Rate-Constrained Ranking and the Rate-Weighted AUC

Louise A.C. Millard[1,2,3], Peter A. Flach[1,3], and Julian P.T. Higgins[2,4]

[1] Intelligent Systems Laboratory, University of Bristol, United Kingdom
[2] School of Social and Community Medicine, University of Bristol, United Kingdom
[3] MRC Integrative Epidemiology Unit, University of Bristol, United Kingdom
[4] Centre for Reviews and Dissemination, University of York, York, United Kingdom
{Louise.Millard,Peter.Flach,Julian.Higgins}@bristol.ac.uk

Abstract. Ranking tasks, where instances are ranked by a predicted score, are common in machine learning. Often only a proportion of the instances in the ranking can be processed, and this quantity, the predicted positive rate (PPR), may not be known precisely. In this situation, the evaluation of a model's performance needs to account for these imprecise constraints on the PPR, but existing metrics such as the area under the ROC curve (AUC) and early retrieval metrics such as normalised discounted cumulative gain (NDCG) cannot do this. In this paper we introduce a novel metric, the rate-weighted AUC (rAUC), to evaluate ranking models when constraints across the PPR exist, and provide an efficient algorithm to estimate the rAUC using an empirical ROC curve. Our experiments show that rAUC, AUC and NDCG often select different models. We demonstrate the usefulness of rAUC on a practical application: ranking articles for rapid reviews in epidemiology.

1 Introduction and Motivation

The work reported in this paper was motivated by the task of undertaking rapid reviews of clinical trials. A rapid review should follow the broad principles of a systematic review, where a medical research question is asked (such as the effect of a drug on a disease) and the evidence from all relevant research articles is compiled to give a better estimate of the drugs effect than each individual study provides. However, a rapid review needs to be performed under strict time and resource constraints, so it may not be possible to review all relevant articles. Currently, a rapid review is performed by human reviewers who search online medical research databases for articles reporting clinical trials of a particular research question [7]. In order to retrieve a set of articles that can be reviewed in the allotted time, the reviewer may iteratively refine the search query until the number of articles is deemed manageable.

An important consideration when performing a rapid review is the quality of each study. Low-quality studies are more likely to give a biased estimate of the research question and may need to be excluded from the review or considered with caution [9]. Therefore, the aim of a rapid review can be described as maximising the number of high-quality articles assessed, given the particular time

T. Calders et al. (Eds.): ECML PKDD 2014, Part II, LNCS 8725, pp. 386–403, 2014.

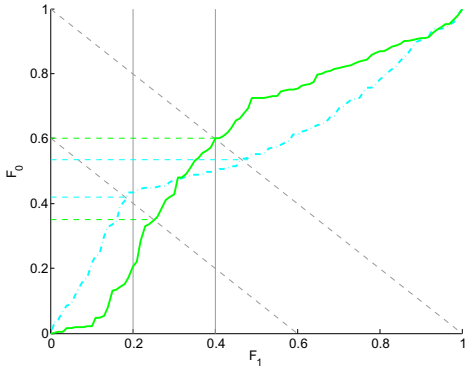


Fig. 1. Two hypothetical ROC curves (x-axis: false positive rate, y-axis: true positive rate), example PPR isometrics (diagonal lines with $PPR = 0.5$ (top) and $PPR = 0.3$ (bottom); the slope of -1 indicates a uniform class distribution) and example partial AUC bounds (vertical lines)

constraints of the review. The iterative search method described above is a rather crude approach that does not consider article quality, and can be thought of as a classification of articles as *included* or *excluded* from the review. We suggest that this can be greatly improved by instead learning a model for estimating the article's study quality, and using the model's scores to rank the studies under review, such that the most reliable research is assessed first. The reviewers can then simply review the articles in decreasing order of estimated quality until they run out of time. There is no need to specify a classification threshold.

This approach suggests that a good model is one that exhibits good ranking behaviour with respect to study quality, with particular emphasis on the proportion of articles that can reasonably be processed. The total amount of time available for a review and the number of articles returned from the initial search query is typically known. Given an estimate of the time it will take a reviewer to assess a single article, the proportion of articles in the search results that is expected to be processed can be inferred. In terms of binary classification this proportion is the predicted positive rate (PPR). If the PPR is known precisely, finding the best model is straight-forward. Figure 1 illustrates this with two hypothetical ROC curves where neither curve dominates the other. The two dashed lines show two example PPR values that could be inferred for a rapid review. We can see that the PPR value affects which model is chosen. The (solid) green model is chosen when $PPR = 0.5$ and the (dashed) blue model is chosen when $PPR = 0.3$, as these models have the highest recall at these respective points on the ROC curves.

However, the PPR inferred depends on the time needed to review a single article, and this is not known precisely. Articles vary in length and difficulty and hence it is only possible to estimate a probability distribution across the

PPR, rather than specify a single value. Therefore, an appropriate measure of rate-constrained ranking would average the true positive rate across each value of PPR, weighted by its probability. In this paper we develop such a measure. In addition to our motivating example, there are many other tasks that are rate constrained, with uncertainty across the rates. In general, these tasks are restricted to a fixed budget of a resource such as time or money, where the exact expenditure for each instance is not known precisely. Another example is telephone sales, which is restricted by the allocated number of person hours, such that when ranking a database of customers to determine those most likely to show interest, it is not known exactly how many customers will be contacted as the time per phone call is variable.

