Available online at www.sciencedirect.com

ScienceDirect

Engineering Applications of Artificial Intelligence 21 (2008) 874–882

Engineering Applications of
ARTIFICIAL INTELLIGENCE

www.elsevier.com/locate/engappai

About the relationship between ROC curves and Cohen's kappa

Arie Ben-David*

Management Information Systems, Department of Technology Management, Holon Institute of Technology,
52 Golomb Street, P.O. Box 305, Holon 58102, Israel

Received 29 January 2007; received in revised form 14 September 2007; accepted 27 September 2007
Available online 26 November 2007

Abstract

Receiver operating characteristic (ROC) curves are very powerful tools for measuring classifiers' accuracy in binary-class problems. However, their usefulness in real-world multi-class problems has not been demonstrated yet. In these frequently occurring multi-class cases, simple accuracy meters that do compensate for random successes, such as the kappa statistic, are needed.

ROC curves are two-dimensional graphs. Kappa is a scalar. Each comes from an entirely different discipline. This research investigates whether they do have anything in common. A mathematical formulation that links ROC spaces with the kappa statistic is derived here for the first time. The understanding of how these two accuracy meters relate to each other can assist in a better understanding of their respective pros and cons.

Keywords: Classification accuracy; ROC curves; Area under ROC curve (AUC); Cohen's kappa; Machine learning

1. Introduction

Due to its simplicity, accuracy is by far the primary meter for assessing classifier accuracy (Lim et al., 2000; Alpaydin, 2004; Witten and Frank, 2005; Demsar, 2006). However, it is a problematic meter; in particular, it does not compensate for successes that are due to mere chance. Consider, for instance, a binary classification problem with equal probabilities of success and failure, and a classifier that makes correct classification in half of the cases. In terms of accuracy this classifier scores exactly 0.5. However, one must recall that a random classifier also scores 0.5 accuracy, a fact that makes both of them practically useless.

In an *m*-class classification problem with equal probabilities one must always measure the added value of a classifier against a random one that scores $1/m$ accuracy. Similarly, the accuracy of any classifier must be compared against that of a random one also in cases where the classes are not evenly distributed. The main problem with accuracy as a meter of performance stems from the fact that it must always be judged against the accuracy of a random classifier on similar data. Unfortunately, most authors who use accuracy as their primary meter fail to do so (Demsar, 2006), making some of their major findings questionable if not meaningless.

There are currently several alternative meters to accuracy that do take randomness into account. Here we discuss two of the major ones, namely receiver operating characteristic (ROC curves) and kappa. This paper shows for the first time that ROC curves and kappa do have a lot in common. The relationship between ROC curves and kappa is studied here, and their major pros and cons are discussed.

ROC curves, are amongst the most common alternatives to accuracy. ROC curves are two-dimensional graphs of true positives (TPs; i.e., successes) versus false positives (FPs; i.e., false alarms). They are rooted in communications back in the mid-fifties. A classifier that generates high TPs but low FPs is preferable to one that does the opposite. By varying the value of some parameter, or a threshold, in a classifier (e.g., sensitivity of a signal detector), one can tune up the model in such a way that the tradeoff between TPs and FPs is optimal (or at least acceptable) in the context of the particular application. The area under the

*Tel.: +972 3 7317977; fax: +972 3 5716481.
E-mail address: hol_abendav@bezeqint.net

0952-1976/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2007.09.009

ROC curve is often used to measure the added value, if any, of one classifier versus another (the latter is often a baseline classifier). ROC curves were found very useful in other disciplines, such as Statistics, Medical diagnosis, Biology, and more recently in Machine learning and Data mining.

Cohen's kappa (Cohen, 1960) is a scalar meter of accuracy. It was first introduced as a measure of agreement between observers of psychological behavior. Originally, Cohen's kappa was used for measuring the degree of agreement between two observers (say, A and B) of a similar phenomenon, while compensating for agreements that can be attributed to chance. It was only later found that Cohen's kappa can also be used as a meter for classifiers' accuracy. If observer A represents reality and observer B stands for a classifier—Cohen's kappa evaluates the degree of agreement (i.e., accuracy) between the classifier and reality.

ROC curves and AUCs have become quite popular in recent years for measuring classifiers accuracy. In contrast, the use of Cohen's kappa is still uncommon in Machine learning. This is despite of the fact that kappa is very popular in many other respectable disciplines such as Statistic, Psychology, Biology, and Medicine.

In principle, both ROC curves and Cohen's kappa measure classifier's accuracy, while compensating for random successes. Each does so in its own unique way. Is there anything in common between ROC curves and Cohen's kappa? This research addresses this question for the first time. The main contribution of this publication stems from the fact that it quantitatively shows that ROC curves and Cohen's kappa are closely related. They can, thus, complement each other; when the use of ROC curves and AUCs is excluded for whatever reason (for instance, in multi-class problems)—kappa is a simple and an intuitive alternative.

