Sets of Receiver Operating Characteristic Curves and their Use in the Evaluation of Multi-Class Classification

Stephan M. Winkler
Upper Austrian University of Applied Sciences,
College of Information Technology at Hagenberg
Hauptstraße 117,
A-4232 Hagenberg, Austria
stephan@heuristiclab.com

Michael Affenzeller
Upper Austrian University of Applied Sciences,
College of Information Technology at Hagenberg
Hauptstraße 117,
A-4232 Hagenberg, Austria
michael@heuristiclab.com

Stefan Wagner
Upper Austrian University of Applied Sciences,
College of Information Technology at Hagenberg
Hauptstraße 117,
A-4232 Hagenberg, Austria
stefan@heuristiclab.com

ABSTRACT

Within the last two decades, Receiver Operating Characteristic (ROC) Curves have become a standard tool for the analysis and comparison of classifiers since they provide a convenient graphical display of the trade-off between true and false positive classification rates for two class problems. However, there has been relatively little work examining ROC for more than two classes.

Here we present an extension of ROC curves which can be used for illustrating and analyzing the quality of multi-class classifiers. Instead of using one single curve, we deal with sets of curves which are calculated for each class separately. These are used for analyzing not only how exactly the classes are separated, but also how clearly the classifier is able to distinguish the given classes. Apart from making it possible to analyze the results graphically, several values describing the classifier's quality can be calculated.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology—*classifier design and evaluation*

General Terms

Measurement, Performance, Standardization

Keywords

Classifier Systems, Data Mining, Machine Learning, Pattern Recognition and Classification

1. INTRODUCTION

Receiver Operating Characteristic (ROC) analysis provides a convenient graphical display of the trade-off between true and false positive classification rates for two class problems [2]. Since its introduction in the medical and signal processing literatures ([3], [5]), ROC analysis has become a prominent method for selecting an operating point.

Here we present an extension to ROC analysis making it possible to measure the quality of classifiers for multi-class problems. Unlike other multi-class-ROC approaches which have been presented recently (see [2] or [4], e.g.) we propose a method based on the theory of ROC curves that creates sets of ROC curves for each class that can be analyzed separately or in combination. Thus, what we get is a convenient graphical display of the trade-off between true and false classifications for multi-class problems.

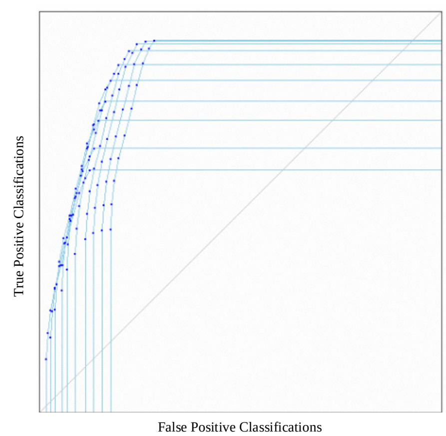

Figure 1: An exemplary Multi-ROC chart.

In the context of two class classification, ROC curves are calculated as follows: For each possible threshold value discriminating two given classes (e.g., 0 and 1 or 'true' and 'false'), the numbers of true and false classifications for one of the classes are calculated. For example, if the two classes 'true' and 'false' are to be discriminated using a given classifier, a fixed set of equidistant thresholds is tested and the true positives (TP) and the false positives (FP) are counted for each of them. Each pair of TP and FP values produces a point of the ROC curve. This method is very useful for analyzing the quality of two class classifiers, but unfortunately it is not directly applicable for more than two classes. This is why we here propose using sets of ROC curves for each class separately.

The work described here was done within the Translational Research Project L282 "GP-Based Techniques for the Design of Virtual Sensors" sponsored by the Austrian Science Fund (FWF).

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.
ACM 1-59593-186-4/06/0007.

Furthermore, ROC analysis often includes the calculation of the area under the ROC curve (AUC). Here we develop a generalization of this AUC analysis for multi-class problems which gives the operator the possibility to see not only how accurately, but also how clearly classes can be separated from each other.