There are, of course, several existing metrics often used to evaluate ranking tasks. The area under the ROC curve (AUC), which estimates the probability that a random positive is ranked higher than a random negative, measures ranking performance across the entire ROC curve, treating all regions as equally important [5]. An alternative to the AUC, the partial AUC (pAUC), has previously been suggested [4]. This metric also weights uniformly, but restricts to a range of false positive rate (or true positive rate) values. For instance, the two solid vertical lines of Figure 1 show example pAUC bounds, constrained to false positive rates between 0.2 and 0.4. We require a metric that weights the area under the ROC curve with respect to the PPR, but the AUC weights the area uniformly and the pAUC can only weight across true positive or false positive rates, components of the PPR, and not the PPR itself.

Early retrieval tasks are those where examples near the top of the ranking are more important, as these examples are more likely to be processed. Several metrics in several fields have been proposed to address this problem, such as normalised discounted cumulative gain (NDCG) [10]. However, as we demonstrate later this metric and related ones assume that the likelihood of stopping at a particular position in the ranking is always higher nearer the top which is not necessarily the case when rates are constrained.

A key contribution of this paper is the derivation of a new metric, the rate-weighted AUC (rAUC), to evaluate models for rate-constrained ranking tasks (Section 3). We prove that the rAUC and rate-weighted expected recall are linearly related given a fixed class distribution. Furthermore, we provide an efficient algorithm to estimate the rAUC using an empirical ROC curve (Section 4). Finally, we demonstrate that given rate constraints the rAUC chooses the optimal model while the AUC and NDCG metrics often choose a suboptimal model (Section 5).

2 Notation and Basic Definitions

We follow the notation of [8]. We assume a two-class classification problem with instance space $\mathcal{X}$. The positive and negative classes are denoted by 0 and 1, respectively. The learner outputs a score $s(x) \in [0, 1]$ for each instance $x \in \mathcal{X}$, such that higher scores express a stronger belief that x belongs to class 1.

The score densities and cumulative distributions are denoted by f_k and F_k for class $k \in \{0, 1\}$. Given a threshold at score t the true positive rate (also called sensitivity or positive recall) is $P(s(x) \leqslant t|k = 0) = F_0(t)$ and the false positive rate is $P(s(x) \leqslant t|k = 1) = F_1(t)$. The true negative rate, also called specificity or negative recall, is $1 - F_1(t)$.

The proportions of positives and negatives are denoted by π_0 and π_1 respectively. Accuracy acc at threshold t is a weighted average of positive and negative recall:

$$acc(t) = \pi_0 F_0(t) + \pi_1(1 - F_1(t)) \tag{1}$$

Similarly, the proportion of positive predictions at threshold t (the predicted positive rate) is a weighted average of the true and false positive rates:

$$r(t) = \pi_0 F_0(t) + \pi_1 F_1(t) \tag{2}$$

This is the predicted positive rate, which we abbreviate to the rate.

A ROC curve is a plot of true positive rate on the y-axis against false positive rate on the x-axis. The area under the ROC curve (AUC) is the true positive rate averaged over all false positive rates:

$$AUC = \int_0^1 F_0 dF_1 = \int_{-\infty}^{+\infty} F_0(t) f_1(t) dt \tag{3}$$

Alternative parameterisations are possible; in this paper we are particularly interested in a parametrisation by rate.

Metrics such as predicted positive rate can be depicted in ROC space using isometrics – points on ROC space that have the same value for a given metric [6]. For instance, several combinations of false and true positive values result in the same rate (Equation 2), and this can be shown as a straight line drawn in ROC space.

3 The Rate-Weighted AUC

The aim of a rate-constrained ranking task is to maximise the expected true positive rate given a probability distribution across the rates. Common formulations of the AUC are given as an expectation of the true positive rate across all false positive rates or the thresholds (Equation 3). It is not possible to apply a weight across rates using these formulations, because they are given in terms of expectations over F_1 and t, rather than the rate. The following section derives the AUC as an expectation across rates, such that the derived formula can be altered to weight the AUC with respect to the rate.

Accuracy isometrics in ROC space are lines of constant accuracy with slope π_1/π_0 [6]. Similarly, rate isometrics are lines of constant rate with slope $-\pi_1/\pi_0$. Examples are shown in Figures 2a and Figure 2c for uniform and non-uniform class distributions, respectively.

Definition 1. Rate-accuracy space is a plot of rate on the x-axis and accuracy on the y-axis. Rate-recall space is a plot of rate on the x-axis and recall on the y-axis. Where positive recall is used, rate-recall space is denoted rate-$F_0(r)$ space. Where negative recall is used, rate-recall space is denoted rate-$(1-F_1(r))$ space.

We translate the ROC curve to rate-accuracy and rate-recall spaces using a linear transformation, such that the AUC can be calculated in this space instead. The ROC curve of Figure 2a is transformed into the rate-accuracy curve shown in Figure 2b, and the rate-recall curves shown in Figures 2e and 2f, for positive and negative recall respectively. We can see that the transformations into rate-accuracy and rate-recall spaces result in unreachable areas. The upper bounds of the rate-accuracy and rate-recall curves correspond to the ROC curve of a perfect classifier, and the lower bounds to that of a pessimal classifier.

Definition 2. The lower bounds in x-y space are given by a function $f_{min}(x)$ specifying the minimum possible value of y at each value of x. The upper bounds in x-y space are given by a function $f_{max}(x)$ specifying the maximum possible value of y at each value of x.