The paper is organized as follows: a short introduction to ROC curves and Cohen's kappa is given first. Later the relation between the two meters is presented both through examples and quantitatively. Implications of these findings are discussed later.

2. ROC curves

ROC curves describe tradeoffs between TPs and FPs. They originated in signal detection theory about six decades ago, where they were used for the tuning of signal detectors. A typical task in which ROC curves proved very useful was the determination of a threshold for detecting signals that were corrupted with noise. ROC curves were later adapted to other disciplines, such as Statistics, Medical diagnosis, and more recently—to Machine learning (Provost and Fawcett, 1997). Since the idea of using ROC curves is not new, only some basic ROC curves-related concepts that are needed for the discussion here will be mentioned.

Provost et al. (1998) gave a concise definition of ROC curves in the context of Machine learning: "ROC curves describe the predictive behavior of a classifier, independent of class distribution or error costs". In that publication they suggested a way ROC curves should be used for the ranking of classifiers.

ROC curves can be generated and be used in a variety of ways, both for research and for practical applications. These methods will not be discussed here in detail either. A comprehensive tutorial about ROC curves was given at ICML 2004. It included a long list of references, and can be found in Peter Flach's web site (Flach, 2004).

Assume that a data set from a certain application domain is available, and that it has only two equally distributed classes: positives (P) and negatives (N). Assume further that several classifiers were tested and their ROC curves were generated with their respective confidence intervals. In the best-case scenario, one ROC curve dominates all the others, in the sense that for every possible FP, its TP value is the highest or equal relative to all the rest. In a less fortunate, but frequently occurred scenario, ROC curves do intersect each other. Even in such a case it may still be possible to find a ROC curve that excels within some rang of PFs—one which is of interest in the context of the particular application. Another related technique for classifier ranking is to compute the area under the ROC curve (frequently referred to as AUC). The larger this area is—the better. The AUCs of two classifiers can be subtracted from each other, and the difference is a measure for the accuracy-wise added-value of one relative to the other.

ROC curves are basically of two dimensions. This property is clearly a virtue in two-class problems. Unfortunately, many real-world applications have multi-valued class, often with seven, 10, or even more distinct values. Although it is technically possible to draw ROC curves for every binary split of class values (e.g., class 1 versus all the rest, class 2 versus all the rest, etc.), combining many ROC curves (or AUCs) into a single-meaningful result is conceptually not intuitive, and may require expensive computation; In particular when many classes are involved. Mossman (1999) and Heckerling (2001) showed that working with three-dimensional ROC curves is possible in Medical diagnosis by calculating the volume under 3D ROC curves. However, adopting their idea to dimensions higher than three has not been put to a practical test yet, mainly due to computational complexity. We will return to this point later on.

3. Cohen's kappa

Cohen's kappa is used as a measure of classifiers accuracy in disciplines such as Statistics, Psychology, Biology and Medicine for some decades by now. Since it received only very little attention by the Machine learning community, its very basic formulae are shown here.

Cohen's kappa is defined as

$$K = \frac{P_0 - P_c}{1 - P_c}, \tag{1}$$

where P_0 is the total agreement probability, or the accuracy, and P_c is the agreement probability which is due to chance.

$$P_c = \sum_{i=1}^{I} \mathrm{P}(x_{i.})\mathrm{P}(x_{.i}), \tag{2}$$

where I is the number of class values, and $\mathrm{P}(x_{.i}), \mathrm{P}(x_{i.})$ are the columns and rows marginal probabilities, respectively.

By expressing the accuracy in terms of the confusion matrix's main diagonal probabilities and substituting (2) in (1), one gets following expression for Cohen's kappa:

$$K = \frac{\sum_{i=1}^{I}\mathrm{P}(x_{ii}) - \sum_{i=1}^{I}\mathrm{P}(x_i)\mathrm{P}(x_i)}{1 - \sum_{i=1}^{I}\mathrm{P}(x_{i.})\mathrm{P}(x_{.i})x_{.i})}, \tag{3}$$

where I, $\mathrm{P}(x_{.i}), \mathrm{P}(x_{i.})$ are as in (2), and $P(x_{ii})$ are the successful hit probabilities on the main diagonal of the confusion matrix.

Cohen's kappa ranges from −1 to 1. The theoretical range of the kappa statistic (i.e., from −1 to +1) is different from that of accuracy (0–1). However, most "reasonable" classifiers do at least as good as random or as majority-based classifiers on most real-world datasets, so by definition they score kappa higher than zero (Margineantu and Dietterich, 1997), a fact that makes the comparison of accuracy and kappa easier.

Kappa has some interesting properties that have been discussed in Ben David (2007). It has been shown there that

(A) Chance plays a significant role in typical classification problems. A benchmark of 15 datasets has shown that on the average, more than one third (!) of the hits could be attributed to chance alone.

