Rank Measures for Ordering

Jin Huang and Charles X. Ling

Department of Computer Science,
The University of Western Ontario,
London, Ontario, Canada N6A 5B7
{jhuang33, cling}@csd.uwo.ca

Abstract. Many data mining applications require a ranking, rather than a mere classification, of cases. Examples of these applications are widespread, including Internet search engines (ranking of pages returned) and customer relationship management (ranking of profitable customers). However, little theoretical foundation and practical guideline have been established to assess the merits of different rank measures for ordering. In this paper, we first review several general criteria to judge the merits of different single-number measures. Then we propose a novel rank measure, and compare the commonly used rank measures and our new one according to the criteria. This leads to a preference order for these rank measures. We conduct experiments on real-world datasets to confirm the preference order. The results of the paper will be very useful in evaluating and comparing rank algorithms.

1 Introduction

Ranking of cases is an increasingly important way to describe the result of many data mining and other science and engineering applications. For example, the result of document search in information retrieval and Internet search is typically a ranking of the results in the order of match. This leaves two issues to be addressed. First, given two orders of cases, how do we design or choose a measure to determine which order is better? Second, given two different rank measures of ordering, how do we tell which rank measure is more desirable?

In previous research, the issue of determining which order is better is usually addressed using accuracy and its variants, such as recall and F-measures, which are typically used in information retrieval. More recently, AUC (Area Under Curve) of the ROC (Receiver Operating Characteristics) has gained an increasing acceptance in comparing learning algorithms [1] and constructing learning models [2,3]. Bradley [4] experimentally compared popular machine learning algorithms using both accuracy and AUC, and found that AUC exhibits several desirable properties when compared to the accuracy.

However, accuracy is traditionally designed to judge the merits of classification results, and AUC is simply used as a replacement of accuracy without much reasoning for why it is a better measure, especially for the case of ordering. The main reason for this lack of understanding is that up to now, there has been no theoretical study on whether any of these measures work better than others, or whether there are even better measures in existence.

A. Jorge et al. (Eds.): PKDD 2005, LNAI 3721, pp. 503–510, 2005.

In this paper, we first review our previous work [5] on general criteria to compare two arbitrary single-number measures (see Section 2.1). Then we compare six rank measures for ordering using our general criteria. Our contributions in this part consist of a novel measure for the performance of ordering (Section 2.4), and a preference order discovered for these measures (Section 3.1). The experiments on real-world datasets confirm our analysis, which show that better rank measures are more sensitive in comparing rank algorithms (see Section 3.2).

2 Rank Measures for Ordering

In this section, we first review the criteria proposed in our previous work to compare two arbitrary measures. We then review five commonly used rank measures, and propose one new rank measure, OAUC. Then based on the comparison criteria, we will make a detailed comparison among these measures, which leads to a preference order of the six rank measures. Finally, we perform experiments with real-world data to confirm our conclusions on the preference order. The conclusions of the paper are significant for future machine learning and data mining applications involving ranking and ordering.

2.1 Review of Formal Criteria for Comparing Measures

In [5] the *degree of consistency* and *degree of discriminancy* of two measures are proposed and defined. The degree of consistency between two measures f and g, denoted as $\mathbf{C_{f,g}}$, is simply the fraction (probability) that two measures are consistent over some distribution of the instance space. Two measures are consistent when comparing two objects a and b, if f stipulates that a is better than b, g also stipulates that a is better than b. [5] define that two measures f and g are *consistent* iff the degree of consistency $\mathbf{C_{f,g}} > 0.5$. That is, f and g are consistent if they agree with each other on over half of the cases.

The *degree of discriminancy* of f over g, denoted as $\mathbf{D_{f/g}}$, is defined as the ratio of cases where f can tell the difference but g cannot, over the cases where g can tell the difference but f cannot. [5] define that a measure f is *more discriminant* (or *finer*) than g iff $D_{f/g} > 1$. That is, f is finer than g if there are more cases where f can tell the difference but g cannot, than g can tell the difference but f cannot.

