AUK: a simple alternative to the AUC

Uzay Kaymak, Arie Ben-David, and Rob Potharst

ERIM REPORT SERIES *RESEARCH IN MANAGEMENT*	
ERIM Report Series reference number	ERS-2010-024-LIS
Publication	June 2010
Number of pages	19
Persistent paper URL	http://hdl.handle.net/1765/19678
Email address corresponding author	kaymak@ese.eur.nl
Address	Erasmus Research Institute of Management (ERIM) RSM Erasmus University / Erasmus School of Economics Erasmus Universiteit Rotterdam P.O.Box 1738 3000 DR Rotterdam, The Netherlands Phone: + 31 10 408 1182 Fax: + 31 10 408 9640 Email: info@erim.eur.nl Internet: www.erim.eur.nl

Bibliographic data and classifications of all the ERIM reports are also available on the ERIM website:
www.erim.eur.nl

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT

REPORT SERIES

RESEARCH IN MANAGEMENT

ABSTRACT AND KEYWORDS	
Abstract	The area under Receiver Operating Characteristic (ROC) curve, also known as the AUC-index, is commonly used for ranking the performance of data mining models. The AUC has many merits, such as objectivity and ease of interpretation. However, since it is class indifferent, its usefulness while dealing with highly skewed data sets is questionable, to say the least. In this paper, we propose a simple alternative scalar measure to the AUCindex, the Area Under an Kappa curve (AUK). The proposed AUK-index compensates for the above basic flaw of the AUC by being sensitive to the class distribution. Therefore it is particularly suitable for measuring classifiers' performance on skewed data sets. After introducing the AUK we explore its mathematical relationship with the AUC and show that there is a nonlinear relation between them.
Free Keywords	ROC curve, area under ROC curve, AUC, H-measure, Kappa index, AUK, model ranking, model selection
Availability	
Classifications	

AUK: a simple alternative to the AUC

Uzay Kaymak [a, b] Arie Ben-David [c] Rob Potharst [a]

[a]Econometric Institute, Erasmus School of Economics

P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands

Email: {kaymak,potharst}@ese.eur.nl

[b]School of Industrial Engineering, Eindhoven University of Technology

P.O. Box 513, 5600 MB, Eindhoven, The Netherlands

[c]Technology Management, Holon Institute of Technology

52 Golomb St., P.O. Box 305, Holon 58102, Israel

Email: hol_abendav@bezeqint.net

Abstract

The area under Receiver Operating Characteristic (ROC) curve, also known as the AUC-index, is commonly used for ranking the performance of data mining models. The AUC has many merits, such as objectivity and ease of interpretation. However, since it is class indifferent, its usefulness while dealing with highly skewed data sets is questionable, to say the least. In this paper, we propose a simple alternative scalar measure to the AUC-index, the Area Under an Kappa curve (AUK). The proposed AUK-index compensates for the above basic flaw of the AUC by being sensitive to the class distribution. Therefore it is particularly suitable for measuring classifiers' performance on skewed data sets. After introducing the AUK we explore its mathematical relationship with the AUC and show that there is a nonlinear relation between them.

Keywords

ROC curve, area under ROC curve, AUC, H-measure, Kappa index, AUK, model ranking, model selection.

1 Introduction

Receiver Operating Characteristic (ROC) curves are among the most popular tools which have been proposed over the years for ranking model performance. ROC curves are two dimensional graphs in which true positives are plotted against false positives. Historically, they have been used for finding an optimum operating point for radio and radar transmissions, but the data mining community has found out that they can also be used for studying classifier performance. Originally ROC Curves were used for binary problems, but multi-class extensions of ROC curves have also been explored [9]. More information about ROC curves in data mining can be found in [5][6]. A recent comprehensive survey which covers ROC Curves as well as other graphical methods for performance evaluation can be found in [11].

ROC curves are very useful while assessing model accuracy in classification problems. However, it is often required to encapsulate the core information they provide within a single scalar quantity. For example, when many ROC curves cross each other it is not always trivial to rank the various models. In such cases the use of scalar measures can help. Also, in order to perform statistical ranking tests, scalar performance measures are usually used [10]. Using scalar performance measures often comes at a risk of loosing valuable information, but nevertheless, there are many occasions in which their use is practically unavoidable.

The area under a ROC curve (the so-called AUC index) is one of the most commonly used scalars for ranking model performance. The AUC is very simple to calculate and interpret, so models can naturally be ranked according to their AUC index for any given data set. The AUC is independent of class priors. It also ignores misclassification costs [9]. It is objective, in the sense that every researcher will get the very same value of AUC given a trained classification model and an identical testing data set. However, AUC has its shortcomings as well. Hand has shown in [8] that using the AUC is equivalent to assuming that misclassification costs depend on the classifier, an assumption which clearly does not make sense, since these costs are problem domain dependent rather than classifier dependent. He proposed a new measure called the H-measure which is dependent, among other things, on the class priors.

The H-measure is indeed a major improvement over the AUC as it takes the priors into account. This modification is of particular importance for highly class-imbalanced, or skewed, data sets. Nevertheless, the H-measure makes some assumptions regarding the distribution of the loss functions, which may or may not be valid for some (or many) application domains.