(B) More importantly, it has been shown that ranking classifiers by kappa frequently differs from the ranking by accuracy.

Kappa is a single-scalar meter, so it is less expressive than ROC curves, which have two dimensions. However, in multi-class cases, in particular where there are many classes to consider, its simplicity becomes a virtue. More references about kappa can be found in Maclure and Willett (1987), Thompson and Walter (1988), Cook (1998) and Cicchetti and Feinstein (1990).

4. The relationship between ROC curves and kappa

ROC curves and Cohen's kappa emerged from entirely different disciplines. Both are used for measuring classifiers' accuracy for a couple of decades. Both are considered very successful by more than one respected scientific discipline by now. Is this just a coincidence? The result of a "not-invented-here" syndrome? Or do ROC curves and Cohen's kappa have anything in common that make them so useful? Answering this question may assist in a better understanding of them both.

Let us begin with six short illustrative examples. Since ROC curves are of particular usefulness in the case of binary-class problems (see above), the examples, as well as the mathematical formulation that follows, are restricted to binary-class problems.

Six binary confusion matrices are examined here. The accuracy as well as P_c and the kappa statistic are calculated for each matrix. The first three examples assume that the fraction of TP examples is identical to that of the negatives. It is also assumed that the fraction of positively classified examples is identical to that of negatively classified ones. These assumptions are relaxed later on.

Example 1. A random confusion matrix

True values	Predicted values		
	P	N	Total
P	TP = 0.25	FN = 0.25	$P = 0.50$
N	FP = 0.25	TN = 0.25	$N = 0.50$
Total	$\overline{P} = 0.50$	$\overline{N} = 0.50$	1

The first confusion matrix assumes a random "classifier". This, as well as all the confusion matrices that follow, are given in terms of probabilities. TP and TN indicate the probabilities of being TPs and true negatives, respectively, FP is the probability of false positives, and FN is the probability of false negatives. The probability of a being a positive in reality is indicated by P and that of being a negative is indicated by N .The probability of being predicted as positive is indicated by $\bar{P}$, and of being predicted as negative is indicated by $\overline{N}$.

By applying (1), (2) and (3) one gets the following results for Example 1. Accuracy = 0.5, $P_c = 0.5$, $K = 0$.

Example 1 shows that a balanced number of positive and negative binary examples that are classified at random with equal probabilities of being either positive or negative yields a single point (FP−TP) = (0.25, 0.25) in the FP–TP space, which corresponds to the point (P_c–K) = (0.5, 0) in the P_c−K space.

The second example shows a confusion matrix that results from a smarter-than-random classifier.

Example 2. A confusion matrix of a "Smarter" classifier

True values	Predicted values		
	P	N	Total
P	TP = 0.40	FN = 0.10	$P = 0.50$
N	FP = 0.10	TN = 0.40	$N = 0.50$
Total	$\overline{P} = 0.50$	$\overline{N} = 0.50$	1

Here: accuracy is P_c is 0.5, and K is 0.6.

Example 2 shows that the point (FP, TP) = (0.10, 0.40) in the FP−TP space corresponds to the point (P_c, K) = (0.5, 0.6) in the P_c–K space.

As a third example, consider a perfect classifier that results in the following confusion matrix:

Example 3. A confusion matrix of a perfect classifier

True values	Predicted values		
	P	N	Total
P	TP $=0.50$	FN $=0.00$	$P=0.50$
N	FP $=0.00$	TN $=0.50$	$N=0.50$
Total	$\overline{P}=0.50$	$\overline{N}=0.50$	1

In this case: accuracy $=1.00$, $P_c=0.5$, and $K=1.00$.

For a perfect classifier, the point (FP, TP) $=(0.00, 0.50)$ in the FP–TP space corresponds to the point $(P_c, K)=(0.5, 1.0)$ in the P_c–K space.

We turn now our attention to a case where the positive and the negative examples are not equally distributed in reality. Also, for one reason or another the classifier labels a higher portion as positives relative to their share in reality. Consider, for instance, the case shown in the confusion matrix of Example 4.

Example 4. A confusion matrix of imbalanced classes

True values	Predicted values		
	P	N	Total
P	TP $=0.65$	FN $=0.05$	$P=0.70$
N	FP $=0.15$	TN $=0.15$	$N=0.30$
Total	$\overline{P}=0.80$	$\overline{N}=0.20$	1

In this case: accuracy is 0.80, $P_c=0.625$, and $K=0.474$.

The point (FP, TP) $=(0.15, 0.65)$ in the FP–TP space corresponds to the point $(P_c, K)=(0.625, 0.474)$ in the P_c–K space.

The fifth example is of a perfect hit situation of the above-imbalanced class distribution:

Example 5. Perfect hits of imbalanced classes

True values	Predicted values		
	P	N	Total
P	TP $=0.70$	FN $=0.00$	$P=0.70$
N	FP $=0.00$	TN $=0.30$	$N=0.30$
Total	$\overline{P}=0.70$	$\overline{N}=0.30$	1

Accuracy is 1.00, $P_c\ =0.58$, and $K\ =1.00$.

