Efficient computation of the A' statistic

David Radcliffe

November 12, 2013

Abstract

We present an efficient method for computing the A' statistic for a binary classifier. We prove that our method is equivalent to the slower method of pairwise comparisons even when the classifier assigns equal ratings to some instances.

1 Introduction

The A' statistic is a measure of the accuracy of a binary classifier. It represents the probability that the classifier will rate a randomly chosen positive instance higher than a randomly chosen negative instance. The naive algorithm for computing this statistic is to compare each positive instance with each negative instance. In the worse case, the n instances are divided equally between the positive and negative classes, and the number of comparisons is $n^2/4$. This may be prohibitive with current computers when n is very large.

It is also possible to compute the statistic in time $O(n \log n)$ by regarding it as the area under the ROC curve. However, this approach can lead to inconsistent results when the data contains ties, i.e. positive and negative instances with the same rating. In this note, we will describe an $O(n \log n)$ algorithm that avoids this inconsistency and is fully equivalent to the method of pairwise comparisons.

2 Precise definition of the A' statistic

Suppose that we have a finite set of instances which can be separated into two classes: *positive* and *negative*. Suppose that we have a model that attempts to predict the class. It does so by assigning a *rating* to each instance, with the intention that instances with higher ratings are more likely to belong to the positive class. The ratings can be arbitrary real numbers, and we admit the possibility that difference instances will receive the same rating.

The ratings are used to construct a binary classifier. We choose a threshold value. Instances whose ratings exceed the threshold are predicted to be positive, and instances below the threshold are predicted to be negative. Selecting the threshold requires a careful balancing act. A high threshold will produce many false negatives, but a low threshold will produce many false positives.

The A' statistic is used to measure the ability of the classifier to distinguish between positive and negative instances. The statistic can be interpreted as a probability. We randomly choose one positive instance and one negative instance, and we attempt to distinguish them using the classifier by predicting that the instance with the higher rating is positive. If the ratings are equal then we toss a fair coin. The A' statistic is the probability that this procedure correctly identifies the positive instance.

3 Computing the A' statistic

We use the following notation. Let P and N be the total number of positive and negative instances, respectively. Let r be the number of distinct ratings, and let $x_1 < x_2 < \ldots < x_r$ be the distinct ratings produced by the classifier. Let p_i and n_i be the number of positive and negative instances with rating equal to x_i. Finally, let P_i and N_i be the number of positive and negative instance with rating less than x_i. Then

$$
\begin{aligned}
A' &= \frac{1}{PN}\left\{\left(\sum_{i=1}^{r}\sum_{j=1}^{i-1} p_i n_j\right) + \left(\frac{1}{2}\sum_{i=1}^{r} p_i n_i\right)\right\} && (1)\\
&= \frac{1}{PN}\sum_{i=1}^{r}\left[p_i\left(\sum_{j=1}^{i-1} n_j + \frac{n_i}{2}\right)\right] && (2)\\
&= \frac{1}{PN}\sum_{i=1}^{r}\left[p_i\left(N_i + \frac{n_i}{2}\right)\right] \ . && (3)
\end{aligned}
$$

A JavaScript implementation of this algorithm is available at `http://gotmath.com/doc/APrime.html` and a description in pseudocode is listed below.

Data: A list of r triples (x_i, p_i, n_i) with $x_1 < x_2 < \ldots < x_r$.
Result: The A' statistic.
$A \leftarrow 0$;
$P \leftarrow 0$;
$N \leftarrow 0$;
for $i \leftarrow 1$ **to** r **do**
  $A \leftarrow A + p_i * (N + n_i/2)$;
  $P \leftarrow P + p_i$;
  $N \leftarrow N + n_i$;
end
return $A/(P * N)$

Algorithm 1: An efficient algorithm for computing the A' statistic.

4 Geometric interpretation

The A' statistic can also be interpreted as the area under a curve. We plot the $r+1$ points $(P_1, N_1), (P_2, N_2), \ldots, (P_r, N_r), (P, N)$ in the coordinate plane and join adjacent pairs of points with line segments, as shown in Figure 1.

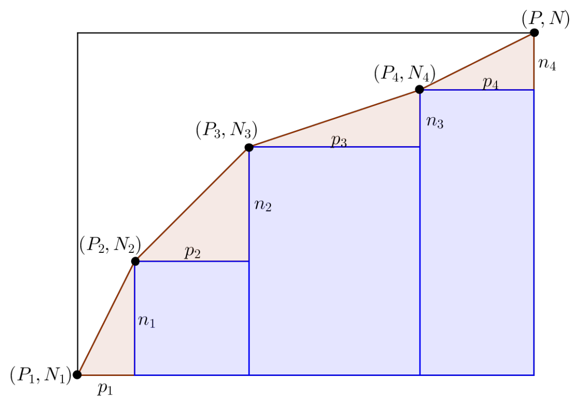

Figure 1: Geometric interpretation of the A' statistic.

Observe that the area under the i-th line segment is equal to $p_i(N_i + n_i/2)$, which is the i-th term of the sum in Equation 3. Therefore, the A' statistic is equal to the area under the curve, divided by the area PN of the enclosing rectangle. This curve is almost identical to the receiver operating characteristic (ROC) curve, except that the latter is scaled so that both coordinates range from 0 to 1. For background information on the ROC curve, see [1] and [2].

References

[1] JA Hanley and BJ McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[2] Wikipedia. Receiver operating characteristic — Wikipedia, The Free Encyclopedia, 2013. [Online; accessed 8-November-2013].