We now focus on rate-accuracy space, but a similar derivation can be given for rate-recall space (given in Theorem 5). In rate-accuracy space, the lower and upper bounds of accuracy at rate r are given by:

$$acc_{min}(r) = |\pi_1 - r| \qquad acc_{max}(r) = 1 - |\pi_0 - r| \tag{4}$$

These are derived from Equation 1 and the fact that acc_{min} corresponds to points with $F_0 = 0$ when $r \leqslant \pi_1$ and points with $F_1 = 1$ when $r \geqslant \pi_1$, and acc_{max} corresponds to points with $F_1 = 0$ when $r \leqslant \pi_0$ and points with $F_0 = 1$ when $r \geqslant \pi_0$.

Clearly, a ROC curve can only cross each rate isometric at a single point, which allows us to reformulate the AUC in terms of accuracy and rates in order to apply a weight across rates. Accuracy difference acc_{dif} is the difference in the accuracy value of the ROC curve with the minimum possible accuracy value for a given rate:

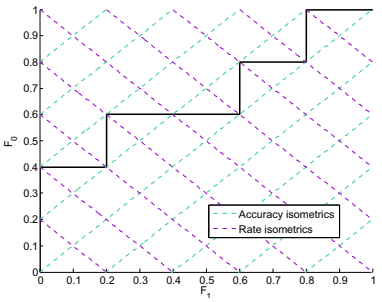
$$acc_{dif}(r) = acc(r) - acc_{min}(r) \tag{5}$$

Theorem 3. *The AUC is equal to the normalised accuracy difference across all rates $r \in [0,1]$:*

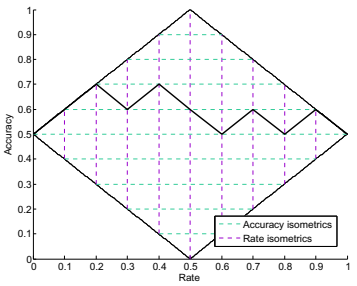
$$AUC = \frac{1}{K_{acc}} \int_0^1 acc_{dif}(r)dr \tag{6}$$

where K_{acc} is constant for a fixed class distribution:

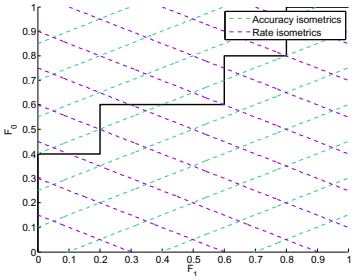
$$K_{acc} = \int_0^1 (acc_{max}(r) - acc_{min}(r))\, dr \tag{7}$$



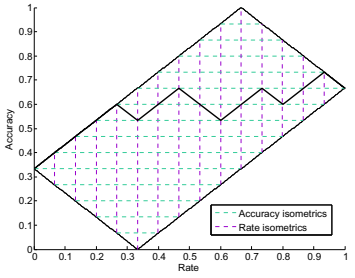
(a) Example ROC curve with rate and accuracy isometrics for $\pi_0 = \pi_1 = \frac{1}{2}$.



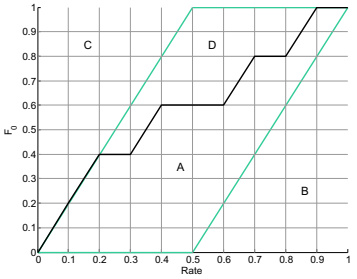
(b) Rate-accuracy curve corresponding to ROC curve shown in Figure 2a.



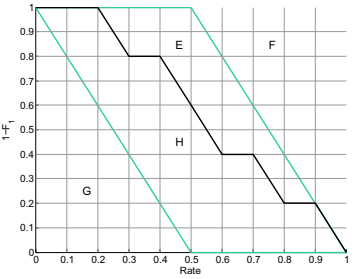
(c) Example ROC curve with rate and accuracy isometrics for $\pi_0 = \frac{2}{3}$, $\pi_1 = \frac{1}{3}$.



(d) Rate-accuracy curve corresponding to ROC curve shown in Figure 2c.



(e) Rate-recall curve for the positive class of ROC curve shown in Figure 2a.



(f) Rate-recall curve for the negative class, of ROC curve shown in Figure 2a.

Fig. 2. Example ROC curves, rate-accuracy curves and rate-recall curves

Theorem 3 holds as transforming a ROC curve from ROC to rate-accuracy space requires only linear transformations such that the relative areas under and above the curve within the transformed bounds of the original ROC space remains the same. This reformulation of AUC in terms of rates allows us to introduce a rate-constrained generalisation.

Definition 4. The rate-weighted AUC of a ROC curve is the AUC weighted across the rates:

$$rAUC = \frac{1}{K_{acc,w(r)}} \int_0^1 w(r) acc_{dif}(r) dr \tag{8}$$

where $w(r)$ is a density over the rate and $K_{acc,w(r)}$ is given by:

$$K_{acc,w(r)} = \int_0^1 w(r) \left(acc_{max}(r) - acc_{min}(r)\right) dr \tag{9}$$

Theorem 5. *The rAUC is equal to the normalised F_0 difference weighted across all rates. With a slight abuse of notation we use $F_k(r)$ to mean $F_k(F^{-1}(r))$.*

$$rAUC = \frac{1}{K_{F_0,w(r)}} \int_0^1 w(r) \left(F_0(r) - F_{0,min}(r)\right) dr \tag{10}$$

where

$$K_{F_0,w(r)} = \int_0^1 w(r) \left(F_{0,max}(r) - F_{0,min}(r)\right) dr \tag{11}$$

and $F_{0,min}(r) = \max\left(0, \frac{r-\pi_1}{\pi_0}\right)$, $F_{0,max}(r) = \min\left(1, \frac{r}{\pi_0}\right)$.