The point (FP, TP) $=(0.00, 0.70)$ in the FP–TP space corresponds the point $(P_c, K)=(0.58, 1.00)$ in the P_c–K space.

Finally, a random classifier for the above examples is expected to produce a confusion matrix similar to the following:

Example 6. A confusion matrix of an imbalanced class random classifier

True values	Predicted values		
	P	N	Total
P	TP $=0.49$	FN $=0.21$	$P=0.70$
N	FP $=0.21$	TN $=0.09$	$N=0.30$
Total	$\overline{P}=0.70$	$\overline{N}=0.30$	1

In which the results are

Accuracy is 0.58, P_c is 0.58, and $K\ =\ 0$.

The point (FP, TP) $=(0.21, 0.49)$ in the TP–FP space corresponds the point $(P_c, K)=(0.58,0)$ in the P_c–K space.

The above examples show that each of the six points in the FP–TP space had a corresponding point in the P_c–K space. But is it always the case? If positive, can one express P_c and K in terms of FP and TP? Again, we restrict this question to the two-dimension case for reasons discussed earlier.

To answer these questions consider the confusion matrix of Table 1. TP, FP, FN, TN, N, P, $\bar{P}$, $\overline{N}$ in Table 1 are all probabilities, exactly as in the six examples so far. The meaning of each of them was explained in Example 1.

By the way the confusion matrix of Table 1 (and Examples 1–6) was built, the following equations do hold:

$$\bar{P}=\mathrm{TP}+\mathrm{FN}, \tag{4}$$

$$N=\mathrm{FP}+\mathrm{TN}, \tag{5}$$

$$\overline{P}=\mathrm{TP}+\mathrm{FP}, \tag{6}$$

$$\overline{N}=\mathrm{FN}+\mathrm{TN}, \tag{7}$$

$$P+N=\overline{P}+\overline{N}=\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}=1. \tag{8}$$

As discussed earlier, a ROC curve, is a two-dimension graph, most commonly of the form

$$\mathrm{TP}=f(\mathrm{FP}). \tag{9}$$

The value of f in terms of K is, thus, of interest to us here.

Table 1
A binary confusion matrix

True values	Predicted values		
	P	N	Total
P	TP	FN	P
N	FP	TN	N
Total	$\bar{P}$	$\bar{N}$	1

The value of P_c is given in (2). Using it and (4)–(8) one gets the following equation for Table 1:

$$P_c = (\mathrm{TP} + \mathrm{FP})(1 - 2N) + N. \tag{10}$$

The derivation of (10) can be found in Appendix A.1.

A special case of (10) is when $P = N = 0.5$, such as in Examples 1–3. In this case $P_c = 0.5$, regardless of the values of TP and FP.

From (10) and (2) a general relationship between TP and FP can be derived. The details can be found in Appendix A.2.

$$\mathrm{TP} = \frac{P\overline{P} + N\overline{N} - N}{1 - 2N} - \mathrm{FP}. \tag{11}$$

Recalling that the accuracy, P_0 in (1) is

$$P_0 = \mathrm{TP} + \mathrm{TN}, \tag{12}$$

the value of kappa can be expressed as

$$K = \frac{\mathrm{TP} - \mathrm{FP} - \overline{P}(1 - 2N)}{P - \overline{P}(1 - 2N)}. \tag{13}$$

For which the details are given in Appendix A.3.

Formula (11) can, thus, be re-written as

$$\mathrm{TP} = \overline{P}(1 - K)(1 - 2N) + KP - \mathrm{FP}. \tag{14}$$

Details can be found in Appendix A.4.

In the special case, where where $P = N = 0.5$

$$K = 2(\mathrm{TP} - \mathrm{FP}) \tag{15}$$

or

$$\mathrm{TP} = 0.5K + \mathrm{FP}, \tag{16}$$

as can be seen in the first three examples.

Fig. 1 gives a graphic view of several cases where $P = N = 0.5$ in the familiar FP–TP space. The leftmost point at the end of the line that is shown in Fig. 1 represents the outcome of an ideal classifier (Example 3). The rightmost point on that line belongs to a perfect disagreement "classifier". The circled point, (FT, TP) = (0.25, 0.25), belongs to a random "classifier" (Example 1). The values of the corresponding kappa statistics are shown near each point. The negative values of K represent worse-than-random outcomes, and they are usually of less interest in Machine learning (Margineantu and Dietterich, 1997). For clarity of representation, Fig. 1 does not show any particular ROC curve. Also note that the horizontal axis and the vertical axis of Fig. 1 shows FP and TP, respectively, instead of the usual false positive rate (FPR, FP/N) and true positive rate (TPR, TP/P) on 0–1 scales, one usually see in ROC curves. These conventions are kept systematic in all the figures that follow.

