Pattern Recognition Letters 34 (2013) 492–495

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance?

D.J. Hand, C. Anagnostopoulos *

Department of Mathematics, South Kensington Campus, Imperial College London, London SW7 2AZ, UK

ARTICLE INFO

Article history:
Received 14 May 2012
Available online 12 December 2012

Communicated by S. Sarkar

Keywords:
Area under the curve
Classification
Gini coefficient
ROC curve
Screening

ABSTRACT

The area under the receiver operating characteristic curve is a widely used measure of the performance of classification rules. This paper shows that when classifications are based solely on data describing individual objects to be classified, the area under the receiver operating characteristic curve is an incoherent measure of performance, in the sense that the measure itself depends on the classifier being measured. It significantly extends earlier work by showing that this incoherence is not a consequence of a cost-based interpretation of misclassifications, but is a fundamental property of the area under the curve itself. The paper also shows that if additional information, such as the class assignments of other objects, is taken into account when making a classification, then the area under the curve is a coherent measure, although in those circumstances it makes an assumption which is seldom if ever appropriate.

1. Introduction

Given a set of objects, for each of which we have a feature vector of descriptive variables and for each of which we also know to which one of a mutually exclusive set of classes it belongs, the aim of supervised classification is to construct a rule which will allow new objects to be assigned to a class using only the information in the feature vectors of the new objects. In what follows, for simplicity, we shall assume there are only two classes, labelled 0 and 1, and then the classification rule is a mapping from the feature vector of each new object to the set of class labels $\{0, 1\}$.

Classifiers are most commonly constructed in two stages. The first stage is a mapping from the feature vectors to a univariate 'score' continuum. In this paper, we are concerned with the theoretical properties of performance measures of classification rules, so we shall assume that the score distributions for both of the classes are known, ignoring issues such as estimation and sampling errors. This means that we can transform the score to be the probability that the object belongs to class 1, in the sense that a proportion s of objects with transformed score s belong to class 1. Henceforth in this paper we shall assume the score has been transformed in this way. This means that all the integrals in what follows are over the interval [0,1]. Note that different classifiers map different sets of objects to score s: by virtue of being different classifiers, they have different contours of the probability of belonging to class 1 over the feature space.

We will denote the distribution of the scores of class 0 objects by $f_0(s)$ and the distribution of the scores of class 1 objects by $f_1(s)$, with corresponding cumulative distribution functions $F_0(s)$ and $F_1(s)$, and with relative class sizes (priors, in the statistical classification literature) π_0 and $\pi_1 = 1 - \pi_0$ respectively.

Since the score, s, associated with an object is the probability that object will belong to class 1, classification is achieved by comparing that probability with some threshold probability, t. Objects with probability greater than t of belonging to class 1 are assigned to class 1, and otherwise to class 0. Note that only the sign of $(s - t)$ matters here, not its magnitude: objects are assigned to class 1 whether their scores are just larger than t or much larger.

In general, for most natural problems, information in the features does not allow the classes to be perfectly separated. This means, in particular, that the score distributions $f_0(s)$ and $f_1(s)$ have overlapping support; that there is no threshold t for which all the class 1 scores are larger than t and all class 0 scores smaller than t. One implication is that whatever value we choose for the classification threshold t, some objects will have scores on the wrong side of the threshold, and will hence be misclassified. This means that it is necessary to devise some way of measuring the degree to which the classification rule fails to perfectly classify objects; some way of measuring the effectiveness of classification rules.

