Evaluation Methods for Ordinal Classification

Lisa Gaudette and Nathalie Japkowicz

School of Information Technology and Engineering, University of Ottawa
lgaud082@uottawa.ca, nat@site.uottawa.ca

Abstract. Ordinal classification is a form of multi-class classification where there is an inherent ordering between the classes, but not a meaningful numeric difference between them. Little attention has been paid as to how to evaluate these problems, with many authors simply reporting accuracy, which does not account for the severity of the error. Several evaluation metrics are compared across a dataset for a problem of classifying user reviews, where the data is highly skewed towards the highest values. Mean squared error is found to be the best metric when we prefer more (smaller) errors overall to reduce the number of large errors, while mean absolute error is also a good metric if we instead prefer fewer errors overall with more tolerance for large errors.

1 Introduction

Ordinal classification, sometimes referred to as ordinal regression, represents a type of multi-class classification where there is an inherent ordering relationship between the classes, but where there is not a meaningful numeric difference between them [1]. This type of problem occurs frequently in human devised scales, which cover many domains from product reviews to medical diagnosis.

In this type of scenario, some errors are worse than others. A classifier which makes many small errors could be preferable to a classifier that makes fewer errors overall but which makes more large errors. This paper is motivated by work in ordinal sentiment analysis of online user reviews. In this domain, small errors are not so important as larger errors; humans are not perfect at detecting a 1 star difference on a 4 or 5 star scale [2], while classifying a 1 star review as a 5 star review is a very serious problem. In addition, this domain is highly imbalanced - a great deal of reviewers will rate a product as 5 stars. This paper will examine various evaluation measures in the context of this scenario.

2 Related Work

In recent years, there has been much discussion on the flaws of accuracy as a metric for comparing performance on machine learning tasks (see [3], among others). In addition to the flaws inherent in using accuracy for binary problems, in the ordinal case, accuracy tells us nothing about the severity of the error and in many applications this is important. Most papers on ordinal classification that

Y. Gao and N. Japkowicz (Eds.): Canadian AI 2009, LNAI 5549, pp. 207–210, 2009.

we have found simply use accuracy as an error measure, without considering whether or not it is an appropriate measure, such as [2, 1] among others. A few papers do use more interesting techniques; the "normalized distance performance measure" is used in [4] for work with image retrieval, and an AUC type measure for the ordinal case is introduced in [5], representing a volume under a surface.

While we have found little work discussing evaluation metrics for ordinal problems, there has been more work done comparing metrics for binary problems. A recent work which includes some multi-class problems is [6], which looks at correlations over a large variety of datasets and algorithms and then tests the metrics for sensitivity to different kinds of noise on an artificial binary problem.

3 Measures

Accuracy (ACC), Mean Absolute Error (MAE), and Mean Squared Error (MSE) are used as they are common, simple, evaluation metrics. Accuracy represents the number of correctly classified examples as a proportion of all examples, while MAE and MSE represent the mean of the absolute or squared difference between the output of the classifier and the correct label over all examples.

Linear Correlation (Correl) measures a linear relationship between two sets of numbers. A strong correlation between the classifier output and the actual labels should represent a good classifier.

Normalized Distance Performance Measure (NDPM) is a measure introduced by [7] that is designed for information retrieval in cases where a user has established relative preferences for documents, without necessarily referring to a predefined scale. NDPM is based on how often the user ranking and the system ranking are contradictory, and how often the user and system rankings are "compatible" - when the user ranking establishes a preference between two documents but the system ranking does not. In this sense it is a truly ordinal measure, and does not assume anything about the magnitude of the error between two classes.

Accuracy within n (ACC1, ACC2, etc) represents a family of measures which are similar to accuracy, however, they allow for a wider range of outputs to be considered "correct". In the case where the correct output is 4 stars, outputs of 3 stars and 4 stars would be considered accurate within 1. When there are k classes, accuracy within $k-2$ includes all outputs as correct except for the worst possible kind of error, that is, mistaking class k with class 1 or vice versa. Traditional accuracy can be referred to as ACC0 in this context.

Used together, these measures provide a more qualitative picture of the performance of the classifier and the magnitude of it's errors, while greatly summarizing the information in a confusion matrix. However it should be noted that used alone these measures suffer from many of the same problems as accuracy, and most often we want a single number to summarize the performance of a system. One thing these measures can do is allow us to define a situation in which a classifier is indisputably superior to another - when accuracy within n is higher for each n, including 0, (or higher for at least one and equal for others).