These assumptions can be replaced by others of course, but in this case the proposed H-measure will loose the fundamental property which is required from all performance measures, that is its objectivity (i.e., different assumptions by different researchers will lead to different H-measure values). Furthermore, the H-measure must be normalized in order to transform it onto a scale that has an intuitive meaning for researchers (such as [0..1]). Although the normalization procedure is fairly simple, it makes the H-measure less intuitively interpretable.

In this work, we propose a new measure as a simple alternative for both the AUC and the H-measure. The newly proposed measure is based upon an old and quite well established measure called Cohen's Kappa [4]. Cohen's Kappa has been used in major disciplines such as Medicine and Statistics for a couple of decades for measuring classifiers' performance, but it is still less common in data mining circles. The relation between ROC curves and Cohen's Kappa has been studied in [2]. In this current work we extend Kappa to two dimensions in a way which resembles how ROC curves are being plotted (i.e., by plotting Kappa versus the false positive rate). Later we calculate the area under the Kappa curve, resulting in a scalar which we call AUK (stands for the Area Under the Kappa curve). It turns out that there is an indirect relation between AUC and AUK. Similar to the H-measure, but unlike AUC, the newly proposed measure, the AUK, is dependent on the priors. However, unlike the H-measure, the AUK does not assume any explicit loss function distribution. Similar to the AUC and the H-measure, the AUK is also objective in the sense mentioned earlier.

Since the AUK has origins in Kappa, it prefers correct classifications of the minority class to those of the majority class. We argue that what might seem a strange feature of a classifier performance measure is actually a virtue for data sets which are highly skewed. Another attractive feature of the newly proposed AUK is the fact that it is simpler to compute and easier to interpret than the H-measure. Another notable added value of the AUK is that it identifies a unique optimal model for every class distribution. In this way, the AUK can be more useful in data mining than AUC.

By showing the relation between the newly proposed AUK and the well-known AUC, as we do later on, the observations from this research can and should contribute to a better understanding and acceptance of both Cohen's Kappa and the AUK within the data mining community. Although we restrict our discussion here to binary problems, we think that extensions of the proposed AUK to multi-class problems may be easier than similar extensions to AUC and the

H-measure. This is due to the fact that multi-class versions of Kappa are already known for decades. However this issue needs to be studied separately.

Section 2 and Section 3 define the preliminaries upon which the analysis depends. The relation between ROC curves and Kappa is discussed in Section 4, where it is proposed to determine the value of Kappa as a function of the false positive rate. This results in a Kappa curve that is related to the ROC curve. The AUK index is introduced in Section 5. The relation between AUC and AUK is explored in Section 6. Guidelines for selecting the optimal model from a Kappa curve are discussed in Section 7. An illustrative numeric example is given in Section 8. Finally, conclusions and suggestions for further research can be found in Section 9.

2 Preliminaries

Consider a binary classification problem, where one has to distinguish between a *positive class* $+$ and a *negative class* $-$. The performance of a classifier can be assessed by studying the so-called *confusion matrix*. Table 1 shows a confusion matrix for a binary classification problem. Indicated by p is the fraction of positive examples that are shown to the classifier, and n is the fraction of truly negatives. The classifier predicts $\hat{p}$ of the examples as positives and $\hat{n}$ as negatives. True positives are indicated by *TP*, false negatives by *FN*, true negatives by *TN*, and false positives by *FP*.

In order to illustrate some concepts later on, Table 1 also shows values of an hypothetical example, in which the $+$ indicates fraudulent credit card transactions, while the $-$ represents legitimate ones. This common type of applications is typified by $p \ll n$ (it is most likely that the company will file for bankruptcy well before $p = n$). The company's main goal is to correctly classify the minority class ($+$), as is the case with many financial, medical, industrial, military and other applications. Nevertheless, both *FP* and *FN* do incur costs: *FP* in terms of lost revenues and/or honest clients' dissatisfaction, and *FN* in fraudulent purchases. We assume here that the exact values of these costs are unknown. All we know is that correctly identifying a $+$ is preferable to correctly identifying a $-$. Unfortunately, this is the case with many real world applications. Since we want our measure to be objective, we avoid making here additional assumptions.

Performance of a classifier can be assessed in terms of *TP*, *TN*, *FN* and *FP*. Accuracy a, for

Table 1: A confusion matrix of classifier A.

	Predicted class		
True Class	$+$	$-$	Total
$+$	*TP = 0.05*	*FN = 0.02*	$p = 0.07$
$-$	*FP = 0.03*	*TN = 0.90*	$n = 0.93$
Total:	$\hat{p} = 0.08$	$\hat{n} = 0.92$	1

example, is computed as

$$a = TP + TN. \tag{1}$$

The accuracy shown in Table 1 is $0.95(0.05+0.90)$. However, it is an unimpressive outcome considering the fact that a random or a majority based classifier would have correctly classified 93% of the cases. More importantly, 37.5% $(3/8)$ of the alarms it raises are false. The disadvantages of accuracy as a performance index will not be discussed here any further since the topic has been well covered in [4][12][10][8][2][5], to mention a few.

Many alternatives for measuring classifiers' performance have been suggested over the years. Among them are error rate, Kolmogorov-Smirnov(KS) statistic, specifity and sensitivity, precision and recall, likelihood ratios, receiver operating characteristic (ROC) curves [3][12], the area under ROC curve (AUC)[3][12], and recently, the H-measure [8].