Fig. 2 gives a graphic view of the kappa statistic versus FP for a case similar to the one of Fig. 1 ($P = N = 0.5$). Fig. 2, thus, provides an alternative way of looking at the information that is being conveyed in Fig. 1. For instance, the point (FP, K) = (0.1, 0.6) corresponds to the point

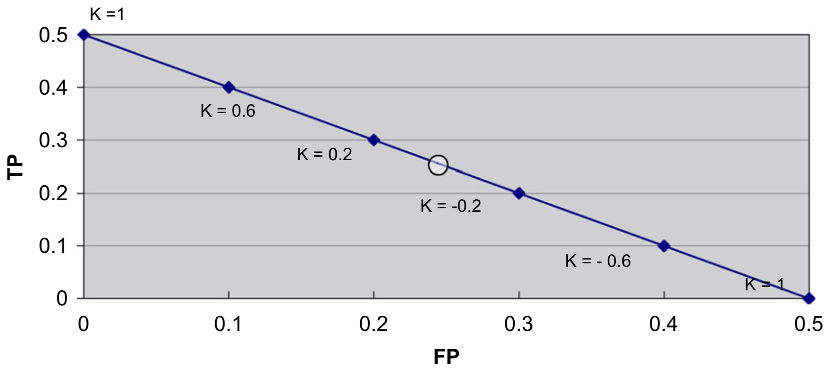

Fig. 1. TP versus FP and kappa values for $P = N = 0.5$.

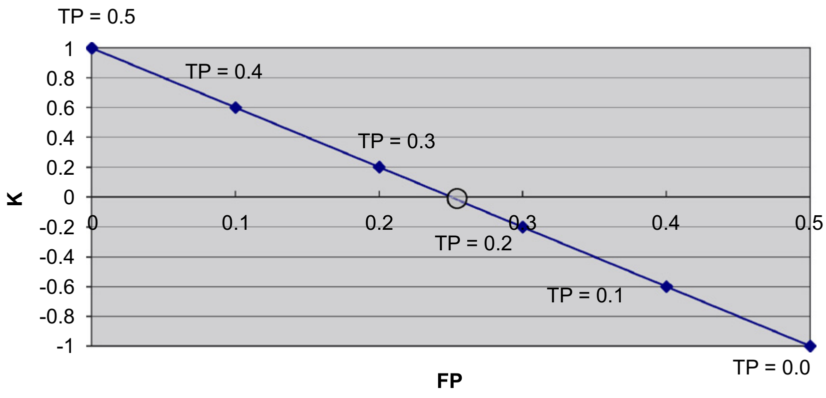

Fig. 2. Kappa versus FP for $P = N = 0.5$.

(FP, TP) = (0.1, 0.4). The values of the TPs are shown near each point. The further a point is on the left section of the line—the more accurate the model is, and vise versa. A perfect classifier is represented by the point (FP, K) = (0, 1), where TP = P = 0.5. A random classifier is represented by the point (FP, K) = (0.25, 0), marked in a circle (TP = 0.25).

One can generate ROC curves in the K–FP space in ways similar to those that are used for generating them in the FP–TP space. Let us call such curves ROC–K curves. Again, they are not shown here for the sake of clarity, but ROC–K curves do share some basic properties with ROC Curves. For example, the further "north-west" the curve is—the more accurate the classifier. The AUC for a ROC–K curve can also be calculated and serve as a measure for model accuracy, etc.

While comparing the vertical axis of Fig. 1 versus that of Fig. 2 (or Fig. 3 versus Fig. 4 later on), it is important to keep in mind that TP and kappa do not measure exactly the same thing: TP, ignores random hits on positive predictions, which is usually not a virtue of an accuracy meter. This is because one typically wants to compensate for random hit. The kappa statistic, on the other hand, does compensate for random hits, but it does so for all the hits; that is—including TNs. In general, the selection of what the vertical (horizontal) axis should represent (TP, TPR, K, or one of the many other variables that were proposed over the years) mainly depends of what should be optimized in any particular application. Also note that if the kappa statistic is to be used in this way—it will essentially become a two-dimensional metric. As such it will gain expressiveness, but on the expanse of simplicity. Based on the discussion later on, it is doubtful whether it is a good way to follow in multi-class cases.

Fig. 3 shows a case where 70 percent of the examples' true values are positive, but for one reason or another the classifier has labeled only 60 percent of the testing data as being positive (P = 0.7 and $\overline{P}$ = 0.6). Similar to Fig. 1, only relevant values of FPs are shown on the X-axis. Seven (FP, TP) points are shown in Fig. 3 with their corresponding kappa values. All the points on the sloped straight line that is shown in Fig. 3 satisfy P = 0.7 and $\overline{P}$ = 0.6.