2.2 Notation of Ordering

We will use some simple notations to represent ordering throughout this paper. Without loss of generality, for n examples to be ordered, we use the actual ordering position of each example as the label to represent this example in the ordered list. For example, suppose that the label of the actual highest ranked example is n, the label of the actual second highest ranked example is $n-1$, etc. We assume the examples are ordered incrementally from left to right. Then the *true-order list* is $l = 1, 2, \ldots, n$. For any ordered list generated by an ordering algorithm, it is a permutation of l. We use $\pi(l)$ to denote the ordered list generated by ordering algorithm π. $\pi(l)$ can be written as $a_1, a_2, \ldots, a_n$, where a_i is the actual ordering position of the example that is ranked ith in $\pi(l)$.

Table 1. An example of ordered lists

l	1	2	3	4	5	6	7	8
	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
$\pi(l)$	3	6	8	1	4	2	5	7

Table 1 gives an instance of ordered lists with eight examples. In this table, l is the true-order list and $\pi(l)$ is the ordered list generated by an ordering algorithm π. In $\pi(l)$ from left to right are the values of a_i. We can find that $a_1 = 3$, $a_2 = 6$, …, $a_8 = 7$.

2.3 Previous Rank Measures for Ordering

We first review five most commonly used rank measures. Later we will invent a new rank measure which we will evaluate among the rest.

We call some of the rank measures "true-order" rank measures, because to obtain the evaluation values, we must know the true order of the original lists. Some other rank measures, however, are not true-order rank measures. They do not need the true order to obtain evaluation values; instead, only a "rough" ordering is sufficient. For example, accuracy and AUC are not true-order rank measures. As long as we know the true classification, we can calculate their values. In a sense, positive examples can be regarded as "the upper half", and negative examples are the "lower half" in an ordering, and such a rough ordering is sufficient to obtain AUC and accuracy.

1. **Euclidean Distance (ED)**
   If we consider the ordered list and the true order as a point $(a_1, a_2, \ldots, a_n)$ and a point $(1, 2, \ldots, n)$ in an n-dimensional Euclidean space, then ED is the Euclidean Distance between these two points, which is $\sqrt{\sum_{i=1}^{n}(a_i - i)^2}$. For simplicity we use the squared value of Euclidean distance as the measure. Then $ED = \sum_{i=1}^{n}(a_i - i)^2$. Clearly, ED is a true-order rank measure.
   For the example in Table 1, It is easy to obtain that $ED = (3-1)^2 + (6-2)^2 + (8-3)^2 + (1-4)^2 + (4-5)^2 + (2-6)^2 + (5-7)^2 + (7-8)^2 = 76$.
2. **Manhattan Distance (MD)**
   This measure MD is similar to ED except that here we sum the absolute values instead of sum squared values. It is also a true-order rank measure. For our order problem $MD = \sum_{i=1}^{n}|a_i - i|$. For the example in Table 1, it is easy to obtain that $MD = |3-1| + |6-2| + |8-3| + |1-4| + |4-5| + |2-6| + |5-7| + |7-8| = 22$.
3. **Sum of Reversed Number (SRN)**
   This is roughly the sum of the reversed pairs in the list. That is, $SRN = \sum_{i=1}^{n} s(i)$. It is clearly a true-order measure.
   For the ith example, its reversed number $s(i)$ is defined as the number of examples whose positions in $\pi(l)$ are greater than i but the actual ranked positions are less than i. For the example in Table 1, we can find that the examples of 1 and 2 are both ranked higher than the first example 3 in $\pi(l)$. Thus $s(1) = 1 + 1 = 2$. Similarly we have $s(2) = 4$, $s(3) = 5$, etc. Therefore the SRN for the ordered list $\pi(l)$ is $SRN = 2 + 4 + 5 + 0 + 1 + 0 + 0 + 0 = 12$.

