Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation*

Marina Sokolova[1], Nathalie Japkowicz[2], and Stan Szpakowicz[3]

[1] DIRO, Université de Montréal, Montreal, Canada
sokolovm@iro.umontreal.ca
[2] SITE, University of Ottawa, Ottawa, Canada
nat@site.uottawa.ca
[3] SITE, University of Ottawa, Ottawa, Canada
ICS, Polish Academy of Sciences, Warsaw, Poland
szpak@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 among researchers. Most measures in use today focus on a classifier's ability to identify classes correctly. We note other useful properties, such as failure avoidance or class discrimination, and we suggest measures to evaluate such properties. These measures – Youden's index, likelihood, Discriminant power – are used in medical diagnosis. We show that they 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 these measures.

1 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.

Table 1. A confusion matrix for binary classification

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

This paper argues that the measures commonly used now (accuracy, precision, recall, F-Score and ROC Analysis) do not fully meet the needs of learning problems in which

* We did this work while the first author was at the University of Ottawa. Partial support came from the Natural Sciences and Engineering Research Council of Canada.

A. Sattar and B.H. Kang (Eds.): AI 2006, LNAI 4304, pp. 1015–1021, 2006.

the classes are *equally important* and where *several algorithms are compared*. Our findings agree with those of [1] who surveys the comparison of algorithms on multiple data sets. His survey, based on the papers published at the International Conferences on ML 2003–2004, notes that algorithms are mainly compared on accuracy.

2 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. Classification 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:

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

Conversely, two measures that separately estimate a classifier's performance on different classes are sensitivity and specificity (often employed in biomedical and medical applications, and in studies which involve image and visual data):

$$sensitivity = \frac{tp}{tp \ + \ fn}; specificity = \frac{tn}{fp \ + tn} \tag{2}$$

Focus on one class prevails in text classification, information extraction, natural language processing and bioinformatics, where 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 classification – or combined into one – binary classification. The measures of choice calculated on the positive class[1] are:

$$precision = \frac{tp}{tp \ + fp}; recall = \frac{tp}{tp \ + \ fn} = sensitivity \tag{3}$$

$$F - measure = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \tag{4}$$

All three measures distinguish the correct classification of labels within different classes. They concentrate on one class (positive examples). Recall is a function of its correctly classified examples (true positives) and its misclassified examples (false negatives). Precision is a function of true positives and examples misclassified 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 classifier performance can be obtained by the ROC:

$$ROC = \frac{P(x|positive)}{P(x|negative)} \tag{5}$$

[1] The same measures can be calculated for a negative class, but they are not reported.

$P(x|C)$ denotes the conditional probability that a data entry has the class label C. An ROC curve plots the classification results from the most positive to the most negative classification. Due to the wide use in cost/benefit decision analysis, ROC and the Area under the Curve (AUC) apply to learning with asymmetric cost functions and imbalanced data sets [2]. 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 defined by one run is widely known as *balanced accuracy*:

$$AUC_b = (sensitivity\ + specificity)/2. \tag{6}$$

3 Critique of the Traditional ML Measures

We argue that performance measures other than accuracy, F-score, precision, recall or ROC do apply and can be beneficial. As a preamble, let us remind the reader that ML borrowed those measures from the assessment of medical trials [3] and from behavioural research [4], 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 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 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 [5,6]. 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 [6]. E-negotiations occur in various domains (for example, labour or business) and involve various users (for example, negotiators or facilitators).

The *Inspire* data [7] 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) [8] with tenfold cross-validation; see Table 2.

Table 2. Traditional classification results

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

Table 3. Classifier capabilities shown by traditional measures

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

4 Search for Measures

In this study, 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 [9,10], 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 (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 (assesses the predictive power of the algorithm);
sensitivity/specificity approximates the probability of the positive/negative label being true (assesses the effectiveness of the algorithm on a single class);
ROC shows a relation between the sensitivity and the specificity of the algorithm;
F-score is a composite measure which favours algorithms with higher sensitivity and challenges those with higher specificity. See Table 3 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.

Three measures that caught our attention have been used in medical diagnosis to analyze tests [3]. The measures are *Youden's index* [11], *likelihood* [12], and *Discriminant power* [13]. They combine sensitivity and specificity and their complements.

