Pattern Recognition 46 (2013) 2249–2255

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A new framework for optimal classifier design

Matías Di Martino*, Guzmán Hernández [1], Marcelo Fiori [2], Alicia Fernández [2]

Universidad de la República, Facultad de Ingeniería, Julio Herrera y Reissig 565, CC 30 - CP, Montevideo 11000, Uruguay

ARTICLE INFO

Article history:
Received 27 September 2012
Received in revised form
17 December 2012
Accepted 7 January 2013
Available online 17 January 2013

Keywords:
Class imbalance
One class SVM
F-measure
Recall
Precision
Fraud detection
Level set method

ABSTRACT

The use of alternative measures to evaluate classifier performance is gaining attention, specially for imbalanced problems. However, the use of these measures in the classifier design process is still unsolved. In this work we propose a classifier designed specifically to optimize one of these alternative measures, namely, the so-called *F*-measure. Nevertheless, the technique is general, and it can be used to optimize other evaluation measures. An algorithm to train the novel classifier is proposed, and the numerical scheme is tested with several databases, showing the optimality and robustness of the presented classifier.

1. Introduction

Evaluation measures have a crucial role in classifier analysis and design. Accuracy, Recall, Precision, *F*-measure, kappa, ACU [10] and some other new proposed measures like Informedness and Markedness [20] are examples of different evaluation measures. Depending on the problem and the field of application one measure could be more suitable than another. While in the behavioral sciences, Specificity and Sensitivity are commonly used, in the medical sciences, ROC analysis is a standard for evaluation. On the other hand, in the Information Retrieval community and fraud detection, Recall, Precision and *F*-measure are considered appropriate measures for testing effectiveness.

In a learning design strategy, the best rule for the specific application will be the one that get the optimal performance for the chosen measure.

Looking for the best decision rule, in a Bayesian framework, implies to minimize the overall risk taking into account the different misclassification costs [7]; in an equal misclassification cost problem we can find this optimal solution, with maximum Accuracy, selecting the class that has the maximum a posteriori probability.

However, finding a decision rule that looks for minimum error rate or maximum Accuracy in an imbalanced domain gives solutions strongly biased to favor the majority class, getting poor performance.

This problem is particularly important in those applications where the instances of a class (the majority one) heavily outnumber the instances of the other (the minority) class and it is costly to misclassify samples from the minority class. For example in information retrieval [15], nontechnical losses in power utilities [6,17,18] or medical diagnosis [8].

Identifying these rare events is a challenging issue with great impact regarding many problems in pattern recognition and data mining. The main difficulty in finding discriminatory rules for these applications is that we have to deal with small data sets, with skewed data distributions and overlapping classes. A range of classifiers that work successfully for other applications (decision trees, neural networks, support vector machines (SVMs), etc.) get a poor performance in this context [24]. For example, in a decision tree the pruning criterion is usually the classification error, which can remove branches related with the minority class. In back-propagation neural networks, the expected gradient vector length is proportional to the class size, and so the gradient vector is dominated by the prevalent class and consequently the weights are determined by this class. SVMs are thought to be more robust to the class imbalance problem since they use only a few support vectors to calculate region boundaries. However, in a two class problem, the boundaries are determined by the prevalent class, since the algorithm tries to find the largest margin and the minimum error. A different approach is taken in one-class learning,

* Corresponding author. Tel.: +598 2711 5444; fax: +598 2711 1630.
E-mail addresses: matiasdm@fing.edu.uy (M. Di Martino), guzmanhc@fing.edu.uy (G. Hernández), mfiori@fing.edu.uy (M. Fiori), alicia@fing.edu.uy (A. Fernández).
[1] Tel.: +598 2711 5444.
[2] Tel.: +598 2711 0974.

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.01.006

for example one class SVM, where the model is created based on the samples of only one of the classes. In [21] the optimality of one-class SVMs over two-class SVM classifiers is demonstrated for some important imbalanced problems.

Recently, great effort has been done to give better solutions to class imbalance problems (see [24,11,12] and references therein). In most of the approaches that deal with an imbalanced problem, the idea is to adapt the classifiers that have good Accuracy in balanced domains. A variety of ways of doing this have been proposed: changing class distributions [4,5,13], incorporating costs[3] in decision making [2,1], and using alternative performance metrics instead of Accuracy in the learning process with the standard algorithms [10].

In this work we propose a different approach to this problem, designing a classifier based on an optimal decision rule that maximizes a chosen evaluation measure, in this case the *F*-measure [27]. More specifically, if Ω is the feature space, we are looking for the classifier $u:\Omega\to\mathbb{R}$ that maximizes the *F*-measure. Here, given the feature vector x, the classifier (or decision function) u assigns the class ω_+ if $u(x)>0$, and class ω_- if $u(x)<0$. We address this problem by proposing an energy $E[u]$ such that its minimum is achieved for the optimal classifier u (in the sense of the *F*-measure). We solve this optimization problem using a gradient descent flow, inspired by the level-set method [19]. Although the analysis is made for *F*-measure, it could be extended to the other measures. In the particular case when the chosen measure is the Accuracy the proposed algorithm is equivalent to the Bayes approach.