Fig. 4 is basically similar to Fig. 2. The five straight lines that are shown in Fig. 4 share a common feature with

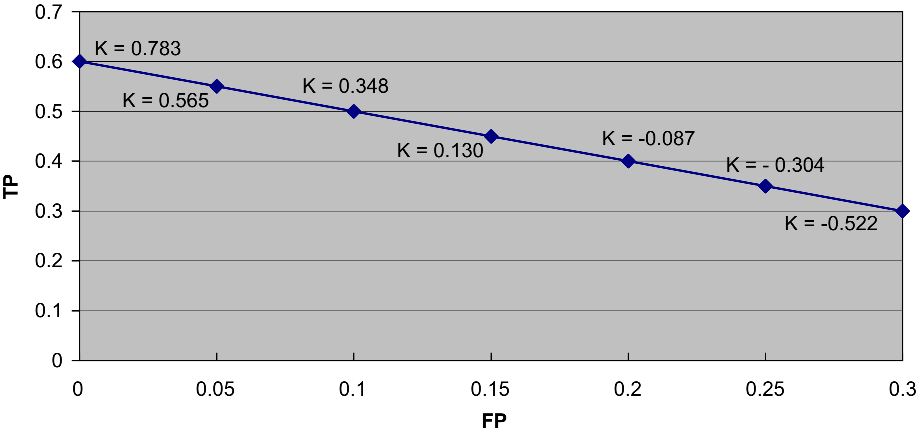

Fig. 3. TP versus FP and kappa values for P = 0.7 and $\overline{P}$ = 0.6.

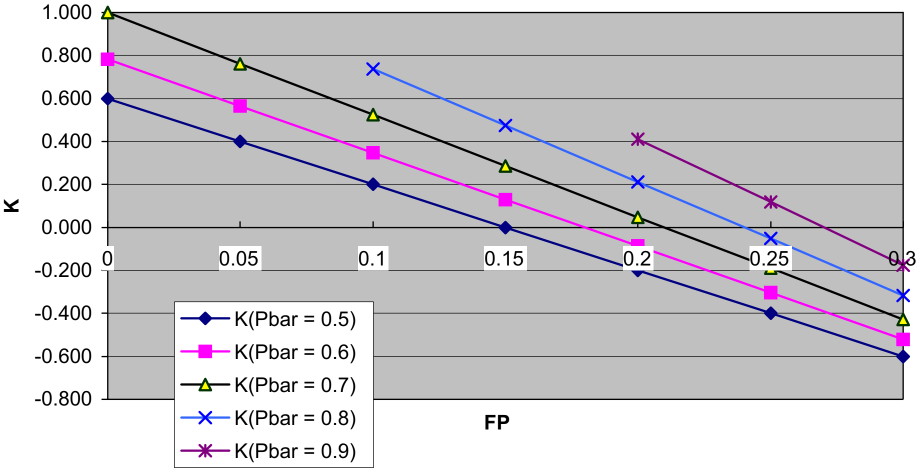

Fig. 4. Kappa versus FP for P = 0.7 and selected values of $\overline{P}$.

Fig. 3, that is $P = 0.7$ for all of them. However, the value of $\overline{P}$ varies, ranging from 0.5 (the lowest line) to 0.9 (highest line) in increments of 0.1. It is the second lowest line (the one with the square dots) that conveys the information of Fig. 3. The values of TPs of the points on this second lowest line were omitted for clarity of representation, but they can easily be found in Fig. 3.

6. Discussion

Formulae (13)–(16) show for the first time the mathematical relationship between ROC spaces and Cohen's kappa. Both ROC curves and the kappa statistic serve a very similar purpose; they are mainly used for ranking classifiers. Being of two dimensions, ROC curves can convey more information than what is possible via a single scalar such as Cohen's kappa. However, the simplicity of Cohen's kappa can be a virtue when the use of ROC curves is cost-ineffective, and/or when the interpretation of ROC curves is rather difficult in multi-class problems.

The goal of a typical Machine learning application is to find the most efficient Machine learning model for a particular data set in a cost-effective way. Consider, for example, a high-tech company, that is interested in finding a model that well predicts the success of candidates for a certain type of R&D jobs, according to a data set of past candidate profiles and their degree of success within the firm. Suppose that "success" is measured on a 10-valued scale, and that (as happens many times) the cost of errors can only be roughly estimated or it is practically unknown. An important question that the company needs to answer is whether it is worthwhile to generate ROC curves in the first place for this particular application's multi-valued data.

When used for classifier ranking, it frequently happens that ROC curves do intersect each other, such that no dominating curve can be identified (Provost and Fawcett, 2001). In this case, the area under curve method (AUC) can be used. However, while comparing many ML models (just Weka (WEKA) and Yale (YALE) have dozens of classifiers), this work can easily get out-of-hand, such that arriving at an optimal model may not been guaranteed for practical reasons.