A large number of measures have been proposed. They fall into two broad classes: those for which the value of the classification threshold is given, and those for which it is not. Clearly this threshold must be specified by the time the classifier is used, but it is

* Corresponding author. Tel.: +44 (0) 20 7594 2752; fax: +44 (0) 20 7594 8517.
E-mail addresses: d.j.hand@imperial.ac.uk (D.J. Hand), canagnos@imperial.ac.uk (C. Anagnostopoulos).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.12.004

often the case that it has not been given at the time the classifier is being evaluated. For example, the intention might be to use the classifier in medical screening in different populations or clinics, which could require different thresholds, or in credit scoring under different economic circumstances which might again require different thresholds. In such cases, the classifier must be evaluated before the threshold has been chosen. For completeness, we should parenthetically remark here that sometimes a third class is defined, based on the accuracy of the classifier's estimates of the probability that an object will belong to class 1. However, measures of this last kind are not strictly measures of *classification accuracy* (see, for example, Friedman, 1997).

When the classification threshold is specified and known beforehand, performance is based on the (possibly weighted) counts in the four cells of the table of cross-classifications of the true class by the predicted class – and many measures have been based on this table, including recall (or sensitivity), specificity, error rate, proportion correctly classified, precision, the *F*-measure, and others (see Hand, 2012).

When the classification threshold, t, has not been given at the time that the performance has to be estimated, things are more difficult. One common strategy is to implicitly aggregate over a distribution of possible values for t, producing a portmanteau measure. The distribution indicates how likely we believe it is that each value of t will be used when the classifier is used in practice, and the aggregate measure is an average performance over possible t values. Again there are multiple ways to do this, but this paper focuses on one particularly popular such measure, the area under the receiver operating characteristic (ROC) curve. This curve shows a plot of $F_0(s)$ on the vertical axis against $F_1(s)$ on the horizontal axis (see Krzanowski and Hand, 2009 for a detailed discussion). The area under this curve, which we denote AUC, is $\int F_0(s)dF_1(s)$. The *Gini* coefficient is sometimes used as an alternative to the AUC. The Gini coefficient is a linear transformation of the AUC standardised so that chance classification accuracy has a score of 0.

As we illustrate below, the AUC is a particularly widely used measure of classifier performance. However, Hand (2009, 2010) (and see also Hilden, 1991) showed that when interpreted in terms of a balance of the relative costs of the two kinds of misclassification, the AUC is incoherent – in the sense that it requires that the relative costs of the two kinds of misclassification differ from classifier to classifier. Briefly, the argument in those papers is as follows. It begins from the premise (which this paper relaxes) that an appropriate choice for threshold should be based on balancing the cost due to misclassifying class 1 objects against the cost due to misclassifying class 0 objects. This balance of costs cannot be a function of the classifiers used, but is an external property of the problem: the relative severity of misdiagnosing someone with stomach cancer as having indigestion, compared with the reverse kind of misdiagnosis, is the same whether one uses logistic regression or a tree-based classifier to make the diagnosis. Then, given the score distributions of the two classes, there is a mapping between the ratio of these costs and the classification threshold which minimises the overall misclassification loss: given the cost ratio, this mapping specifies the optimal threshold, in the sense that it leads to minimum overall loss. Unfortunately, often the cost ratio cannot be specified beforehand. One strategy for overcoming this is to propose a distribution of likely values for this cost ratio, and calculate a mean classification loss, integrating over this distribution. This distribution of cost ratios corresponds to a distribution of optimal classification threshold values, via the mapping just mentioned. The AUC is equivalent to calculating this mean classification loss, integrating over the distribution of optimal classification threshold values. Unfortunately, since the mapping depends on the classifier, the distribution used in calculating the mean also depends on the classifier – and is equivalent to using a different distribution over the cost ratio for different classifiers. As we saw above, however, this is inappropriate: the cost ratio distribution should be the same for all classifiers.

Furthermore, as is clear from the definition and as has been explored by various authors (e.g., McLish, 1989; Walter, 2005; Wieand et al., 1989), removing the dependence between the relative misclassification costs and the classification threshold t means that the AUC is equivalent to assuming a uniform distribution over the proportion classified as belonging to class 1. This would seem to be an inappropriate assumption for real applications. These conclusions are important because they mean that the AUC can yield misleading inferences about the relative merits of different classification rules, and this matters because of the popularity of the AUC.