Clearly, we can derive an analogous result using negative recall $(1 - F_1(r))$ instead of positive recall $(F_0(r))$. The area under the rate-recall curve is the expected recall (positive or negative) given a uniform distribution across the rates. This makes the formulation of the rAUC in rate-recall space particularly interesting, as we can infer the relationship between $\mathbb{E}[F_0]$ – the quantity we intend to maximise in rate-constrained ranking – and the rAUC.

Rate-recall space, as shown in Figures 2e and 2f can be divided into 4 distinct regions, for both positive and negative recall (labelled A-D and E-H respectively). We use A both to label the area A and as the mass of this area.

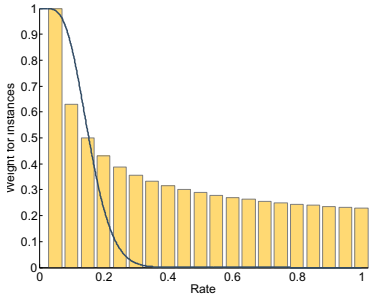
Theorem 6. *The rate-weighted expected true positive rate is related to the rAUC, given a distribution over the rates, by:*

$$\mathbb{E}[F_0] = (1 - B - C) rAUC + B \tag{12}$$

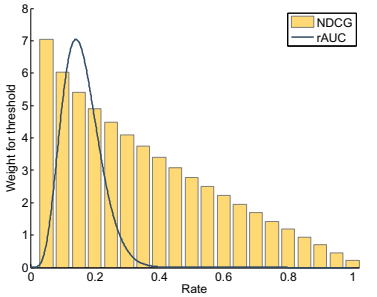
where $C = \int_0^{\pi_0} w(r) \left[\frac{\pi_0 - r}{\pi_0}\right] dr$ *and* $B = \int_{\pi_1}^1 w(r) \frac{r-\pi_1}{\pi_0} dr$.

Proof. Rate-F_0 space is bounded by $r = 0$, $r = 1$, $F_0 = 0$ and $F_0 = 1$, such that the total weighted mass of this area $\int_0^1 w(r) dr = 1$, hence $A + B + C + D = 1$. As $rAUC = \frac{A}{A+D}$, it follows that:

$$\mathbb{E}[F_0] = \frac{A + B}{A + B + C + D} = A + B = rAUC(A + D) + B = (1 - B - C) rAUC + B \tag{13}$$



(a) Weights across instances, representing the likelihood an instance will be processed.



(b) Weights across thresholds, representing the likelihood a rate will be the threshold position (the instance at this rate will be the last to be processed).

Fig. 3. NDCG discrete weights (using log base 2) assuming 20 instances and rAUC continuous weights using beta distribution ($\alpha = 6.23$, $\beta = 32.80$). Weights across instances in left figure are equivalent to weights across thresholds in right figure, respectively.

Area C is the triangular region bounded by the lines $r = 0$, $F_0 = 1$ and $F_0 = \frac{r}{\pi_0}$. The weighted mass of C is given by:

$$C = \int_0^{\pi_0} w(r) \frac{\pi_0 - r}{\pi_0} dr \tag{14}$$

Area B is the triangular region bounded by the lines $r = 1$, $F_0 = 1$ and $\frac{r-\pi_1}{\pi_0}$. The weighted mass of B is given by:

$$B = \int_{\pi_1}^{1} w(r) \frac{r - \pi_1}{\pi_0} dr \tag{15}$$

□

This completes the proof.

B and C depend only on the class and weight distributions, which implies that the relationship between $\mathbb{E}[F_0]$ and rAUC depends only on these and not the shape of the ROC curve. Therefore, maximising $\mathbb{E}[F_0]$ is equivalent to maximising $\mathbb{E}[rAUC]$, which means that rAUC is a suitable metric to evaluate models for rate-constrained ranking.

3.1 Comparing the Weights of NDCG and rAUC

Normalised discounted cumulative gain (NDCG) is given by:

$$NDCG = \frac{1}{K} \cdot \sum_{i=1}^{n} \frac{1}{log_b(i+1)} rel_i \tag{16}$$

where $rel_i \in [0, 1]$ is the label of example at rank i, which can be continuous or binary and denotes the relevance of the example. K is the maximum possible DCG for a ranking of size n: $K = \sum_{i=1}^{n} 1/log_b(i+1)$.

NDCG weights each point in the ranking according to the probability that this instance will be processed. In contrast, the rAUC weights each point in the ranking according to the probability this point will be the threshold index, such that processing will terminate at this point in the ranking. These formulations are closely related, since the probability that an instance at position i is processed is the probability that an instance at a position after i is the threshold index. For example, if a person is processing 20 articles, the probability they will review the article at rank position 10 equals the probability they will stop processing articles at a position between articles 10 and 20. Hence, the relationship between the two weighting methods is given by:

$$w_{instance}(i) = 1 - CDF_{w_{threshold}}(i) \tag{17}$$

where i is the position in the ranking and CDF denotes the cumulative distribution function.

The instance weights of NDCG are shown in Figure 3a, and the equivalent threshold weights are shown in Figure 3b. Here we use $rel_i = 1 - k_i$, where $k_i \in \{0, 1\}$. We can see that the weight of each threshold index decreases as we move further down the ranking. This is a key restriction of the NDCG (and related metrics), as it is not always the case that a ranking is more likely to be processed up to the rank positions nearer the top, as in our motivating example. Figure 3b also shows an example density across thresholds, using a beta distribution, where processing is most likely to stop at a rate of 14%. The corresponding instance weights are shown in Figure 3a. Note that by shifting the beta distribution to the right we will create a situation where a number of top-ranked instances will receive the highest weight, something which is not possible with NDCG.

