Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 27 (2006) 908–917
Pattern Recognition Letters
www.elsevier.com/locate/patrec

The interaction between classification and reject performance for distance-based reject-option classifiers

Thomas C.W. Landgrebe *, David M.J. Tax, Pavel Paclík, Robert P.W. Duin

Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628CD Delft, The Netherlands

Available online 10 January 2006

Abstract

Consider the class of problems in which a *target* class is well-defined, and an *outlier* class is ill-defined. In these cases new *outlier* classes can appear, or the class-conditional distribution of the *outlier* class itself may be poorly sampled. A strategy to deal with this problem involves a two-stage classifier, in which one stage is designed to perform discrimination between known classes, and the other stage encloses known data to protect against changing conditions. The two stages are, however, interrelated, implying that optimising one may compromise the other. In this paper the relation between the two stages is studied within an ROC analysis framework. We show how the operating characteristics can be used for both model selection, and in aiding in the choice of the reject threshold. An analytic study on a controlled experiment is performed, followed by some experiments on real-world datasets with the distance-based reject-option classifier.

Keywords: Ill-defined classification problems; Unseen classes; Reject-option; Model selection; ROC analysis

1. Introduction

In pattern recognition, a typical assumption made during the design phase is that the various classes involved in a particular problem can be sampled reliably. However, in some problems, new classes or clusters may appear in the production phase that were not present during the design/training. In other problems, some classes may be sampled poorly, leading to inaccurate class models. Examples of applications that are affected by this are for instance:

- Diagnostic problems in which the objective of the classifier is to identify abnormal operation from normal operation (Dubuisson and Masson, 1993). It is often the case that a representative training set can be gathered for one of the classes, but due to the nature of the problem, the other class cannot be sampled in a representative manner. For example, in machine fault diagnosis (Ypma et al., 1999) a destructive test for all possible abnormal states may not be feasible or very expensive.
- Recognition systems that involve a rejection and classification stage, for example, road sign classification. Here a classifier needs not only to discriminate between examples of road sign classes, but must also reject non-sign class examples (Paclík, 2004). Gathering a representative set of non-signs may not be possible. Similarly face detection (Pham et al., 2002), where a classifier must deal with well-defined face classes, and an ill-defined non-face class, and handwritten digit recognition (Liu et al., 2002), where non-digit examples are a serious issue.

For simplicity we consider the problem as one in which there is a well-defined *target* class, and a poorly defined *outlier* class. The primary objective is to maintain a high

* Corresponding author. Address: P.O. Box 5031, 2600 GA Delft, Mekelweg 4, 2628CD Delft, The Netherlands. Tel.: +31 (0)15 27 88433; fax: +31 (0)15 27 81843.

E-mail addresses: t.c.w.landgrebe@ewi.tudelft.nl (T.C.W. Landgrebe), d.m.j.tax@ewi.tudelft.nl (D.M.J. Tax), p.paclik@ewi.tudelft.nl (P. Paclík), r.p.w.duin@ewi.tudelft.nl (R.P.W. Duin).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.10.015

classification performance between known classes, and simultaneously to protect the classes of interest from new/unseen classes (or changes in expected conditions, reflected in the change of distribution of these classes). We refer to the latter performance measure as *rejection* performance. Classification performance is defined between a well-defined *target* class ω_t, and some partial knowledge existing for the *outlier* class ω_o. Rejection performance is defined between ω_t and a new (unseen) cluster/class from the *outlier* class ω_r that is not defined precisely in training.

Several strategies have been proposed. The first strategy to cope with this situation was proposed in (Dubuisson and Masson, 1993), called the *distance*-based reject-option. Here a reject-rule was proposed to reject distant objects (with respect to the *target* class) post-classification. This evaluation differs considerably from the second strategy, the *ambiguity* reject-option (defined in (Dubuisson and Masson, 1993)) as proposed in (Chow, 1970). In ambiguity reject, a threshold is included to reject objects occurring in the *overlap* region between two known classes. It is assumed that all classes have been sampled in a representative manner. This is in contrast to this study, in which it is assumed that classes are poorly sampled or not sampled at all.