ROC Curves and AUC have become very popular in recent years [10][8]. They are central to our discussion later, so we present the notation we use throughout this paper.

The following relations hold by the way the confusion matrix has been defined:

$$p = TP + FN \tag{2}$$

$$n = FP + TN \tag{3}$$

$$\hat{p} = TP + FP \tag{4}$$

$$\hat{n} = FN + TN. \tag{5}$$

The true positive rate t and the false positive rate f are defined as:

$$t = \frac{TP}{p}, \tag{6}$$

$$f = \frac{FP}{n}. \tag{7}$$

The true positive rate is usually referred to as the sensitivity of a model (to predict the class of interest, which is indicated here by $+$). By the same token, the false positive rate is the

complement of the specificity of the model. The more specific a model is, the better it classifies the class of interest. A specific model has a small number of false negatives, while a sensitive model predicts a large number of true positives. Ideally, the model should be both sensitive *and* specific, but these two goals often conflict, so one does need to make a trade-off between the two objectives.

In an ROC curve R, the true positive rate t (usually on the vertical axis) is plotted against the false positive rate f.When a classifier computes a continuous valued output, it can be turned into a binary one by applying a threshold to the output. If the output value is above the threshold, the pattern is classified to the $+$ class. Otherwise, it is labeled as belonging to the $-$ class. In that case, the ROC curve for the classifier can be computed by varying the value of the threshold and recording the false positive and true positive rates for all possible thresholds. Many binary classifiers fall into this category. Many classifiers also return probability estimates of a class belonging either to the $+$ or to the $-$ class, so both t and f can be computed from the corresponding cumulative distributions.

An ROC curve R is, thus, a function g of f. In other words, R is characterized by

$$t = g(f). \tag{8}$$

Example ROC curves are shown in Figure 1(a).

ROC Curves are very powerful visualization tools for comparing the performance of two or more classifiers. When a ROC curve of classifier A dominates the one of B (i.e., it has higher t values for all values of f), one can conclude that A is preferred to B. This case is demonstrated in Figure 1(a). ROC Curves can generally convey more information than what is possible by a single scalar metric. However, often ROC curves do intersect, making such a judgment less obvious. Also their use for multi-class problems is even more complicated [9]. Very often, such as when one wants to perform a ranking statistical test, a simple objective scalar metric is needed[10].

The area under an ROC curve (the so-called AUC index) is a well-known scalar measure of the overall performance of a classifier, averaging across different thresholds that can be used to generate a classifier. In recent years it has gained much popularity in the machine learning and data mining literature. According to our notation, an AUC is defined as

$$\mathrm{AUC} = \int_0^1 t \, df. \tag{9}$$

In general, a model with a larger AUC is preferred to a model with a smaller one. The AUC of a random classifier is 0.5, since the ROC curve of a random classifier is the straight line $t = f$. Since one is typically interested in classifiers which have some positive added value over random classifiers, the AUC range which is of interest [0.5..1] can be linearly transformed to the range [0..1], resulting in the Gini coefficient. The Gini coefficient is equal to twice the area between the ROC curve and the diagonal. Simple geometry dictates that

$$\mathrm{GINI} = 2\mathrm{AUC} - 1. \tag{10}$$

The AUC has some attractive features, most notably its objectivity. Given a testing data set and a classifier, every researcher will get the same AUC. The AUC is equivalent to the Mann-Whitney-Wilcoxon statistic and it also has some very intuitive statistical interpretations which will not be discussed here [8]. However, the AUC does suffer from some drawbacks. For instance, when two (or more) ROC curves cross each other, it is not clear whether the classifier with the larger AUC is always preferable. Moreover, a recent study by Hand [8] has pointed at a much more fundamental problem with AUC, one which has not attracted almost any attention so far. Hand showed that the AUC uses different classification cost distributions for different classifiers. He also suggested an alternative to the AUC, called the H-measure. Here we propose another alternative to the AUC, called the AUK (the Area Under Kappa curve), one which solves the AUC's key problem without resorting to prior explicit assumptions about loss function distributions as the H-measure does.

In the following sections, we develop the new measure, the AUK, and discuss some of its key features. We also show how the AUK relates to the AUC, and how by using the AUK one can find an optimal classifier in a straightforward manner. We later show an example comparing it with the AUC. Although outside the scope of this paper, we also shortly discuss the multi-class case, arguing that by adopting the proposed measure, using a splitting method (one against all, etc.) is not required at all.

Let us begin with a short introduction to Cohen's Kappa, which our proposed measure relies upon.

3 Cohen's Kappa

Since its original proposal in 1960 [4] Cohen's Kappa has gained an increasing popularity as a performance measure in disciplines such as medicine, psychology, and statistics. It is less popular in data mining circles though. For this reason we devote here a section which re-introduces it and highlights some of its key features.

Cohen's Kappa is a scalar defined as:

$$\kappa = \frac{a - p_c}{1 - p_c}. \tag{11}$$

Herein a is the accuracy as defined in (1), and p_c is defined as the probability of predicting the correct class due to chance:

$$p_c = p\hat{p} + n\hat{n}. \tag{12}$$

Note that the assumption behind (12) is that each test case is randomly classified as $+$ or $-$ with probabilities $\hat{p}$ and $\hat{n}$ respectively.