4. **Area Under Curve (AUC)**
   The Area Under the ROC Curve, or simply AUC, is a single-number measure widely used in evaluating classification algorithms, and it is not a true-order measure for ranking. To calculate AUC for an ordered list, we only need the true classification (positive or negative examples). For a balanced ordered ranked list with n examples (half positive and half negative), we treat any example whose actual ranked position is greater than $\frac{n}{2}$ as a positive example; and the rest as negative. From left to right we assume the ranking positions of positive examples are $r_1, r_2, \ldots, r_{\lceil \frac{n}{2} \rceil}$. Then $AUC = \frac{\sum_{a_{r_i} > n/2}(r_i - i)}{n^2}$ [6].
   In Table 1, 5, 6, 7, and 8 are positive examples positioned at 2, 3, 7, and 8 respectively. Thus, $AUC = \frac{(2-1)+(3-2)+(7-3)+(8-4)}{4\times 4} = \frac{5}{8}$.
5. **Accuracy (acc)**
   Like AUC, accuracy is also not a true-order rank measure. Similar to AUC, if we classify examples whose rank position above half of the examples as positive, and the rest as negative, we can calculate accuracy easily as $acc = \frac{tp+tn}{n}$, where tp and tn are the number of correctly classified positive and negative examples respectively. In the ordered list $\pi(l)$ in Table 1, 5, 6, 7, and 8 are positive examples, others are negative examples. Thus $tp = 2$, $tn = 2$. $acc = \frac{2+2}{8} = \frac{1}{2}$.

2.4 New Rank Measure for Ordering

We propose a new measure called Ordered Area Under Curve (OAUC), as it is similar to AUC both in meaning and calculation. The only difference is that each term in the formula is weighted by its true order, and the sum is then normalized. Thus, OAUC is a true-order measure. This measure is expected to be better than AUC since it "spreads" its values more widely compared to AUC.

OAUC is defined as follows:

$$OAUC = \frac{\sum a_{r_i}(r_i - i)}{\lfloor \frac{n}{2} \rfloor \sum_{i=1}^{\lceil \frac{n}{2} \rceil} (\lfloor \frac{n}{2} \rfloor + i)}$$

In the ordered list in Table 1, the positive examples are 5, 6, 7, 8 which are positioned at 7, 2, 8 and 3 respectively. Thus $r_1 = 2$, $r_2 = 3$, $r_3 = 7$, $r_4 = 8$, and $a_{r_1} = 6$, $a_{r_2} = 8$, $a_{r_3} = 5$,$a_{r_4} = 7$. $OAUC = \frac{6(2-1)+8(3-2)+5(7-3)+7(8-4)}{4((4+1)+(4+2)+(4+3)+(4+4))} = \frac{31}{52}$.

3 Comparing Rank Measures for Ordering

We first intuitively compare some pairs of measures and analyze whether any two measures satisfy the criteria of consistency and discriminancy. To begin with, we consider ED and MD because these two measures are quite similar in their definitions except that ED sums the squared distance while MD sums the absolute value. We expect that these two measures are consistent in most cases. On the other hand, given a dataset with n examples there are a total of $O(n^3)$ different ED values and $O(n^2)$ different MD values. Thus ED is expected to be more discriminant than MD. Therefore we expect that ED is consistent with and more discriminant than MD.

For AUC and OAUC, since OAUC is an extension of AUC, intuitively we expect that they are consistent. Assuming there are n_1 negative examples and n_0 positive examples, the different values for OAUC is $n_1 \sum_{i=1}^{n_0} (n_1 + i)$, which is greater than the different values of AUC ($n_0 n_1$). We can also expect that OAUC is more discriminant and therefore better than AUC.

However for the rest of the ordering measures we cannot make these intuitive claims because they have totally different definitions or computational methods. Therefore, in order to perform an accurate and detailed comparison and to verify or overturn our intuitions, we will conduct experiments to compare all measures.

3.1 Comparing Rank Measures on Artificial Datasets

To obtain the average degrees of consistency and discriminancy for all possible ranked lists, we use artificial datasets which consist of all possible ordered list of length 8. [1] We assume that the ordered lists are uniformly distributed. We exhaustively compare all pairs of ordered lists and calculate the degree of consistency and degree of discriminancy between two rank measures for ordering.

Table 2 lists the degree of consistency between every pair of six rank measures for ordering. The number in each cell represents the degree of consistency between the measures in the same row and column of the cell. We can find that the degree of consistency between any two measures are greater than 0.5, which indicates that these measures are "similar" in the sense that they are more likely to be consistent than inconsistent.