The present paper takes this discussion further, by demonstrating that the incoherence is not dependent on the cost-based argument, but is much more fundamental. Thus, in the next section we relate the AUC to the proportion of objects correctly classified, and show how the incoherence arises separately from cost considerations. In Section 3 we describe a related but distinct problem. Section 4 discusses the implications. First, however, in the remainder of this section, we report the results of a literature search on the use of the AUC as a measure of classifier performance. The numerical results reported below are only estimates (doubtless we missed some papers) but we took care to ensure that 'area under

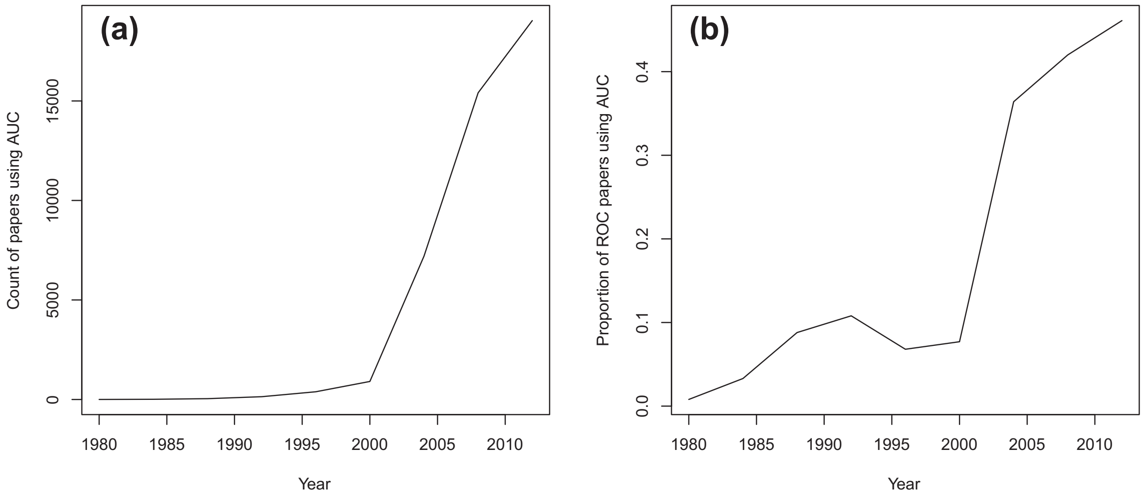

Fig. 1. In three year intervals (a) number of papers using the area under the curve to summarise the receiver operating characteristic curve, and (b) proportion of papers using receiver operating characteristic curves which use area under the curve as the summary measure.

the curve' referred to receiver operating characteristic curves (and not, for example, to concentration curves or other meanings). In general, when in doubt, we erred on the side of caution.

Fig. 1a shows the number of scientific papers which used the area under the curve as a classification rule performance measure for each three year interval from 1981 to 2012 (where the last value is an estimate, scaled by the rate of publication in the interval to the end of 2011). The rate is currently over 6000 usages per year. Fig. 1b shows the proportion of papers using receiver operating characteristic curves which used the AUC to summarise the curves. Currently about half of papers using receiver operating characteristic curves use AUC as a summary measure.

It is apparent from these figures that the AUC is a critically important measure of the performance of classification rules. What is not apparent from the figures is the wide range of application domains in which the AUC is used, including medicine, banking and credit risk assessment, web applications such as spam detection, fraud detection, and a host of others. Furthermore, the figures refer only to academic publications: they do not indicate the importance of the measure in commercial applications. For example, the Gini coefficient, mentioned above, is the most widely used performance measure for evaluating scorecards in the retail banking sector in the UK and the EU, and the extent of this usage is not demonstrated by the academic figures.