Srinivasan (1999) showed that the optimal classifier is found on the convex hull in the ROC space. However, the dimensionality of these ROC-hyperspaces grows quadratically with the number of classes (Provost and Fawsett, 2001)—too complicated for our running example. A more recent approach, the ROCCH–hybrid method, proposed by Provost and Fawcett (2001) alleviates the need for generating all possible ROC curves, but it only works on two-valued class problems; Not a good candidate approach in our case. Mossman (1999) and Heckerling (2001) showed that working with three-dimension ROC curves is possible in Medical diagnosis by calculating the volume under 3D ROC curves. Our problem, however, is of a much a higher dimension. Due to computation complexity issues, adopting Mossman's approach to dimensions higher than three has not been put to practical tests yet.

Hand and Till (2001) proposed what they called "A simple generalization of the AUC for multiple class classification problems". They approximated a multi-class AUCs by projecting from high-dimension ROC space to two-dimension ROC curves, and by averaging for all pairs of class values. Their results were tested on ten UCI data sets. As in most ROC curve approaches, an identical error-cost assumption has been made. However, there have been no reports of implementations of their method on data sets in real-world environments, so it is impossible to get any real feedback whether their approach is successful or not, even on a qualitative level, when compared to alternative approaches.

Methods for ranking multi-class classifiers that are not based on ROC curves do exist as well. They are published in the literature, usually under the general title of "Cost-Sensitive Classification". These methods, which are often simpler and more intuitive than ROC curve-based techniques, are not central to our discussion, and will not be discussed any further. In this respect, there are also many cost-sensitive versions of the kappa statistic that are to be considered. They do have a relatively long history in other disciplines such as Decision-making and Medical diagnosis. "Weighted Kappa" Fleiss (1981) assumed equal, linear, or quadratic error-cost functions and it had many modifications thereafter. Unfortunately, there have been no comprehensive studies so far that compare the various approaches with each other on real-world multi-class data sets (ROC curve-based and weighted kappa-based methods included). It is, therefore, really impossible to point at any winning approach, if such exists at all. What can be said is that although ROC curves are very powerful tools in two-class situations—their superiority over alternative, usually simpler, meters (such as the kappa statistic) in real-world multi-class situations has not been established, nor demonstrated yet.

It is, therefore, argued here, that for the profit-seeking organization of our running example, with the current state-of-the-art body of knowledge in Machine learning, ROC curves are not necessarily always the optimal (i.e., the most cost-effective) choice. It is, thus, a very a good idea to be familiar with some alternative accuracy meters as well—the kappa statistic included. The latter has a very long history in disciplines such as Communications, Statistics, Decision-making, Biology and Medicine, and there is no apparent reason why it will not find a successful niche in Machine learning as well.

A concept similar to the ROC curves' AUC, the area under the *K*-curve in the FP–*K* space can be used for similar purposes. This idea worth further investigation, but it is not clear whether it will have any added value over the ROC curves' AUC concept. What is pretty obvious is that by doing so kappa will become a two-dimensional ROC curve-like metric. Undoubtedly, this feature will enhance its expressiveness, but on the expanse of one of its major assets—its simplicity.

7. Conclusions

It has been shown here that despite of the fact that ROC curves and Cohen's kappa were rooted in entirely different disciplines, they are very closely related concepts. A formula that expresses Cohen's kappa (among other things) in terms of TPs and FPs has been derived here for the first time.

Unlike ROC curves, Cohen's kappa is a scalar meter of accuracy. The latter expresses important properties of a point in a ROC curve space rather than describing the curve in full. ROC curves are sometimes expensive to generate; in particular, when many multi-parameter classifiers are to be tested. Furthermore, the usefulness of ROC curves has not been fully investigated nor demonstrated yet on real-world multi-class problems.

Familiarity with what has been shown here to be a closely related concept to ROC curves, the kappa scalar statistic, can be an asset when ROC curves are too complicated to work with, and/or too expensive to generate. The understanding of both ROC curves and Cohen's kappa, and in particular the close relationship that exists between them, will undoubtedly assist in correctly evaluating each meter's pros and cons' making the selection of the right accuracy meter for any particular classification problem a well-informed and an easier task than before.

Appendix A

A.1. Derivation of formula (10):

By (2) for a two-class confusion matrix

$$Pc = P\overline{P} + N\overline{N} = P\overline{P} - (1-\overline{N})N + N$$

by applying (8)

$$= P\overline{P} - \overline{P}N + N = \overline{P}(P-N) + N = \overline{P}((P+N){-}2N) + N$$

by applying (8) again

$$= \overline{P}(1-2N) + N$$

and by using (6) one gets

$$P_{\mathrm{c}} = (\mathrm{TP}+\mathrm{FP})(1-2N) + N \qquad (10)$$

A.2. Derivation of formula (11):

From (2) and (10) one can write

$$\begin{aligned} P\overline{P} + N\overline{N} &= \mathrm{TP} - 2\mathrm{TP}^{*}N + \mathrm{FP} - 2\mathrm{FP}^{*}N + N \\ &= \mathrm{TP}(1-2N) + \mathrm{FP}(1-2N) + N. \end{aligned}$$