4 Algorithm to Calculate the rAUC of an Empirical ROC Curve

We now use rate-accuracy space to compute the rAUC. A similar algorithm could be implemented in rate-recall space (of either positive or negative recall). Algorithm 1 estimates the rAUC from an empirical ROC curve, where the number of positive N^+ and negative N^- instances is known ($N = N^- + N^+$). This algorithm is similar to the standard AUC $O(N)$ algorithm [5] where the ROC space is processed one vertical (or horizontal) slice at a time. As can be seen in Figure 4, the area under the ROC curve in rate-accuracy space is composed of a series of vertical slices of width $\frac{1}{N}$, each corresponding to an instance. Ties triangles may also exist, each of which corresponds to a set of instances with the same score. The rAUC is calculated as a summation of the weighted mass of all vertical slices and ties triangles, normalised by the weighted mass of the whole rate-accuracy space. The algorithm we propose has four functions: $rAUC$,

$SAUC$, $VAUC$ and $TAUC$. The $rAUC$ function is the main function that iterates through the ranking of instances counting the number of positive and negative instances with the same score (which we call a ties section), and calling the $SAUC$ function when a new score is reached.

The $SAUC$, $VAUC$ and $TAUC$ functions calculate the mass of each ties section (which may consist of only one instance if it has a unique score). The $SAUC$ function simply calls the VAUC and TAUC functions. The negative instances are processed before the positives as when there is a ties triangle the shape of the area under this triangle in rate-accuracy space is given by the area of the negative instances, followed by the positive instances in this ties section. The $VAUC$ function computes the mass of a vertical slice of the area under the curve, using two equations depending whether the current instance is positive or negative. The accuracy difference equation is used, which is computed in terms of r and either F_0 or F_1 depending if the instance is negative or positive respectively (as for instance, if the instance is positive the value of F_1 stays constant). The $TAUC$ function computes the mass of the ties triangle (which is not shown

Algorithm 1. The rAUC algorithm. *scores*: list of scores of instances, in decreasing magnitude. x: list of class labels corresponding to the instances of *score*. N^+: number of positive instances. N^-: number of negative instances.

procedure RAUC($scores, x, N^+, N^-$)
  $\pi_0 \leftarrow N^+/(N^+ + N^-)$; $\pi_1 \leftarrow N^-/(N^+ + N^-)$; $N \leftarrow N^+ + N^-$
  $a_u \leftarrow 0$; $TP \leftarrow 0$; $FP \leftarrow 0$
  $N^+_{ties} \leftarrow 0$; $N^-_{ties} \leftarrow 0$; $score_{ties} \leftarrow -1$
  for $i = 1$ to N **do**
    if $score_{ties} = scores(i)$ **then**
      if x_i is POSITIVE **then**
        $N^+_{ties} \leftarrow N^+_{ties} + 1$
      else
        $N^-_{ties} \leftarrow N^-_{ties} + 1$
      end if
    else
      $[FP, TP, a_u] \leftarrow SAUC(a_u, N^-_{ties}, N^+_{ties}, FP, TP, N^-, N^+)$
      if x_i is POSITIVE **then**
        $N^+_{ties} \leftarrow 1$; $N^-_{ties} \leftarrow 0$
      else
        $N^+_{ties} \leftarrow 0$; $N^-_{ties} \leftarrow 1$
      end if
      $score_{ties} \leftarrow score(i)$
    end if
  end for
  $[FP, TP, a_u] \leftarrow SAUC(a_u, N^-_{ties}, N^+_{ties}, FP, TP, N^-, N^+)$
  $a \leftarrow K(w, \pi_0, \pi_1)$
  $rAUC \leftarrow \frac{a_u}{a}$
  Return $rAUC$
end procedure

Algorithm 1. (continued)

procedure SAUC(a_u, N^-_{ties}, N^+_{ties}, FP,TP, N^-, N^+)
 if $N^-_{ties} \geqslant 1$ **then**
  $FP_{prev} \leftarrow FP$; $FP \leftarrow FP + N^-_{ties}$
  $a_u \leftarrow a_u + VAUC(FP_{prev}, TP, FP, TP, N, 0)$
 end if
 if $N^+_{ties} \geqslant 1$ **then**
  $TP_{prev} \leftarrow TP$; $TP \leftarrow TP + N^+_{ties}$
  $a_u \leftarrow a_u + VAUC(FP, TP_{prev}, FP, TP, N, 1)$
 end if
 if $N^+_{ties} \geqslant 1$ & $N^-_{ties} \geqslant 1$ **then**
  $a_u \leftarrow a_u + TAUC(FP, TP, N^-_{ties}, N^+_{ties}, N^-, N^+)$
 end if
 Return $[FP, TP, a_u]$
end procedure

procedure VAUC($FP_{prev}, TP_{prev}, FP, TP, N, label$)
 $start \leftarrow FP_{prev} + TP_{prev}$, $end \leftarrow FP + TP$
 $f_1 \leftarrow \frac{FP_{prev}}{nTotalMinus}$; $f_0 \leftarrow \frac{TP_{prev},}{nTotalPlus}$
 for $i = start$ to end **do**
  if $label = 0$ **then**
   $FP_{prev} \leftarrow FP_{prev} + 1$; $f_1 \leftarrow \frac{FP_{prev}}{nTotalMinus}$
   $a_u \leftarrow a_u + \int_{\frac{i}{nTotal}}^{\frac{i+1}{nTotal}} (w(r)(2\pi_1 f_0 + \pi_1 - r - |r - \pi_1|)dr$
  else
   $TP_{prev}, \leftarrow TP_{prev}, +1$; $f_0 \leftarrow \frac{TP_{prev},}{nTotalPlus}$
   $a_u \leftarrow a_u + \int_{\frac{i}{nTotal}}^{\frac{i+1}{nTotal}} (w(r)\ (2\ (r\ -\ \pi_1\ f_1) +\ \pi_1\ -\ r\ -\ |\ r\ -\ \pi_1\ |\ )\ dr$
  end if
 end for
 Return a_u
end procedure

in Algorithm 1 due to space constraints). A ties triangle T is composed of 2 sub-triangles T_A and T_B where $T = T_A \cup T_B$. T_A and T_B adjoin on line H, where H is fixed along the rate isometric that passes through the right angled corner of T (see Figure 4).