Kappa has some characteristics that make it suitable for assessing classifier performance. It takes into account the priors, a property which is shared by the H-measure, but not the by the AUC and the Gini coefficient. As can be seen, Kappa is simple to calculate and to interpret. Kappa ranges from -1 to +1. Any random and majority based classifier results in $\kappa = 0$. Negative Kappa values indicate worse than random performance, so these values are usually of no interest to the data mining and machine learning communities. Similar to the Gini coefficient, on the positive range Kappa measures the added value of a classifier over a random or a majority based one. Unlike the H-measure, there is no need to normalize it separately onto the scale $[0 \ldots 1]$, a fact which simplifies it and makes it very intuitive.

Multi-class versions of Kappa as well as their cost-sensitive variants do exist since the Eighties. The basic idea is to add a weighing coefficient w_{ij} (which may be linear or non-linear) to (11), reflecting the severity or cost of an error of classifying an object i to class j. A detailed discussion of this topic is outside the scope of this paper, but it worth mentioning here that these cost-sensitive extensions to Kappa are encapsulated in a single formula, which does not require the splitting of a data set into (potentially) many binary problems as suggested for example in [9]. This property makes the multi-class, cost-sensitive, version of Kappa relatively simple to calculate and interpret. The interested reader can find more details in [7].

Table 2: A confusion matrix of classifier B.

	Predicted class		
True Class	$+$	$-$	Total
$+$	*TP = 0.04*	*FN = 0.02*	$p = 0.06$
$-$	*FP = 0.03*	*TN = 0.91*	$n = 0.94$
Total:	$\hat{p} = 0.07$	$\hat{n} = 0.93$	1

Table 3: A confusion matrix of classifier C.

	Predicted class		
True Class	$+$	$-$	Total
$+$	*TP = 0.06*	*FN = 0.02*	$p = 0.08$
$-$	*FP = 0.03*	*TN = 0.89*	$n = 0.92$
Total:	$\hat{p} = 0.09$	$\hat{n} = 0.91$	1

Perhaps one of the major reasons Kappa has not gained a wide acceptance within the data mining community is the well known fact that it favors correct classifications of the minority class over those of the majority one. We demonstrate this feature through an example. We have seen the confusion matrix of classifier A in Table 1. Suppose now that another classifier, say B, results in the confusion matrix shown in Table 2, and that another classifier, C, gives the confusion matrix shown in Table 3. Despite of the fact that all three classifiers, A, B, and C correctly classified 95 percent of the test cases, the Kappa values are not identical: $\kappa_A = 0.640$, $\kappa_B = 0.589$, and $\kappa_C = 0.679$. Note that classifier B was penalized, while the Kappa value resulting from C improved relative to those of A.

Recalling that (11) does not explicitly include any cost sensitive component whatsoever, the results we have just got may seem rather weird at a first glance. Although we have not explicitly quantified any trade-off between the benefits of correct classifications, the way Kappa compensates for random successes did have an effect. These implicit trade-offs may be modified of course in the cost sensitive versions of Kappa discussed earlier. However, we argue that this property of Kappa is actually a virtue for many real world classification problems in which it is more important to correctly classify the minority rather than the majority class, while the precise values of these trade-offs are rather vague. Having assumed here that all classification costs or benefits are unknown, and that we only prefer to correctly classify the minority class

over correct classifications of the majority class, Kappa seems to make sense. In the coming sections we will use the basic definition of Kappa for binary problems, as shown in (11). In the next section we study the relation between Kappa and ROC.

4 The relation between ROC and Kappa

The relation between ROC and Cohen's Kappa has been studied in [2]. That analysis has formulated a relation between ROC and Kappa given the percentage $\hat{p}$ of positively classified samples. This number, however, is dependent on the model and can be different for each classifier. In this paper, we extend the analysis further by also quantifying the relation between ROC and Cohen's Kappa in terms of the percentage p of positive samples in the data set.

We re-write now accuracy as:

$$a = tp + n(1 - f). \tag{13}$$

Substituting (13) and (12) into (11) gives

$$\kappa = \frac{pt + n(1 - f) - p\hat{p} - n\hat{n}}{1 - p\hat{p} - n\hat{n}}. \tag{14}$$

By using the relations (2) to (5), we can re-write (14) as

$$\kappa = \frac{pt - (1 - p)f - (2p - 1)\hat{p}}{p - (2p - 1)\hat{p}}. \tag{15}$$

Hence, for each point in the ROC space, we can compute the corresponding κ value using (15).