Table 3 shows the degree of discriminancy among all 6 rank measures. The number in the cell of the *i*th row and the *j*th column is the degree of discriminancy for the measure in *i*th row over the one in *j*th column.

From these two tables we can draw the following conclusions. First, these results verified our previous intuitive conclusions about the relations between ED and MD, and between AUC and OAUC. The degree of consistency between ED and MD is 0.95, and between AUC and OAUC 0.99, which means that ED and MD, and AUC and OAUC are highly consistent. The degree of discriminancy for ED over MD, and for OAUC over AUC are greater than 1, which means that ED is better than MD, and OAUC is better than AUC.

Table 2. Degree of consistency between pairs of rank measures for ordering

	AUC	SRN	MD	ED	OAUC	acc
AUC	1	0.88	0.89	0.87	0.99	0.98
SRN	0.88	1	0.95	0.98	0.89	0.91
MD	0.89	0.95	1	0.95	0.90	0.95
ED	0.87	0.98	0.95	1	0.88	0.90
OAUC	0.99	0.89	0.90	0.88	1	0.97
acc	0.98	0.91	0.95	0.90	0.97	1

[1] There are $n!$ different ordered lists for length n, so it is infeasible to enumerate longer lists.

Table 3. Degree of discriminancy between pairs of rank measures for ordering

	AUC	SRN	MD	ED	OAUC	acc
AUC	1	0.88	1.42	0.21	0.0732	14.0
SRN	1.14	1	1.84	0.242	0.215	9.94
MD	0.704	0.54	1	0.117	0.116	6.8
ED	4.76	4.13	8.55	1	0.87	38.2
OAUC	13.67	4.65	8.64	1.15	1	94.75
acc	0.071	0.10	0.147	0.026	0.011	1

Second, since all values of the degree of consistency among all measures are greater than 0.5, we can decide which measure is better than another only based on the value of degree of discriminancy. Recall (Section 2.1) that a measure f is better than another measure g iff $\mathbf{C_{f,g}} > 0.5$ and $\mathbf{D_{f/g}} > 1$. The best measure should be the one whose degrees of discriminancy over all other measures are greater than 1. From Table 3 we can find that all the numbers in the OAUC row are greater than 1, which means that the measure OAUC's degrees of discriminancy over all other measures are greater than 1. Therefore OAUC is the best measure. In the same way we can find that ED is the second best measure, and SRN is the third best. The next are AUC, MD, and acc is the worst.

Finally we can obtain the following preference order of for all six rank measures for ordering:

$$OAUC \succ ED \succ SRN \succ AUC \succ MD \succ acc$$

From the preference order we can conclude that OAUC, a new measure we design based on AUC, is the best measure. ED is the close, second best. The difference for these two measures are not very large (the degree of discriminancy for OAUC over ED is only 1.15). Therefore we should use OAUC and ED instead of others to evaluate ordering algorithms in most cases. Further, the two none-true-order classification measures AUC and accuracy do not perform well as compared with the true-order measures ED and SRN. This suggests that generally we should avoid using classification measures such as AUC and accuracy to evaluate ordering. Finally, MD is the worst true-order measure, and it is even worse than AUC. It should be avoided.

3.2 Comparing Rank Measures with Ranking Algorithms

In this section, we perform experiments to compare two classification algorithms in terms of the six rank measures. What we hope to conclude is that the better rank measures (such as OAUC and ED) would be more sensitive to the significance test (such as the t-test) than other less discriminant measures (such as MD and accuracy). That is, OAUC and ED are more likely to tell the difference between two algorithms than MD and accuracy can. Note that here we do not care about which rank algorithm predicts better; we only care about the sensitivity of the rank measures that are used to compare the rank algorithms. The better the rank measure (according to our criteria), the more sensitive it would be in the comparison, and the more meaningful the conclusion would be for the comparison.

We choose Artificial Neural Networks (ANN) and Instance-Based Learning algorithm (IBL) as our algorithms as they can both accept and produce continuous target. The ANN that we use has one hidden layer; the number of nodes in the hidden layer is half of the input layer (the number of attributes). We use real-world datasets to evaluate and compare ANN and IBL with the six rank measures. We select three real-world datasets *Wine*, *Auto-Mpg* and *CPU-Performance* from the UCI Machine Learning Repository [7].