2. The AUC as an average of correct classifications

When the classification threshold t is unknown, one way we might evaluate a classifier is to use the expected proportion of cases it correctly classifies, averaging over some distribution for the possible values of $t, g(t)$. Since the score scale has been transformed to be the probability of belonging to class 1, $g(t)$ should be the same distribution for all classifiers we are evaluating (for a particular problem). For example, it would be irrational to say that, if logistic regression were to be used, then we would be very likely to choose probability 0.9 as our classification threshold, whereas if a random forests classifier were to be used we would be very *unlikely* to choose probability threshold 0.9. The choice of probability threshold we think is likely to be adopted when the classifier is applied in practice is a property of the problem, not the classifier.

The AUC is defined as $\int F_0(t)f_1(t)dt$, the average proportion of class 0 objects which are correctly classified if the threshold is randomly drawn from the class 1 score distribution $f_1(t)$. Using the properties $\int F_0(t)f_0(t)dt = \int F_1(t)f_1(t)dt = 1/2$ and $\int F_1(t)f_0(t)dt = 1 - \text{AUC}$ it is easy to show that $\int M(F_0(t), F_1(t))m(f_0(t), f_1(t))dt$ is a linear function of the AUC whenever M is a linear function of the probability distribution functions of the classes and m is a linear function of the probability density functions of the classes.

In particular, for example, let us take M to be the proportion of cases correctly classified, $\pi_0 F_0(t) + \pi_1(1 - F_1(t))$, and m the overall population score mixture density, $\pi_0 f_0(t) + \pi_1 f_1(t)$. These particular choices for M and m show that the AUC is a linear transformation (with coefficients which depend only on the class sizes) of the expected proportion of cases correctly classified, $\int M(t)m(t)dt = \pi_0^2/2 + \pi_1^2/2 + 2\pi_1\pi_0$ AUC. But the distribution over which this expectation is taken is $\pi_0 f_0(t) + \pi_1 f_1(t)$, which varies from classifier to classifier, and we have already seen that this distribution, called $g(t)$ above, should be the same for all classifiers. To let $g(t)$ differ between the different classifiers being compared is irrational. As noted elsewhere, it is analogous to comparing different classifiers on the basis of error rate, but using different threshold probabilities for different classifiers.

It is clear from the above, that the problem can be resolved by choosing the distribution $g(t)$ to be independent of the empirical score distributions of the classifiers. This idea is developed in (Hand and Anagnostopoulos, 2012; Hand, 2009, 2010). *But it does not lead to the AUC.*

3. Screening: a different problem

The criticism of the AUC as a measure of classifier performance presented above assumed that the only information to be used in assigning an object to a class was the score of the object (the probability that the object belongs to class 1) and the threshold probability. In that case, the AUC is an incoherent measure of classifier performance in the sense that the relative probabilities given to different choices of the threshold probability varied from classifier to classifier.

Sometimes, however, situations arise in which there are external constraints on the classification problem, and then using different threshold probabilities can be a sensible thing to do. By definition, an external constraint means that extra information is being used, in addition to the probability of class membership of the object to be classified. An example is a situation in which we will choose a classifier which optimises some performance measure subject to the constraint that a certain proportion, P, of all objects are classified into class 1. Examples of situations where such a procedure would be appropriate are medical screening or fraud detection in which resources are limited, so that one can afford to investigate closely only a certain number of people or transactions (here assuming 'number' to be equivalent to a 'proportion' of the presenting objects, for simplicity assuming a known population size). This rule is achieved by adopting that classification threshold t such that a proportion P of the objects have scores larger than t.