Classifiers with the reject-rule differ from conventional classifiers in that two thresholds are used to specify the target area, namely a classification threshold θ, and a rejection threshold t_d (we define the target area to be the region in the feature space in which all examples are labelled *target*). A limitation of the distance-reject criterion is that the threshold itself has no direct relationship with the distribution of the known classes, as discussed in (Muzzolini et al., 1978). Thus a modified reject-rule was proposed in (Muzzolini et al., 1978), involving computing the probability of a new object belonging to any of the known classes, based on covariance estimates. The threshold can then be based on a degree of model-fit to the known classes.

In (Landgrebe et al., 2004) we presented a third reject strategy, involving *combinations* of one-class (Tax, 2001) and supervised classifiers. This scheme allowed different models to be specifically designed for the purposes of classification or rejection. It was argued that a model optimised for the sake of classification may differ from that optimised for rejection, and that combining both optimised models can improve the overall combined classification/rejection performance. Experiments showed that this strategy outperforms the other reject-rules in some situations. It was also observed that a relation between the classification and rejection performance exists, and that optimising either performance is at the detriment of the other.

Each of the strategies has a classification and rejection threshold. In both (Dubuisson and Masson, 1993; Muzzolini et al., 1978), it has been shown how the distance-reject-rule can be applied in practise, involving distance- or class-conditional probability-thresholding of new incoming objects. In the case of the ambiguity reject-option, the classifiers can be evaluated and optimised since it is assumed that all classes have been sampled, as shown in (Chow, 1970) for known costs, and applied to imprecise environments in (Ferri and Hernandez-Orallo, 2004; Tortorella, 2004) to name a few. However, in the case of the distance-based reject-option, a challenging problem posed is that the distribution of the unseen class is by definition absent, and thus standard cost-sensitive evaluations and optimisations become ill-defined, lacking a closed Bayesian formalism.

In (Landgrebe et al., 2004), the ill-defined class problem was tackled by deriving strategies that can be used to study the way in which classification and rejection performance interact, based on the assumption that a new unseen class could occur anywhere in feature space. The rationale is that a minimal target area provides, in general, the most robust solution to an unseen class that could occur anywhere in feature space.[1] The methodology involved the artificial generation of the unseen class by assuming a uniformly distributed unseen class. Based on this methodology, it was observed that similar to the ambiguity-reject case, there is interaction between classification and rejection performance.

This paper is concerned with evaluating and optimising classifiers taking into account this interaction between classification and rejection. For this, receiver operating characteristic (ROC) curves will be used. ROC analysis (Metz, 1978), is a tool typically used in the evaluation of two-class classifiers in imprecise environments, plotting detection rate (true-positive rate) against the false positive rate. We extend this analysis to the unseen class problem by including an additional dimension that is related to the general robustness of the classifier to an unseen class. A similar 3-dimensional ROC analysis has been applied elsewhere, such as in (Ferri and Hernandez-Orallo, 2004; Mossman, 1999; Dreisetl et al., 2000), but in these cases this did not involve the ill-defined class problem. Our approach attempts to minimise the volume of the classes of interest in the feature space for robustness against unseen classes. It allows models to be compared (in a relative sense, since an absolute measure cannot be obtained) and provides insight into the choice of a reject threshold, that does not impact too much on classification performance.

In Section 2, an example is studied analytically to investigate the nature of the relation between classification and rejection rates, and the extended ROC analysis is presented. In Section 3, a criterion is proposed for the comparison of the extended ROC's. This criterion is applied to a synthetic 2-dimensional example with three different models. Finally, we discuss how to optimise an operating point (i.e. choose a classification and rejection threshold). Section 4 consists of a number of experiments to demonstrate the methodology in some realistic scenarios. Conclusions are given in Section 5.

[1] Rather than assuming that unseen classes can occur anywhere in feature space, it may be better to consider the nature of each problem, incorporating prior knowledge with respect to natural bounds in this space. To keep the discussion general, for now we assume a uniform, maximum entropy distribution.