We calculate the mass of a ties triangle by first finding the length H and the rate at each corner of T, labelled P_1, P_2 and P_3 in Figure 4. The weighted mass of the ties triangle is then the summation of T_A and T_B, which are given by:

$$T_A = \int_{r_1}^{r_2} w(r) \cdot H \cdot \frac{r - r_1}{r_2 - r_1} dr \qquad T_B = \int_{r_2}^{r_3} w(r) \cdot H \cdot \frac{1 - (r - r_2)}{r_3 - r_2} dr \tag{18}$$

where r_1, r_2 and r_3 are the rates at P_1, P_2 and P_3 respectively. The rAUC of a ROC curve is computed in $O(N)$ time. Algorithm 4 appears more lengthy compared to the standard AUC algorithm that is calculated in ROC space because each step across rate-accuracy space corresponds to a negative or positive

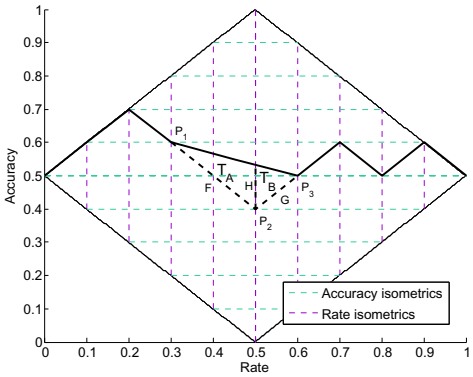


Fig. 4. Example rate-accuracy curve with a ties section (set of instances with the same score). Lengths and angles used to calculate rAUC are labelled.

instance and the height of the curve changes within this step.The standard AUC algorithm makes a step only when the instance is (for example) positive and (given this instance is not tied with another) the height of the ROC curve is constant within this step. The change in height at each step in rate-accuracy space means that the mass of the positive and negative vertical sections (and ties triangle) can only be calculated after the ties section has ended, hence the $SAUC$ function is needed to do this.

5 Experimental Evaluation

We used 5 UCI datasets (vote, autos, credit-g, breast-w and colic) to generate a set of models using 3 learning algorithms (naive Bayes, decision trees and one-rule). We chose a binary variable for each dataset as the label, and learnt 10 models with each dataset/model pair using bootstrap samples of 54% of the data, resulting in 150 generated models. We computed the AUC and NDCG metrics, and rAUC for each of these models, for 5 beta distributions with alpha and beta (α, β) values: $(3, 19)$, $(7, 15)$, $(11, 11)$, $(15, 7)$, and $(19, 3)$, shown in Figure 5. We use NDCG with log base 10.

Figure 6 shows the AUC and NDCG values, compared with the rAUC values, for each model. Each model is shown by 5 points with a single AUC / NDCG value and variable rAUC value (for each of the 5 rate distributions of Figure 5). The variance of the rAUC for each ROC curve across the 5 beta distributions ranges from 0 to 0.260 for these datasets. Spearman's rank correlations between the model rankings using each rate distribution are given in Table 1. The correlation of the rAUC with the AUC varied between 0.872 and 0.975, depending on the rate distribution. We should note that a proportion of the generated models have very high AUC values, and therefore very high rAUC values for most rate distributions (see Figure 6a). To correct for this inflation of the correlation values, Table 1 also shows reduced correlations when restricted to models

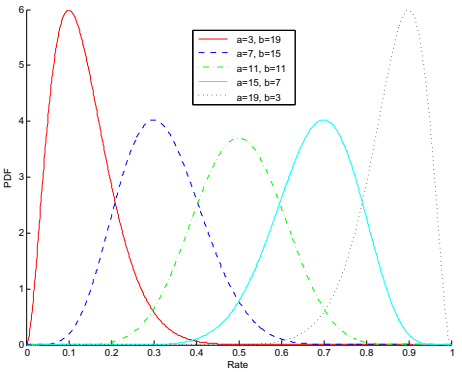


Fig. 5. Beta distributions across the rate

with $AUC \leqslant 0.95$. Correlations between the NDCG and rAUC metrics decrease dramatically when the mode of the beta distribution increases as expected.

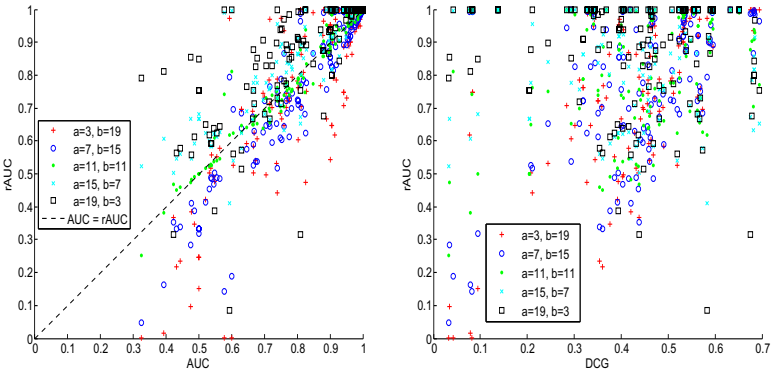
The correlation between rAUC metrics using rate distributions that weight different portions of ROC space is low in general. For instance, the rAUC values using rate distributions with $\alpha = 3$, $\beta = 19$ and $\alpha = 19$, $\beta = 3$ have a Spearman's rank correlation of 0.610. This highlights the importance of using a rate distribution with an appropriate degree of uncertainty, as if it is incongruous with the true probability distribution a suboptimal model may be chosen.

