddone, E. 1995. Noninvasive carotid artery testing. *Ann Intern Med* 122:360–367.

Boparai, J., and Kay, J. 2002. Supporting user task based conversations via email. In *Proc 7th Australasian Document Computing Symposium*.

Brodley, C., ed. 2004. *Proc 21st International Conf on Machine Learning ICML'04*. AAAI Press.

Chawla, N.; Japkowicz, N.; and Kolcz, A., eds. 2004. *Special Issue on Learning from Imbalanced Data Sets*, volume 6(1). ACM SIGKDD Explorations.

Cherkassky, V., and Muller, F. 1998. *Learning from Data*. Wiley.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cozby, P. 2006. *Methods in Behavioral Research*. McGraw-Hill.

Demsar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7:1–30.

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

Duda, R.; Hart, P.; and Stork, D. 2000. *Pattern Classification*. Wiley.

Fawcett, T., and Mishra, N., eds. 2003. *Proc 20th International Conf on Machine Learning ICML'03*. AAAI Press.

Ferry, C.; Lachiche, N.; Macskassy, S.; and Rakotomamonjy, A., eds. 2005. *Proc ICML '05 workshop on ROC Analysis in Machine Learning*.

Gentile, C., and Warmuth, M. K. 1999. Linear hinge loss and average margin. In *Advances in Neural Information Processing Systems 11 (NIPS'98)*, 225–231. MIT Press.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proc 10th ACM SIGKDD International Conf on Knowledge Discovery and Data Mining KDD'04*, 168–177.

Isselbacher, K., and Braunwald, E. 1994. *Harrison's Principles of Internal Medicine*. McGraw-Hill.

Kersten, G., et al. 2006. Electronic negotiations, media and transactions for socio-economic interactions. http://interneg.org/enegotiation/ (2002-2006).

Li, M., and Sleep, M. 2004. Improving melody classification by discriminant feature extraction and fusion. In *Proc International Symposium on Music Information Retrieval ISMIR'04*, 11–14.

Marchand, M., and Sokolova, M. 2005. Learning with decision lists of data-dependent features. *Journal of Machine Learning Research* 6(Apr):427–451.

Mishne, G. 2005. Experiments with mood classification in blog posts. In *Proc 1st Workshop on Stylistic Analysis of Text for Information Access (Style2005)*. staff.science.uva.nl/gilad/pubs/style2005-blogmoods.pdf.

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

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

Nastase, V., and Szpakowicz, S., eds. 2006. *Group Decision and Negotiations, Special Issue on Formal and Informal Information Exchange in E-negotiations*, volume 15(2).

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proc Empirical Methods of Natural Language Processing EMNLP'02*, 79–86.

Raedt, L. D., and Wrobel, S., eds. 2005. *Proc 22nd International Conf on Machine Learning ICML'05*. AAAI Press.

Rüping, S., and Scheffer, T., eds. 2005. *Proc ICML 2005 Workshop on Learning With Multiple Views*. http://www-ai.cs.uni-dortmund.de/MULTIVIEW2005/MultipleViews.pdf.

Salzberg, S. L. 1999. On comparing classifiers: A critique of current research and methods. *Data Mining and Knowledge Discovery* 1:1–12.

Saul, L.; Weiss, Y.; and Bottou, L., eds. 2005. *Advances in Neural Information Processing Systems, NIPS'04*, volume 17. MIT Press.

Sokolova, M.; Nastase, V.; Shah, M.; and Szpakowicz, S. 2005. Feature selection for electronic negotiation texts. In *Proc Recent Advances in Natural Language Processing RANLP'05*, 518–524.

Sokolova, M. 2006. Learning from communication data: Language in electronic business negotiations. PhD thesis, School of Information Technology and Engineering, University of Ottawa, Canada.

Tan, P.; Kumar, V.; and Srivastava, J. 2004. Selecting the right objective measure for association analysis. *Information Systems* 29(4):293–313.

Thrun, S.; Saul, L.; and Schölkopf, B., eds. 2004. *Advances in Neural Information Processing Systems, NIPS'03*, volume 16. MIT Press.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer.

Weiss, Y.; Schölkopf, B.; and Platt, J., eds. 2006. *Advances in Neural Information Processing Systems, NIPS'05*, volume 18. MIT Press.

Witten, I., and Frank, E. 2005. *Data Mining*. Morgan Kaufmann.

Youden, W. 1950. Index for rating diagnostic tests. *Cancer* 3:32–35.