While this concept is not entirely novel, it is not used frequently. For example, [2] uses a similar idea they call "Rating Difference" when describing human performance on their task.

Accuracy + Correlation (ACC+Correl) is simply the mean of accuracy and correlation. The motivation behind this combination is that including the information provided by correlation should improve on accuracy by providing some indication of the severity of the errors. We are not aware of any other work using this combination as a metric.

4 Experiments

In order to test the performance of the measures, we build classifiers using the DVD reviews in the dataset from [8]. We use reviews from the large "unlabeled" file, which does still include labels of 1, 2, 4, and 5 stars. In order to create features, we first use an algorithm which scores the words found in a set of 2500 randomly selected reviews based on how often they appear in positive vs. negative documents, and keep the top and bottom 25% of words as features. We then create a bag of words feature set based on the appearance of those words in a different set of 2500 documents. We then train classifiers on the features using WEKA [9] with default settings to perform 10 fold cross validation. We used 4 different classifiers: SMO, J48, and the OrdinalClassClassifier method introduced in [1] with each of SMO and J48 as base classifiers.

We also performed tests on artificial data, however, these results have been omitted for space reasons.

5 Results

Correlations between the different measures are shown in Table 1. MSE correlates most strongly with ACC1, ACC2, and ACC3, which shows that is is best at

Table 1. Correlations (absolute values) between measures - 4 classifiers, imbalanced dataset

	MAE	MSE	Correl	NDPM	ACC + Correl	ACC	ACC1	ACC2	ACC3
MAE	1	0.944	0.256	0.065	0.384	0.523	0.904	0.874	0.775
MSE	0.944	1	0.234	0.014	0.272	0.226	0.923	0.944	0.898
Correl	0.256	0.234	1	0.821	0.955	0.307	0.006	0.122	0.364
NDPM	0.065	0.014	0.821	1	0.819	0.366	0.285	0.117	0.222
ACC+Correl	0.384	0.272	0.955	0.819	1	0.576	0.069	0.138	0.343
ACC	0.523	0.226	0.307	0.366	0.576	1	0.205	0.104	0.096
ACC1	0.904	0.923	0.006	0.285	0.069	0.205	1	0.924	0.696
ACC2	0.874	0.944	0.122	0.117	0.138	0.104	0.924	1	0.744
ACC3	0.775	0.898	0.364	0.222	0.343	0.096	0.696	0.744	1

capturing differences in the accuracies within n while combining the information into a single measure. MAE also correlates reasonably well with the accuracies within n while correlating better with simple accuracy. However, all of the other measures are very far behind, while in the artificial tests they seemed much closer. In particular, correlation appears to be a terrible measure in practice while it was promising in the artificial tests.

6 Conclusions

Given the imbalanced dataset studied, MSE and MAE are the best performance metrics for ordinal classification of those studied. MSE is better in situations where the severity of the errors is more important, while MAE is better in situations where the tolerance for small errors is lower. This is despite the fact that neither of these measures are truly ordinal by design.

For future work, we would like to expand this analysis across a wider collection of datasets and methods in the domain of ordinal sentiment classification. We are also interested in exploring the evaluation of this problem in a more formal framework.

References

1. Frank, E., Hall, M.: A simple approach to ordinal classification. Technical Report 01/05, Department of Computer Science, University of Waikato (2001)
2. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 2005) (2005)
3. Provost, F., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998) (1998)
4. Wu, H., Lu, H., Ma, S.: A practical SVM-based algorithm for ordinal regression in image retrieval. In: Proceedings of the eleventh ACM international conference on Multimedia (MM 2003) (2003)
5. Waegeman, W., Baets, B.D., Boullart, L.: ROC analysis in ordinal regression learning. Pattern Recognition Letters 29, 1–9 (2008)
6. Ferri, C., Hernández-Orallo, J., Modroiu, R.: An experimental comparison of performance measures for classification. Pattern Recognition Letters 30, 27–38 (2009)
7. Yao, Y.Y.: Measuring retrieval effectiveness based on user preference of documents. Journal of the American Society for Information Science 46, 133–145 (1995)
8. Blitzer, J., Dredze, M., Pereira, F.: Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007) (2007)
9. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)