5.1 Application to Screening for Rapid Reviews

We demonstrate the rAUC using our motivating example described in the introduction: ranking research articles for rapid reviews in epidemiology. We formulate this task in terms of a rate-constrained ranking problem. To reiterate,

Table 1. Spearman's rank correlations comparing the rankings of the 150 models, ranked using the rAUC (with rate distributions of Figure 5), NDCG and AUC

	$\alpha = 3$ $\beta = 19$	$\alpha = 7$ $\beta = 15$	$\alpha = 11$ $\beta = 11$	$\alpha = 15$ $\beta = 7$	$\alpha = 19$ $\beta = 3$
NDCG	0.565	0.438	0.235	0.092	0.018
AUC	0.872	0.951	0.975	0.927	0.886
$AUC \leqslant 0.95$	0.725	0.902	0.961	0.829	0.703
$\alpha = 3$ $\beta = 19$		0.923	0.791	0.676	0.610
$\alpha = 7$ $\beta = 15$			0.931	0.823	0.764
$\alpha = 11$ $\beta = 11$				0.964	0.925
$\alpha = 15$ $\beta = 7$					0.982



(a) AUC and rAUC **(b)** NDCG and rAUC

Fig. 6. Comparison of metrics for 150 models generated with various learners, datasets and distributions over rates

the articles are ranked by estimated study quality, and the objective is to maximise the number of high-quality articles the reviewer assesses given the rate constraints. In this setting, the rate is the proportion of articles that the team reviews, which is not known precisely. The search will return M articles and the reviewers are allotted T hours to complete the review. We use elicitation to determine appropriate parameters for the rate distribution, a method commonly used in epidemiology to establish feasible parameters for a distribution where there is no data from which to infer this. For simplicity, we consider the case of only one reviewer, who estimated the minimum (t_0) and maximum (t_1) time per article, t, the number of minutes they will on average expect to take to assess a single article.

We model t as a inverse beta distribution (with bounds $[\frac{T}{M}, \infty]$), having 0.95 probability of being in the range $[t_0, t_1]$. The rate (the proportion of articles that are reviewed) is given by: $r = \frac{T}{M \cdot t}$. This relationship with t infers a beta distribution across the rates.

We suppose a hypothetical and realistic rapid review where the search returns $M = 2,500$ articles and a reviewer is given 120 person hours ($T = 7,200$ minutes) in which to perform the review. We imagine that the reviewer states they will take between 10 and 45 minutes to assess a single article, which we use to specify two quantiles of t ($0.025 = CDF_t(0, 10)$ and $0.975 = CDF_t(0, 45)$) which we convert to equivalent quantiles of r ($0.975 = CDF_r(0, 0.288)$ and $0.025 = CDF_r(0, 0.064)$). We use the *beta.select* function of the *LearnBayes* R package [1] to find the α and β parameters with these quantiles, giving $\alpha = 6.23$ and $\beta = 32.80$ (shown in Figure 3 (right)).

We use a dataset consisting of 315 full-text articles reporting the results from randomised controlled trials, each labelled with a binary value denoting whether blinding – an indicator of study quality – has been adequately carried out (as described in the article). There were an approximately equal number of articles

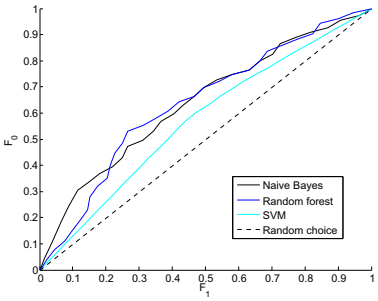


Fig. 7. Consensus ROC curves (using rate-averaging) predicting the blinding risk of bias value of research articles

of each class. We created a set of preliminary models using a bag of words representation, and evaluate these using 10 fold cross validation.

We generated consensus ROC curves for 3 learning algorithms: naive Bayes, decision tree and support vector machine (SVM), shown in Figure 7. A consensus curve represents an average across the ROC curves of all folds [13]. We used *rate-averaging* to generate our consensus curves, previously referred to as pooling [2], where the average of the true and false positive rates at each rate are calculated and then used to generate a single curve. This is appropriate for our rate-constrained task as the points of the consensus curves are the average performance given a particular rate constraint.

The random forest, naive Bayes and SVM models gave a mean rAUC (AUC) of 0.689 (0.636), 0.781 (0.639), and 0.639 (0.570), respectively, across the 10 folds. A two-tailed paired t-test of the AUC values of each model across the 10 cross validation folds, found no difference between the random forest and naive Bayes models ($p = 0.884$). A t-test using the rAUC values found the naive Bayes model is better than the random forest model for this rate distribution ($p = 0.021$). The random forest and naive Bayes models clearly dominate the SVM model such that the SVM model would be inferior for any rate distribution. However, we have shown that while the random forest and naive Bayes models are similar in terms of ranking performance across the entire ranking, the naive Bayes model is much better than the random forest when considering which rate values are more likely for this particular rapid review.