By re-arranging (15), it is possible to express the ROC curve in terms of Kappa as:

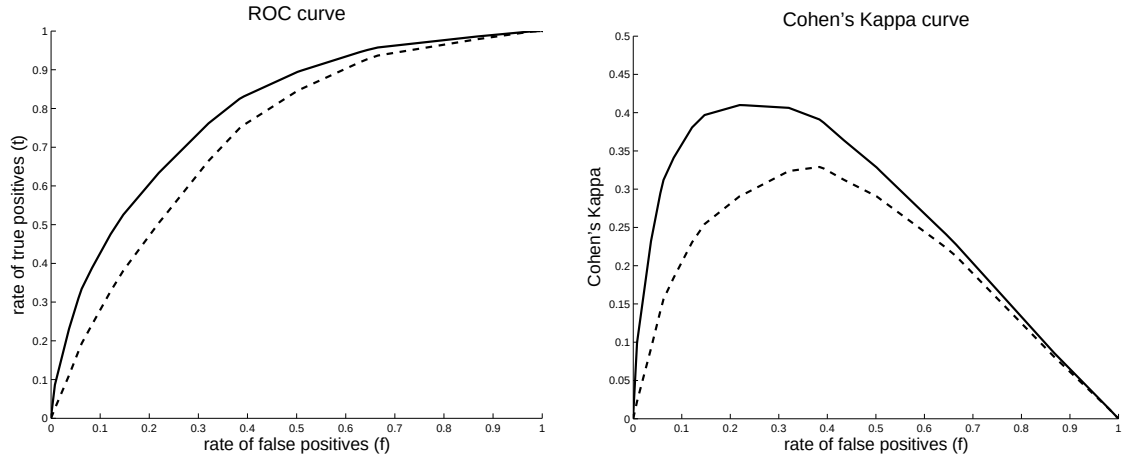
$$t = \frac{1 - p}{p} f + (1 - \kappa)\frac{2p - 1}{p}\hat{p} + \kappa. \tag{16}$$

In (15), Kappa is expressed as a function of ROC in terms of the percentage positive samples predicted by the model. Since this number is different for each model, comparison is easier if Kappa is expressed in terms of the percentage of positive samples in the data. Note that

$$\hat{p} = pt + (1 - p)f = f + p(t - f). \tag{17}$$

Then, Kappa can be written as

$$\kappa = \frac{pt - (1 - p)f - (2p - 1)(pt + (1 - p)f)}{p - (2p - 1)(pt + (1 - p)f)}. \tag{18}$$



(a) (b)

Figure 1: Two ROC curves and the corresponding Cohen's Kappa curves.

Although the above equations are useful for converting ROC into Kappa, they are not very illuminative for understanding the relation between ROC and Kappa. This relationship is clearer if we re-arrange the terms in (18) to yield

$$\kappa = \frac{2p(1-p)(t-f)}{p + (1-2p)f + p(1-2p)(t-f)} = h(t-f), \tag{19}$$

where h is a non-linear function. In other words, Kappa can be seen as a nonlinear transformation of the difference between the true positive rate and the false positive rate. Note, however, that the false positive rate is a parameter of the non-linear function h. Also, the line $t = f$ is the ROC of a random model. Hence, Kappa can also be interpreted as a nonlinear transformation of the difference between the ROC of the model to be evaluated and the ROC of a random one. Therefore, Kappa is related to the performance improvement that a classifier achieves over the performance of a random classifier with the same false positive rate.

Figure 1 shows the ROC curves for two classifiers together with their corresponding Kappa curves. In this simple example, the classifier corresponding to the solid line dominates the classifier corresponding to the dashed line, since the solid ROC curve is closer to the upper left corner of the ROC space than the dashed ROC curve. Similarly, we observe that the solid Kappa curve dominates the dashed Kappa curve as it is higher than the dashed curve for all values of the false positive rate. Nevertheless, in realistic examples, such as the one to be discussed in Section 8, ROC curves frequently do intersect each other, making the visual rankings of the various models rather difficult if not impossible.

The special case of $p = n = 0.5$ for the relation between the ROC and Kappa should be

mentioned separately. In that case, the following relation holds:

$$\kappa = t - f. \tag{20}$$

Hence, when there are as many positive samples in the data as there are negative samples, Cohen's Kappa is equal to the ROC improvement that a classifier brings over a random model. In this case, Cohen's Kappa provides exactly the same information as the ROC curve.

5 The AUK index

The above discussion exposes the close relation between an ROC curve (true positive rate plotted against the false positive rates) and the Kappa curve (Kappa values plotted against the false positive rate). Just like the area under the ROC curve (AUC) is an indication of the overall performance of a classifier, the area under a Kappa curve can also be seen as such an indication. Therefore, we propose to plot Kappa against the false positive rate as an alternative way of analyzing the performance of a classifier and to assess its overall performance by computing the area under the Kappa curve. We call this new index the AUK index (AUK stands for the Area Under the Kappa curve).

The AUK is defined as:

$$\text{AUK} = \int_0^1 \kappa(f)\, df. \tag{21}$$

Note that AUK already accounts for the uninteresting area below the main diagonal. Hence, it is a measure that is more focused than the AUC index. Furthermore, Kappa inherently takes into account the class skewness in the data, and so the AUK is expected to be a more powerful index than the AUC.

Since Kappa is a transformation of the difference between the ROC curve of a classifier and the ROC curve of a random model, AUK can be seen as a transformation of the area between the ROC curve and the ROC of a random model, one which compensates for skewness and for random successes.

6 The relation between AUC and AUK

Since AUC is the area under an ROC curve and ROC is related to κ, AUK is related to the AUC. AUK is a transformation of the area under ROC that is above the main diagonal. It can therefore

be expected that AUK is a function of the difference between the model AUC and the AUC of a random classifier (which is equal to the constant 0.5).