Therefore

$$\mathrm{TP}(1-2N) = P\overline{P} + N\overline{N} - \mathrm{FP}(1-2N) - N.$$

So

$$\mathrm{TP} = \frac{P\overline{P} + N\overline{N} - N}{1-2N} - \mathrm{FP}. \qquad (11)$$

A.3. Derivation of formula (13)

Using (1), (3) and (12) one can write

$$K = \frac{\mathrm{TP} + \mathrm{TN} - \overline{P}(1-2N) - N}{1 - \overline{P}(1-2N) - N}$$

By using (8) and (5)

$$= \frac{\mathrm{TP} + N - \mathrm{FP} - \overline{P}(1-2N) - N}{P - \overline{P}(1-2N)}.$$

So

$$K = \frac{\mathrm{TP} - \mathrm{FP} - \overline{P}(1-2N)}{P - \overline{P}(1-2N)}. \qquad (13)$$

A.4. Derivation of formula (14)

From (13) one can write

$$\mathrm{TP} = KP - K\overline{P}(1-2N) + \mathrm{FP} + \overline{P}(1-2N)$$

$$= (\overline{P} - K\overline{P})(1-2N) + KP - \mathrm{FP}$$

$$= \overline{P}(1-K)(1-2N) + KP - \mathrm{FP}.$$

Therefore

$$\mathrm{TP} = \overline{P}(1-K)(1-2N) + KP - \mathrm{FP}. \qquad (14)$$

References

Alpaydin, E., 2004. Introduction to Machine Learning. MIT Press.

Ben David, A., 2007. A lot of randomness is hiding in accuracy. Engineering Applications of Artificial Intelligence 20 (7), 875–885.

Cicchetti, D.V., Feinstein, A.R., 1990. High agreement but low Kappa, in two parts. Journal of Clinical Epidemiology 43 (6), 543–558.

Cohen, J.A., 1960. Coefficient of agreement for nominal scales. Educational and Psychological Measurement, 37–46.

Cook, R.J., 1998. Kappa and its dependence on marginal rates. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of BioStatistics. Wiley, New York, pp. 2166–2168.

Demsar, J., 2006. Statistical comparisons of classifiers over multiple datasets. Journal of Machine Learning Research 7, 1–30.

Flach, P.A., 2004. The Many Faces of ROC Analysis in Machine Learning, a tutorial, ICML 2004, 〈www.cs.bris.ac.uk/~flach/〉.

Fleiss, J.L., 1981. Statistical Methods for Rates and Proportions, second ed. Wiley.

Hand, D.J., Till, R.J., 2001. A simple generalization of the area under ROC curve for multiple class classification problems. Machine Learning 45, 171–186.

Heckerling, P.S., 2001. Parametric three-way receiver operating characteristic surface analysis using Mathematica. Medical Decision Making 21, 409–417.

Lim, T.S., Loh, W.Y., Shih, Y.S., 2000. A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms. Machine Learning 40, 203–229.

Maclure, M., Willett, W.C., 1987. Misinterpretation and misuse of the Kappa statistic. American Journal of Epidemiology 126 (2), 161–169.

Margineantu, D.D., Dietterich, T.G., 1997. Bootstrap methods for the cost-sensitive evaluation of classifiers. In: Proceedings of the Seventh International Conference on Machine Learning, Morgan Kaufmann, pp. 582–590.

Mossman, D., 1999. Three-way ROCs. Medical Decision Making 19, 78–89.

Provost, F., Fawcett, T., 1997. Analysis and visualization of classifier performance-comparison under imprecise class and cost distribution. In: Heckerman, D., Mannila, H., Pregibon, D., Uthurusamy, R., (Eds.), Proceedings of the Third International Conference on Knowledge Discovery and Data Mining. AAAI Press, CA.

Provost, F., Fawcett, T., Koavi, R., 1998. The case against accuracy estimation for comparing classifiers. In: Proceedings of the 15th International Conference on Machine Learning (ICML-98).

Provost, F., Fawcett, T., 2001. Robust classification for imprecise environments. Machine Learning 4, 203–231.

Srinivasan, A., 1999. Note on the location of optimal classifiers in N-Dimensional ROC space. Technical Report, PRG-TR-2-99, Oxford.

Thompson, W.D., Walter, S.D., 1988. A reappraisal of the Kappa coefficient. Journal of Clinical Epidemiology 41, 949–958.

YALE-Yet another learning environment, CS Department, University of Dortmund, Genrmany, 〈http://www-ai.cs.uni-dortmund.de/SOFTWARE/YALE/index.html〉.

WEKA-Machine Learning Project, CS Department, University of Waikato, Hamilton, New Zealand, 〈http://www.cs.waikato.ac.nz/ml/〉.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, second ed. Academic Press.