Since different classifiers have different score distributions, it is likely that the threshold value t for which a proportion P of the objects have higher scores will vary from classifier to classifier. That is, it would be perfectly reasonable in such a situation for an object which had a probability p of belonging to class 1 to be assigned to class 0 by one classifier and class 1 by another, even if they agreed on the value p. This new situation contrasts with the situation described in Section 2 because in that section objects are assigned to classes solely on the basis of their estimated probability of belonging to class 1. In particular, in Section 2, if the threshold is 0.9, say, then an object with estimated class 1 probability of 0.91 will be assigned to class 1 regardless of how other objects are classified. In the screening case of this section, however, an object with estimated class 1 probability of 0.91 may or may not be assigned to class 1, depending on how other objects are classified. If a proportion of objects greater than P have estimated class 1 probabilities greater than 0.91 then this new object will not be assigned to class 1. Thus in the screening case, the class to which a given object is assigned depends on the scores given to other objects. In the screening case, using the AUC is equivalent to calculating a mean performance, over a distribution of threshold values, just as in Section 2, but now the distribution does not depend on the score distributions of the classes. This independence means that the AUC is coherent when used in this way – the same measuring instrument is used for evaluating all classifiers. The threshold distribution will now be chosen on the basis of one's belief about the values of P which are relevant to the problem. In terms of Section 2, the m function is independent of f_0 and f_1.

Of course, if P were to be known beforehand, then it would be equivalent to knowing the classification threshold for each classifier, and then, as described in Section 1, count-based measures of performance should be used. Portmanteau measures such as the AUC are only relevant if the threshold cannot be specified. That is, they are relevant in this situation only if we know that some proportion P of the objects will be assigned to class 1, but we do not (yet) know what P will be.

4. Discussion

The evaluation of classification performance is central to the construction and selection of classification rules. Because of its importance, a huge literature has accumulated, spanning data analytic domains including statistics, pattern recognition, machine learning, and data mining, and also application areas such as medicine, credit risk, speech recognition, signal detection, fault identification, and target identification. Apart from the many papers, several books have also been written on this topic (e.g., Gönen, 2007; Hand, 1997; Krzanowski and Hand, 2009; Pepe, 2003; Zhou et al., 2002, amongst others) so that a full review of this literature would be impossible. However, a few measures are particularly popular. These include the misclassification or error rate (Jamain, 2004 claims that the vast majority of comparative studies of classification rules report error rate), the Kolmogorov Smirnov statistic, and the area under the receiver operating characteristic curve (or, equivalently, the Gini coefficient). This paper focuses on the last of these measures.

The area under the receiver operating characteristic curve, $\text{AUC} = \int F_0(s)dF_1(s)$, is a popular measure for evaluating and comparing classification rules when one cannot decide a priori what classification threshold will be used. It is a measure of the separability of the score distributions of the two classes. However, for many, perhaps most, classification problems it is incoherent in the way it treats different classifiers. In particular, given that a classifier is to assign all objects with class 1 probability greater than some threshold probability t to class 1, and all others to class 0, using the area under the curve is equivalent to supposing that one's belief about the likely values of the probability t varies from classifier to classifier. In many classification situations this is irrational: the classification of objects should depend on their probability of belonging to class 1, and not on how that probability was estimated.

This core incoherence of the area under the ROC curve as a measure of classifier performance has been explored in (Hand, 2009, 2010) (and our attention has recently been drawn to Hilden, 1991, for an exploration closely related to that in (Hand, 2009)), in terms of the relative costs of misclassifying class 0 points into class 1, and vice versa. Those papers show that the area under the curve is equivalent to assuming different cost distributions for different classifiers. This is generally inappropriate: the distribution of relative misclassification costs should be a function of the problem, and not of the instrument used to make the classification. As (Hand, 2009) puts it, using the area under the ROC curve is equivalent to evaluating different classifiers using different metrics, and a fundamental tenet of comparative evaluation is that one uses the same measuring instrument on the things being compared: I do not measure your 'size' using a weighing scale calibrated in grams, and mine using a metre rule calibrated in centimetres, and assert that you are 'larger' because your number is greater. The contribution of this present paper is to sidestep the need to refer to relative misclassification costs, showing that the problem is a fundamental one.