We also show that, in contrast with common solutions, the proposed algorithm does not need to change original distributions or arbitrarily assign misclassification costs to find an appropriate decision rule for severe imbalanced problems. Although there is consensus about the need of using suitable evaluation measures for classifier design, to the best of our knowledge no technique has been proposed that optimizes these alternative measures over all decision frontiers.

The rest of the paper is organized as follows. In Section 2 the optimal classifier for the *F*-measure is proposed, and a numerical scheme to obtain it is presented. Experimental results are shown in Section 3, and we conclude in Section 4.

2. Proposed classifier formulation

In this paper we assume that there are two classes, one called here the *negative* class that represents the majority class, usually associated with the normal scenario, and the other called the *positive* class that represents the minority class. We define $C=\{\omega_+,\omega_-\}$ as the set of possible classes, being *TP* (true positive) the number of $x\in\omega_+$ correctly classified, *TN* (true negative) the number of $x\in\omega_-$ correctly classified, *FP* (false positive) and *FN* (false negative) the number of $x\in\omega_-$ and $x\in\omega_+$ misclassified respectively. Let us also Recall some related well known definitions:

$$\text{Accuracy}: \quad \mathcal{A}=\frac{TP+TN}{TP+TN+FP+FN},$$

$$\text{Recall}: \quad \mathcal{R}=\frac{TP}{TP+FN},$$

$$\text{Precision}: \quad \mathcal{P}=\frac{TP}{TP+FP},$$

$$F\text{-measure}: \quad F_\beta=\frac{(1+\beta^2)\mathcal{R}\mathcal{P}}{\beta^2\mathcal{P}+\mathcal{R}}.$$

[3] The misclassification cost can be set by experts or learned [24].

Precision and Recall are two important measures to evaluate the performance of a given classifier in an imbalance scenario. The Recall indicates the true positive rate, while the Precision indicates the positive predictive value. The *F*-measure combines them with a parameter $\beta\in[0,+\infty)$. With $\beta=1$, F_β is the harmonic mean between Recall and Precision, meanwhile with $\beta\gg1$ or $\beta\ll1$, the F_β approaches the Recall or the Precision respectively. A high value of F_β ensures that both Recall and Precision are reasonably high, which is a desirable property since it indicates reasonable values of both true positive and false positive rates. The best β value for a specific application depends on the adequate relation between Recall and Precision for each particular problem [15].

The task of finding a classifier consists in defining the regions Ω_+ and Ω_- of Ω, such that if x belongs to Ω_+/Ω_-, it will be classified as belonging to the positive/negative class. To train the classifier to maximize a given performance measure, we must therefore find the regions Ω_+ and Ω_- that give maximal performance measure for the available data set.

In order to find the classifier that maximizes a given performance measure, we must be able to express the quantities *FN*, *FP* and *TP* in terms of Ω_+ and Ω_-. These can be calculated by computing which points of the training data set belong to the regions Ω_+ and Ω_-. However, for the realization of the proposed algorithm, we will estimate these quantities in terms of probability densities for the positive and negative classes. To this end, we suppose that we have estimates for certain density functions, $f_+(x)$ and $f_-(x)$, such that in terms of these functions, we have the following approximations for the quantities *FN*, *FP*, *TP* and *TN*:

$$FN=P\int_{\Omega_-}f_+(x)\,dx, \tag{1}$$

$$FP=N\int_{\Omega_+}f_-(x)\,dx, \tag{2}$$

$$TP=P\int_{\Omega_+}f_+(x)\,dx, \tag{3}$$

$$TN=N\int_{\Omega_-}f_-(x)\,dx, \tag{4}$$

where P and N are the number of positive and negative instances in the training database, and the distribution functions $f_+(x)$ and $f_-(x)$ satisfy

$$\int_\Omega f_\pm(x)\,dx=1. \tag{5}$$

If these functions are known, the task of finding the optimal classifier consists in finding the regions Ω_+ and Ω_- that maximize the chosen measure. As was mentioned before, this choice depends on the particular problem or application considered. In this paper we have chosen *F*-measure as the evaluation measure, and in the next subsection we present an algorithm to determine the optimal boundaries for this measure. However, the framework is general, and the generalization to other evaluation measures that combine *FN, FP, TN* and *TP* is straightforward.

2.1. Optimal boundary determination for F-measure

It can be seen that maximizing *F*-measure is equivalent to minimizing the quantity:

$$\epsilon=\frac{\beta^2FN+FP}{TP}. \tag{6}$$

The quantities *FN*, *FP*, and *TP* can be expressed in terms of the functions $f_\pm(x)$, as was defined in the previous section. Therefore the task of training a classifier that maximizes *F*-measure (and minimizes ϵ) can be approached as finding the regions Ω_+ and Ω_-