Youden's index. The avoidance of failure complements accuracy, or the ability to correctly label examples. Youden's index γ [11] evaluates the algorithm's ability to avoid failure – equally weights its performance on positive and negative examples:

$$\gamma = sensitivity - (1 - specificity) \tag{7}$$

Youden's index has been traditionally used to compare diagnostic abilities of two tests [12]. It summarizes sensitivity and specificity and has linear correspondence balanced accuracy (a higher value of γ means better ability to avoid failure):

$$\gamma = 2AUC_b - 1. \tag{8}$$

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:

$$\rho_+ = \frac{sensitivity}{1 - specificity}; \rho_- = \frac{1 - sensitivity}{specificity} \tag{9}$$

A higher positive 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 [12]. 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.

- $\rho_+^A > \rho_+^B$ 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;

Fig. 1. Likelihoods and algorithm performance

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) [13]:

$$DP = \frac{\sqrt{3}}{\pi}(\log X + \log Y), \tag{10}$$

$$X = sensitivity/(1 - sensitivity), Y = specificity/(1 - specificity). \tag{11}$$

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. The algorithm is a poor discriminant if $DP < 1$, limited if $DP < 2$, fair if $DP < 3$, good – in other cases.

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 4.

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

Table 4. New classification results

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. Comparison of the classifiers' abilities by new measures

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

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[2]. 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 5.

NB is marginally superior to SVM: we confirm our hypothesis that the superiority of an algorithm is related to how evaluation is measured.

The results reported in Tables 2, 3, 4 and 5 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.

5 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 thus to avoid failure in classification. We have argued this has not yet been done 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

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

more general than ML. We note that the particularly apt problems include knowledge-based non-topic text classification (mood classification [14], classification of the outcomes of person-to-person e-negotiations [6], opinion and sentiment analysis [15], and so on) and classification of email conversations [16]. All these studies involve data sets gathered with specific purposes in a well-defined environment: researchers discussing the time and venue of a meeting [16], bloggers labelling with their moods blogs posted on a Web site [14], participants of e-negotiations held by a negotiation support system [6]. All cited papers employ commonly used ML measures. For example, [6] and [15] report accuracy, precision, recall and F-score; other papers, especially on sentiment analysis, report only accuracy, for example [5].

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.

References

1. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7** (2006) 1–30
2. Chawla, N., Japkowicz, N., Kolcz, A., eds.: Special Issue on Learning from Imbalanced Data Sets. Volume 6(1). ACM SIGKDD Explorations (2004)
3. Isselbacher, K., Braunwald, E.: Harrison's Principles of Internal Medicine. McGraw-Hill (1994)
4. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale, NJ (1988)
5. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proc Empirical Methods of Natural Language Processing EMNLP'02. (2002) 79–86
6. Sokolova, M., Nastase, V., Shah, M., Szpakowicz, S.: Feature selection for electronic negotiation texts. In: Proc Recent Advances in Natural Language Processing RANLP'05. (2005) 518–524
7. Kersten, G., et al.: Electronic negotiations, media and transactions for socio-economic interactions (2006) http://interneg.org/enegotiation/ (2002-2006).
8. Witten, I., Frank, E.: Data Mining. Morgan Kaufmann (2005)
9. Cherkassky, V., Muller, F.: Learning from Data. Wiley (1998)
10. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley (2000)
11. Youden, W.: Index for rating diagnostic tests. Cancer **3** (1950) 32–35
12. Biggerstaff, B.: Comparing diagnostic tests: a simple graphic using likelihood ratios. Statistics in Medicine **19**(5) (2000) 649–663
13. Blakeley, D., Oddone, E.: Noninvasive carotid artery testing. Ann Intern Med **122** (1995) 360–367
14. Mishne, G.: Experiments with mood classification in blog posts. In: Proc 1st Workshop on Stylistic Analysis of Text for Information Access (Style2005). (2005) staff.science.uva.nl/gilad/pubs/style2005-blogmoods.pdf.
15. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proc 10th ACM SIGKDD International Conf on Knowledge Discovery and Data Mining KDD'04. (2004) 168–177
16. Boparai, J., Kay, J.: Supporting user task based conversations via email. In: Proc 7th Australasian Document Computing Symposium. (2002)