For the special case where $p = n = 0.5$, the AUK can be computed straightforwardly as

$$\text{AUK} = \int_0^1 \kappa \, df = \int_0^1 (t - f) \, df \tag{22}$$

$$\text{AUK} = \text{AUC} - 0.5. \tag{23}$$

Hence, in this special case, AUK and AUC differ only by a constant. Furthermore, by substituting (10) and re-arranging it is found that

$$\text{AUK} = \frac{1}{2}\text{GINI}. \tag{24}$$

In other words, AUK is equal to half the Gini coefficient when there is no class skew in the data set.

7 Selecting the optimal threshold

There are a number of advantages to using the kappa curve and the AUK. First of all, Kappa as a measure accounts for correct classifications due to chance, so that the added value of a classifier above a random one can be assessed immediately. Secondly, the Kappa curve inherently accounts for class skewness. From (19) it can be seen that the percentage of positive class samples appears explicitly in the transformation. This introduces a correction for class skew. Thirdly, the Kappa curve usually has a unique maximum that can be used to select the optimal model (by Kappa) from the set of models that have been used to generate the ROC and/or Kappa curves. Although it can happen that there are multiple maxima to the Kappa curve, in many practical problems a unique optimum will be present to select the optimal threshold for generating an optimal classifier that maximizes the Kappa index.[1] Of course, one has to be careful about sampling effects and be aware that the training data is just one realization of the underlying distribution. However, such considerations can be overcome by re-sampling techniques such as bagging. In that case, the expected Kappa curve can be generated by averaging over a large number of samples from the same underlying population.

[1] A fourth advantage can be mentioned that Kappa can be applied for multi-class problems, but this aspect, briefly mentioned earlier, has not been studied here.

Finding the maximum of the Kappa curve can be achieved by taking the derivative of the curve and setting it equal to zero. Consider first the special case without class skew, i.e $p = 0.5$. Then,

$$\frac{d\kappa}{df} = \frac{d(t-f)}{df} = \frac{dt}{df} - 1 = 0. \tag{25}$$

Hence, the value of κ is maximized when the gradient of the ROC curve equals 1. This is a well-known method of selecting the optimal model from an ROC curve [6][5]. The above equation shows that it is optimal in the sense that it maximizes Cohen's Kappa when the data set is balanced.

When there is a class skew in the data set, the derivative of the Kappa curve becomes more complex. By taking the derivative of the Kappa curve and setting it equal to zero one finds that the following condition should be satisfied to maximize κ:

$$\frac{dt}{df} = 1 + \frac{(1-2p)(t-f)}{p+(1-2p)f}. \tag{26}$$

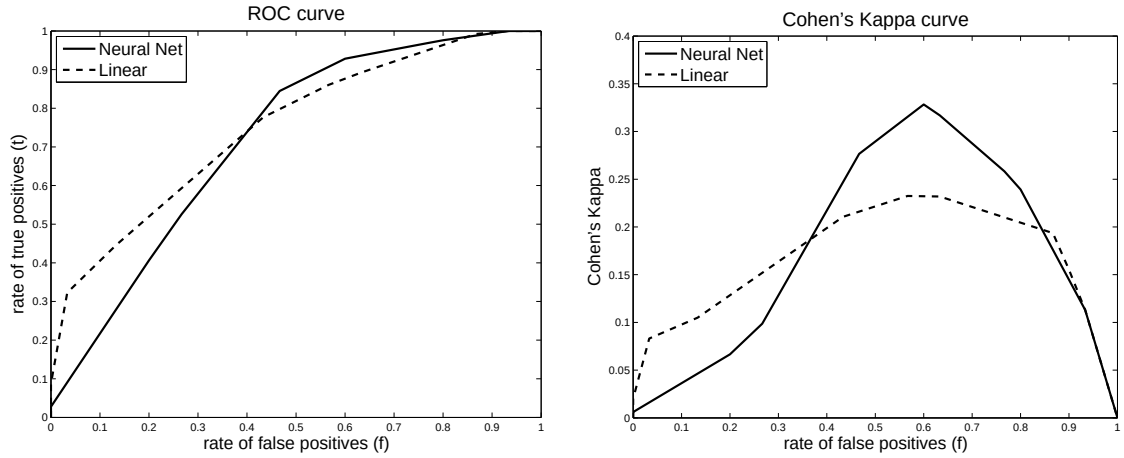
Equation (26) shows that when $p < 0.5$ (which is usually the case), the second term is positive, and hence the optimal point that maximizes κ shifts towards the left of the ROC curve so that the classifier needs to be more discriminative.[2] By maximizing κ it now becomes possible to compare different classifiers as well since a unique condition for optimality has been established. In other words, we can find the classifier that maximizes accuracy (discounted by the chance events) in a very elegant way.

8 An example

In this section we consider an example to illustrate Kappa curves and the AUK index. We have used the Statlog German Credit data set from the UCI Machine Learning Repository [1]. Since the AUK measure takes into account the skewness of the data, we produced a highly skewed version of the original UCI data set by randomly deleting samples of the minority class until they were 11% of the total. This example demonstrates that although AUK and AUC are related, the rankings they produce may differ.

The data set contains 768 samples. We have sampled the data randomly to divide it into a training set consisting of 500 patterns and a test set consisting of the remaining examples. By

[2]This assumes that one has a model that is better than a random one, i.e. $(t-f) > 0$, which is usually the case in data mining.