2. MULTI-CLASS ROC ANALYSIS USING SETS OF ROC CURVES

The main idea presented here is that for each given class c_i the numbers of true and false classifications are calculated for each possible pair of thresholds between the classes c_{i-1} and c_i as well as between c_i and c_{i+1}. This is in fact done under the assumption that the n classes can be represented as real numbers and that $c_i < c_{i+1}$ holds for every $i \in [1, (n-1)]$ (with n being the number of classes). For a given class c_i, the corresponding TP and FP values (on the basis of the N original values $\vec{o}$ and estimated values $\vec{e}$) are calculated as:

$$\forall\ (\ \langle t_a, t_b \rangle \mid (c_{i-1} < t_a < c_i)\ \&\ (c_i < t_b < c_{i+1})\ ):$$
$$TP(t_a, t_b) = |\{e_j : (t_a \le e_j \le t_b)\ \&\ (t_a \le o_j \le t_b)\}|$$
$$FP(t_a, t_b) = |\{e_j : (t_a \le e_j \le t_b)\ \&\ (o_j < t_a \vee t_b < o_j)\}|$$

The resulting tuples of (FP,TP) values are stored in a matrix which can be plotted easily. This obviously yields a set of points which can be interpreted analog to the interpretation of 'normal' ROC curves: the closer the point are located to the left upper corner, the higher is the quality of the classifier at hand. For getting sets of ROC curves instead of ROC points, the following change is introduced: An arbitrary threshold t_a between the classes c_{i-1} and c_i is fixed and the FP and TP values for all possible thresholds t_b between c_i and c_{i+1} are calculated. This produces one single ROC curve; it is executed for all possible values of t_a (i.e., for all possible thresholds between c_{i-1} and c_i). Thus, what we get is a set of ROC curves. Please note that this procedure could also be executed the other way around, i.e. one could also choose an arbitrary threshold t_b between c_i and c_{i+1}, calculate all corresponding ROC points and repeat this for all values for all possible values of t_a. There is in fact no scientific reason for preferring one of the variants; either can be used because it does not significantly change the results. This is because the ROC points are not influenced by this decision, the only difference is that the ROC lines are arranged a little bit differently since the sets of ROC points joint to curves will differ.

An example showing 10 ROC curves is given in Figure 1; this MROC chart was generated for a classifier learned for a synthetical data set storing 2000 samples divided into 6 classes. A GP-based classification algorithm [6] combined with enhanced offspring selection concepts [1] was used for evolving this classifier.

Of course this procedure cannot be executed in exactly this way for the classes c_1 and c_n. For c_1 it is only possible to calculate the ROC points (and therefore the ROC curve) for all possible thresholds between c_1 and c_2, for c_n this is done analogically with all possible thresholds between c_{n-1} and c_n. This is why sets of ROC curves can be calculated for the classes $c_2 \ldots c_{n-1}$ whereas only simple ROC curves can be produced for c_1 and c_n.

As already mentioned, the most common quantitative index describing an ROC curve is the area under it (AUC). The bigger the area under a ROC curve is, the better the discriminator model is; if the two classes can be ideally separated, the ROC curve goes through the upper left corner and thus, the area under it reaches its maximal possible value which is exactly 1.0.

In the context of multi-class ROC (MROC) curves the two following values can be calculated assuming that all m ROC curves for a given class have already been calculated:

- The maximum AUC ($MaxAUC$) is the maximum of all areas under the ROC curves calculated for a specific class; it measures how exactly this class is separated from the others using optimal thresholds:

$$MaxAUC = \max_{i=1..m} AUC(ROC_i)$$

- The average AUC ($AvgAUC$) is calculated as the mean value of all areas under the ROC curves for a specific class; it measures how clearly this class is separated from the others since it takes into account all possible thresholds parameter settings:

$$AvgAUC = \frac{\sum_{i=1..m} AUC(ROC_i)}{m}$$

3. CONCLUSION

We have here presented a simple, but surely very useful and intuitive approach extending ROC analysis so that it can be used also in the context of multi-class classification. In addition to a graphical display, the average as well as the maximum area under the resulting ROC curves can be considered for evaluating multi-class classifiers. We are convinced that this enhanced ROC approach can be very useful in the context of investigating classification problems in various fields; recent discussions with experts in various scientific fields have strongly sustained this opinion.

4. REFERENCES

[1] M. Affenzeller, S. Wagner, and S. Winkler. Goal-oriented preservation of essential genetic information by offspring selection. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO) 2005*, 2:1595–1596, 2005.

[2] J. E. Fieldsend and R. M. Everson. Formulation and comparison of multi-class ROC surfaces. *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, 2005.

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

[4] A. Srinivasan. Note on the location of optimal classifiers in n-dimensional ROC space (technical report PRG-TR-2-99). Technical report, Oxford University Computing Laboratory, 1999.

[5] M. H. Zweig. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39:561–577, 1993.

[6] S. Winkler, M. Affenzeller, and S. Wagner. Using enhanced genetic programming techniques for evolving classifiers in the context of medical diagnosis - An empirical study. *Proceedings of the GECCO 2006 Workshop on Medical Applications of Genetic and Evolutionary Computation (MedGEC 2006)*, 2006.