The distinction between the two strategies for choosing the threshold distribution described in Sections 2 and 3 has been made elsewhere, though as far as we know comparative discussions follow the cost-based argument. Examples of such papers are Hand (2010) and Flach et al. (2011). Flach et al. (2011) choose the thresholds independently of the costs, from the population score mixture distribution, as described in Section 3. For example, Flach et al. (2011) say (their Section 3) 'we uniformly select an instance x, and set the threshold to the score of that instance' – and this is the choice implicit in the AUC. To us this choice seems unlikely to be useful in most practical situations. In a screening application, for example, it would be equivalent to saying that one thought it equally likely that one would want to screen out hardly any of the population or almost all of the population for subsequent close examination. So, when used in this way, the AUC is coherent, but would appear to be inappropriate. Recognition of this inappropriateness has motivated the development of alternatives to the AUC, such as the partial area under the curve (e.g., McLish, 1989; Walter, 2005), measures based on other distributions on the proportions classified into class 1 (e.g., Wieand et al., 1989), and measures based on specifying the distribution of relative severities of the misclassification costs (e.g., Hand, 2009, 2010). We are grateful to a referee for pointing out that, for the first of these three cases, the partial AUC may be coherent or incoherent, depending on how the interval over which the partial AUC is evaluated is chosen, and on the way the classification threshold is chosen within the interval.

In summary, the AUC would appear to be either incoherent, requiring different probability threshold distributions for different classifiers, or inappropriate, assuming a uniform distribution over the proportion of (for example) class 1 which are classified as class 1, and we recommend that an alternative measure which overcomes these shortcomings is adopted in place of the AUC. One such alternative measure, the H measure, is described in (Hand, 2009, 2010), and an improved version thereof in (Hand and Anagnostopoulos, 2012). Papers (Hand, 2009, 2010) also give numerical comparisons between the AUC and this alternative measure, as does the package vignette for the respective R package, available from the CRAN repository (see http://www.hmeasure.net for more details).

Acknowledgement

We would like to express our appreciation to Peter Flach for a stimulating discussion about the AUC and its properties, which helped us to refine this paper. We would also like to thank the anonymous referees for helpful suggestions which, we believe, improved the clarity of the paper.

References

Flach, P., Hernández-Orallo, J., Ferri, C., 2011. A coherent interpretation of the AUC as a measure of aggregated classification performance. In: Proceedings of 28th International Conference on Machine Learning, Bellevue, WA, pp. 657–664.

Friedman, J., 1997. On bias, variance, 0/1-loss, and the curse of dimensionality. Data Mining Knowledge Discovery 1, 55–77.

Gönen, M., 2007. Analyzing Receiver Operating Characteristic Curves with SAS. Technical Report. SAS Institute, Cary, NC.

Hand, D.J., Anagnostopoulos, C., 2012. A better Beta for the H measure of classification performance. arXiv:1202.2564v1, http://arxiv.org/abs/1202.2564v1

Hand, D., 1997. Construction and Assessment of Classification Rules. Wiley, Chichester.

Hand, D., 2009. Measuring classifier performance: A coherent alternative to the area under the ROC curve. Mach. Learn. 77, 103–123.

Hand, D., 2010. Evaluating diagnostic tests: The area under the ROC curve and the balance of errors. Stat. Med. 29, 1502–1510.

Hand, D.J., 2012. Assessing the performance of classification methods. Int. Stat. Rev. 80, 400–414.

Hilden, J., 1991. The area under the ROC curve and its competitors. Med. Decis. Making 11, 95–101.

Jamain, A., 2004. A meta-analysis of classification methods, Ph.D. Thesis. Department of Mathematics, Imperial College, London.

Krzanowski, W., Hand, D., 2009. ROC Curves for Continuous Data. Chapman and Hall.

McLish, D., 1989. Analyzing a portion of the ROC curve. Med. Decis. Making 9, 190–195.

Pepe, M., 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, Oxford.

Walter, S., 2005. The partial area under the summary ROC curve. Stat. Med. 24, 2025–2040.

Wieand, S., Gail, M., James, B., James, K., 1989. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika 76, 585–592.

Zhou, X.H., Obuchowski, N., McClish, D., 2002. Statistical Methods in Diagnostic Medicine. Wiley, New York.