(a) (b)

Figure 2: Convex ROC curves (a) and Kappa curves (b) for the Statlog German Credit data.

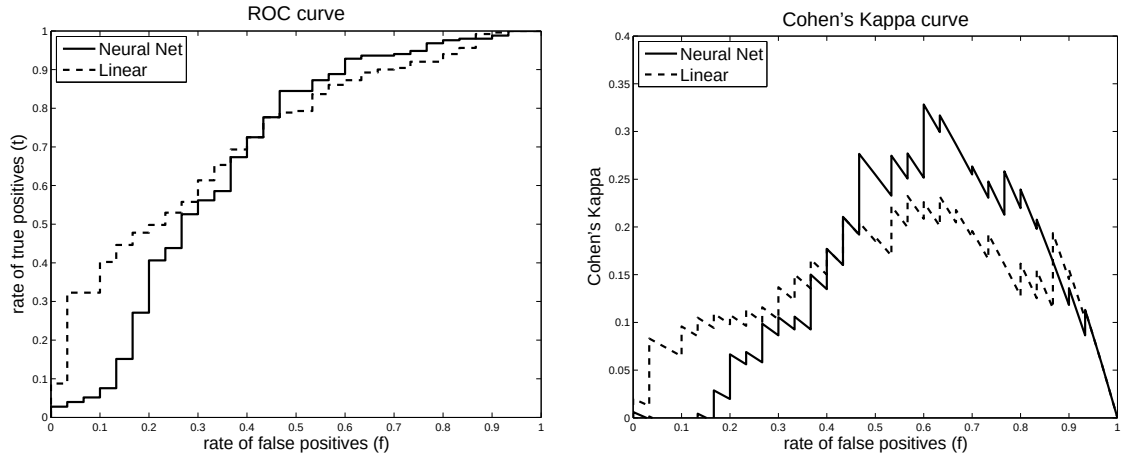
Table 4: AUC and AUK corresponding to curves shown in Figure 2.

	Linear	Neural Net
AUC	0.7493	0.7186
AUK	0.1672	0.1753

using the training data, we have generated two models. The first model is a linear regression. The second is a neural network with one hidden layer containing five neurons. We used the standard logistic function as the transfer function both in the hidden and the output layer's nodes. The experiments were performed in Matlab by using Mathworks' Neural Networks Toolbox. The ROC and the Kappa curves have also been computed in Matlab.

Figure 2(a) shows the convex-hull ROC curves of the two models computed by using the algorithm described in [5]. Figure 2(b) shows the Kappa curves computed from the ROC curves by using relation (19). Note that both the ROC curves and the Kappa curves intersect. Hence, it is not trivial to determine which model is preferable. In order to rank the models, we have computed the AUC and the AUK. The results are shown in Table 4.

Table 4 shows that according to AUC, the linear model is preferable to the neural network. However, AUK shows the reverse ranking. This demonstrates the fact that although there is a relationship between AUK and AUC, the rankings of AUC and AUK may differ from each other. This is not always the case, of course, since AUK and AUC are correlated. For instance, for the original Statlog German Credit data set (with 30 percent minority class) the rankings by AUK and AUC were identical, and they are not shown here.



(a) (b)

Figure 3: Non-convex ROC curves (a) and kappa curves (b) for the Statlog German Credit data.

Table 5: AUC and AUK corresponding to curves shown in Figure 3.

	Linear	Neural Net
AUC	0.7227	0.6781
AUK	0.1404	0.1449

Similar results were obtained when the ROC and Kappa curves have not made convex, as shown in Figure 3. These ROC curves have also been computed according to the method discussed in [5]. The corresponding values of AUC and AUK are shown in Table 5. Note that the rank reversal shown Table 4 is also observed here. The rankings of the models when applied to the original data set remained identical also in this version. These results are also omitted here for the sake of clarity.

The Kappa values from Figure 2(b) can be used for identifying the optimal operating point (and the corresponding threshold) for the best performing model (the neural network). The optimal false positive rate that maximizes Kappa is 0.60. This corresponds to a maximum Kappa value of 0.328. At this operating point, the true positive rate is 0.928 as can be seen in Figure 2(a).

9 Conclusions and further research

A new measure for classifier performance, nicknamed AUK, has been at the focus of this paper. The relationship between ROC curves and Kappa has been discussed. Our analysis is an exten-

sion of the work presented in [2]. We have shown that there is a one-to-one relation between ROC and Kappa values, so one can be easily converted to the other. In fact, kappa is a nonlinear transformation of the difference between a model's ROC value and the ROC of a random model. When the data set is balanced (i.e., $p = 0.5$), the relation between ROC and Kappa is the simplest: Kappa is then equal to the difference between a model's ROC and the ROC of a random model.

We have introduced here the Kappa curve, which plots Kappa values against false positive rates. The close relation between the ROC and Kappa curves implies that there is also a close relation between the area under an ROC curve and the area under a Kappa curve. To quantify the area under the Kappa curve, we have proposed a new index, the AUK. It has been shown that AUC and AUK are related to one another. AUK can be used for model performance evaluation in a similar way the AUC does. The difference is mainly that AUK accounts for class skewness in the data, while AUC does not. Therefore it seems that AUK is a better measure of a model's performance. The example has demonstrated that using AUK instead of AUC may lead to different rankings. When the data set is balanced, the new index, AUK, is simply half the Gini coefficient. In this particular case the AUK and the AUC differ by a constant.