We thus clearly see that the weight distribution for rate-weighted AUC can be derived directly from the parameters of the rapid review task, in a way that could not be achieved with metrics such as the pAUC.

6 Related Work

The AUC is a popular choice to assess the performance of ranking models, estimating the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance, thus representing ranking performance across the entire dataset. Historically, the AUC has often been used

as a measure of ranking performance without consideration for the particular task at hand. However, when the performance of a learner in particular regions of ROC space has more importance than other areas for a particular task, the AUC is not an appropriate choice.

Alternatives to the AUC have previously been suggested to allow differential importance across true positive or true negative rates, for empirical [4, 12] and analytical [14] ROC curves. As mentioned in the introduction, [4] propose a partial AUC metric (pAUC) to restrict the evaluation of the AUC to a range of false positive or true positive rate values. The pAUC measure is appropriate when it is required that either the true positive or false positive rates fall in a particular range. This metric could be generalised using weights rather than bounds (as we have used for the rAUC), which may be more appropriate where there is a non-uniform probability distribution across either the true or false positive rate. Furthermore, a recent variant of the AUC called the half-AUC was proposed by [3], and evaluates the AUC in only half of the ROC space, either where true positive rate is less than true negative rate or true positive rate is greater than true negative rate, giving two distinct regions that can be assessed.

Several metrics have been suggested for early retrieval tasks, where evaluation focuses on the top of the rankings. Precision@k gives the precision at the top k results of a ranking, thus weighting each example uniformly within this section of the ranking. NDCG [10,11], is one of several metrics that give decreasing weights to examples along the ranking, as discussed in Section 3.1. Others include; robust initial enhancement (RIE) [15], the Boltzmann-enhanced Discrimination of ROC (BEDROC) [17], concentrated ROC (CROC) [16] and sum of the log ranks (SLR) [18]. The instance weights used by these approaches all share the characteristic that they translate into monotonically decreasing rate weights, which as demonstrated before is inappropriate for rate-constrained ranking tasks.

7 Conclusions

In this paper we have introduced a new ranking measure, the rate-weighted AUC (rAUC), to better reflect model performance when the task is constrained by a probability distribution across the predicted positive rate, which we refer to as the rate. The AUC is equivalent to the rAUC given a uniform distribution across the rates. Furthermore, if the rate is fixed then models can be compared by simply comparing the recall at the point on the ROC curve with this rate. We have derived the rAUC from both rate-recall and rate-accuracy space, and introduced rate-recall space as a visualisation of model performance. Furthermore, the rAUC is a linear transformation of rate-weighted expected recall (both the positive and negative respectively), given fixed class and rate distributions. We have described an $O(N)$ algorithm to calculate an estimate of the true rAUC using a data sample.

Our experiments have shown large variability of the rAUC as the rate distribution varies. A comparison with NDCG found low correlations indicating that when the likelihood that the processing will stop at a particular position in the

ranking is lower nearer the top of the ranking than elsewhere, NDCG may be inappropriate. Furthermore, a comparison with the AUC shows that often the rAUC prefers different models. Finally, we have also demonstrated how this approach can be usefully applied to real world tasks, using the example of ranking research articles for rapid reviews in epidemiology.

Acknowledgments. LACM is funded by a UK Medical Research Council studentship. This work was also supported by Medical Research Council grant MC_UU_12013/1-9.

References

1. Albert, J.: Learnbayes: Functions for learning Bayesian inference. R package version 2.12 (2008)
2. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7), 1145–1159 (1997)
3. Bradley, A.P.: Half-AUC for the evaluation of sensitive or specific classifiers. Pattern Recognition Letters 38, 93–98 (2014)
4. Dodd, L.E., Pepe, M.S.: Partial AUC estimation and regression. Biometrics 59(3), 614–623 (2003)
5. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27(8), 861–874 (2006)
6. Flach, P.A.: The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In: Proceedings of the 20th International Conference on Machine Learning, ICML 2003, pp. 194–201 (2003)
7. Ganann, R., Ciliska, D., Thomas, H.: Expediting systematic reviews: Methods and implications of rapid reviews. Implementation Science 5(1), 56 (2010)
8. Hand, D.J.: Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning 77(1), 103–123 (2009)
9. Higgins, J., Altman, D.G.: Assessing risk of bias in included studies. In: Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Book Series, pp. 187–241 (2008)
10. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41–48. ACM (2000)
11. Jarvelin, K., Kekalainen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) 20(4), 422–446 (2002)
12. Jiang, Y., Metz, C.E., Nishikawa, R.M.: A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 201(3), 745–750 (1996)
13. Macskassy, S.A., Provost, F., Rosset, S.: ROC confidence bands: An empirical evaluation. In: Proceedings of the 22nd International Conference on Machine Learning, ICML 2005, pp. 537–544. ACM (2005)
14. McClish, D.K.: Analyzing a portion of the ROC curve. Medical Decision Making 9(3), 190–195 (1989)
15. Sheridan, R.P., Singh, S.B., Fluder, E.M., Kearsley, S.K.: Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. Journal of Chemical Information and Computer Sciences 41(5), 1395–1406 (2001)

16. Swamidass, J., Azencott, C.-A., Daily, K., Baldi, P.: A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. Bioinformatics 26(10), 1348–1356 (2010)
17. Truchon, J.-F., Bayly, C.I.: Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. Journal of Chemical Information and Modeling 47(2), 488–508 (2007)
18. Zhao, W., Hevener, K.E., White, S.W., Lee, R.E., Boyett, J.M.: A statistical framework to evaluate virtual screening. BMC Bioinformatics 10(1), 225 (2009)