In our experiments, we run ANN and IBL with the 10-fold cross validation on the training datasets. For each round of the 10-fold cross validation we train the two algorithms on the same training data and test them on the same testing data. We measure the testing data with six different rank measures (OAUC, ED, SRN, AUC, MD and acc) discussed earlier in the paper. We then perform paired, two-tailed t-tests on the 10 testing datasets for each measure to compare these two algorithms.

Table 4 shows the significance level in the t-test. [2] The smaller the values in the table, the more likely that the two algorithms (ANN and IBL) are significantly different, and the more sensitive the measure is when it is used to compare the two algorithms. Normally a threshold is set up and a binary conclusion (significantly different or not) is obtained. For example, if we set the threshold to be 0.95, then for the artificial dataset, we would conclude that ANN and IBL are statistically significantly different in terms of ED, OAUC and SRN, but not in terms of AUC, MD and acc. However, the actual significance level in Table 4 is more discriminant for the comparison. That is, it is "a better measure" than the simple binary classification of being significantly different or not.

Table 4. The significance level in the paired t-test when comparing ANN and IBL using different rank measures

Measures	Wine	Auto-mpg	CPU
OAUC	0.031	8.64×10^{-4}	1.48×10^{-3}
ED	0.024	1.55×10^{-3}	4.01×10^{-3}
SRN	0.053	8.89×10^{-3}	5.91×10^{-3}
AUC	0.062	5.77×10^{-3}	8.05×10^{-3}
MD	0.053	0.0167	5.97×10^{-3}
acc	0.126	0.0399	0.0269

From Table 4 we can obtain the preference order from the most sensitive measure (the smallest significance level) to the least sensitive measure (the largest significance level) for each dataset is:

- Wine: ED, OAUC, SRN = MD, AUC, acc.
- Auto-mpg: OAUC, ED, AUC, SRN, MD, acc.
- CPU-Performance: OAUC, ED, SRN, MD, AUC, acc.

These preference orders are roughly the same as the preference order of these measures discovered in the last section:

[2] The confidence level for the two arrays of data to be statistically different is one minus the values in the table.

$$OAUC \succ ED \succ SRN \succ AUC \succ MD \succ acc$$

The experimental results confirm our analysis in the last section. That is, OAUC and ED are the best rank measures for evaluating orders. In addition, MD and accuracy should be avoided as rank measures. These conclusions will be very useful for comparing and constructing machine learning algorithms for ranking, and for applications such as Internet search engines and data mining for CRM (Customer Relationship Management).

4 Conclusions

In this paper we use the criteria proposed in our previous work to compare five commonly used rank measures for ordering and a new proposed rank measure (OAUC). We conclude that OAUC is actually the best rank measure for ordering, and it is closely followed by the Euclidian distance (ED). Our results indicate that in comparing different algorithms for the order performance, we should use OAUC or ED, and avoid the least sensitive measures such as Manhattan distance (MD) and accuracy.

In our further work, we plan to improve existing rank learning algorithms by optimizing the better measures, such as OAUC and ED, discovered in this paper.

References

1. Provost, F., Domingos, P.: Tree induction for probability-based ranking. Machine Learning **52:3** (2003) 199–215
2. Ferri, C., Flach, P.A., Hernandez-Orallo, J.: Learning decision trees using the area under the ROC curve. In: Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002). (2002) 139–146
3. Ling, C.X., Zhang, H.: Toward Bayesian classifiers with accurate probabilities. In: Proceedings of the Sixth Pacific-Asia Conference on KDD. Springer (2002) 123–134
4. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition **30** (1997) 1145–1159
5. Ling, C.X., Huang, J., Zhang, H.: AUC: a statistically consistent and more discriminating measure than accuracy. In: Proceedings of 18th International Conference on Artificial Intelligence (IJCAI-2003). (2003) 519–526
6. Hand, D.J., Till, R.J.: A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning **45** (2001) 171–186
7. Blake, C., Merz, C.: UCI Repository of machine learning databases [http://www.ics.uci.edu/] [∼mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science (1998)