Kappa curves can also be used for selecting an optimal model. The Kappa curve often has a unique maximum. In this way, the problem of selecting a suitable threshold for a model can be solved. This is because the threshold that maximizes Kappa is typically unique. For balanced data, Kappa is maximized at the point where the gradient of the ROC curve is equal to 1. In presence of class skewness with $p < 0.5$, the maximum of Kappa corresponds to a point where the gradient of the ROC curve exceeds 1.0.

We have also discussed the fact that Kappa prefers correct classifications of the minority class over those of the majority class. This feature has no effect when the data set is balanced, in which case Kappa still compensates for random successes. However, we have argued that when the data set is highly skewed, it is usually the case that one prefers to correctly classify the minority class. In this very common case, using AUK is more realistic, and preferable to using the AUC as a performance index.

We have also argued that the AUK is very simple and intuitive when compared with the H-measure. There are many real world cases where the exact loss function is unknown. Under these circumstances the implicit assumptions which are inherent to the AUK seem to us

preferable to the explicit assumptions upon which the H-measure is based.

It has also been mentioned that simple cost-sensitive versions of Kappa do exist for a long time. One of the advantages of Cohen's Kappa is that it has been defined and applied for multi-class problems without resorting to any splitting of the data sets (i.e., no one-versus-all or any other splitting method is needed). Although outside the scope of this paper, it seems reasonable to guess that these versions of Kappa can be used for generating cost-sensitive multi-class Kappa curves. Once this is done, it will be possible to extend the AUK in a way similar to what has been done here. While doing so one should be very careful not to introduce application-specific assumptions (such as about the costs) which will eliminate the objectivity of the proposed measure. Our analysis in this paper has been focused on a binary classification problem, where we have only assumed that one prefers to correctly classify the minority over the majority class. No other explicit assumption about benefits or loss functions have been made. The extension of our analysis to multi-class, cost-sensitive classification problems is left for future research.

References

[1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[2] A. Ben-David. About the relationship between ROC curves and Cohen's kappa. *Engineering Applications of AI*, 21:874–882, 2008.

[3] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

[4] J. A. Cohen. A coefficient of agreement for nominal scales, educational and psychological measurement. *Psychological Measurement*, 20:37–46, 1960.

[5] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[6] P. A. Flach. The many faces of ROC analysis in machine learning. ICML 2004 (a tutorial), ICML, 2004.

[7] J. L. Fleiss. *Statistical Methods for Rates and Proportions*. John Wiley and Sons, 2nd edition, 1981.

[8] D. J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77:103–123, 2009.

[9] D. J. Hand and R. J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45:171–186, 2001.

[10] D. Janez. Statistical comparisons of classifiers over multiple datasets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[11] R. C. Prati, G. Baptista, and M. C. Monard. A survey on graphical methods for classification predictive performance evaluation. Personal Correspondance, Mathematics Depts., UFABC and USP Universities, Sao Paulo, Brazil, 2010 (submitted to TKDE), 2010.

[12] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 44:203–231, 2001.

Publications in the Report Series Research* in Management

ERIM Research Program: "Business Processes, Logistics and Information Systems"

2010

Linearization and Decomposition Methods for Large Scale Stochastic Inventory Routing Problem with Service Level Constraints
Yugang Yu, Chengbin Chu, Haoxun Chen, and Feng Chu
ERS-2010-008-LIS
http://hdl.handle.net/1765/18041

Sustainable Passenger Transportation: Dynamic Ride-Sharing
Niels Agatz, Alan Erera, Martin Savelsbergh, and Xing Wang
ERS-2010-010-LIS
http://hdl.handle.net/1765/18429

Visualization of Ship Risk Profiles for the Shipping Industry
Sabine Knapp and Michel van de Velden
ERS-2010-013-LIS
http://hdl.handle.net/1765/19197

Intelligent Personalized Trading Agents that facilitate Real-time Decisionmaking for Auctioneers and Buyers in the Dutch Flower Auctions
Wolfgang Ketter, Eric van Heck, and Rob Zuidwijk
ERS-2010-016-LIS
http://hdl.handle.net/1765/19367

Necessary Condition Hypotheses in Operations Management
Jan Dul, Tony Hak, Gary Goertz, and Chris Voss
ERS-2010-019-LIS
http://hdl.handle.net/1765/19666

Human Factors: Spanning the Gap between OM & HRM
W. Patrick Neumann, and Jan Dul
ERS-2010-020-LIS
http://hdl.handle.net/1765/19668

AUK: a simple alternative to the AUC
Uzay Kaymak, Arie Ben-David, and Rob Potharst
ERS-2010-024-LIS
http://hdl.handle.net/1765/19678

---

* A complete overview of the ERIM Report Series Research in Management:
https://ep.eur.nl/handle/1765/1

ERIM Research Programs:
LIS Business Processes, Logistics and Information Systems
ORG Organizing for Performance
MKT Marketing
F&A Finance and Accounting
STR Strategy and Entrepreneurship