Pattern Recognition 48 (2015) 591–604

Contents lists available at ScienceDirect

Pattern Recognition

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

Quantification-oriented learning based on reliable classifiers

Jose Barranquero, Jorge Díez, Juan José del Coz*

Artificial Intelligence Center (University of Oviedo), Campus de Viesques s/n 33204, Spain

ARTICLE INFO

Article history:
Received 12 July 2013
Received in revised form
13 June 2014
Accepted 31 July 2014
Available online 11 August 2014

Keywords:
Quantification
Class distribution estimation
Performance metrics
Reliability
Multivariate predictions

ABSTRACT

Real-world applications demand effective methods to estimate the class distribution of a sample. In many domains, this is more productive than seeking individual predictions. At a first glance, the straightforward conclusion could be that this task, recently identified as quantification, is as simple as counting the predictions of a classifier. However, due to natural distribution changes occurring in real-world problems, this solution is unsatisfactory. Moreover, current quantification models based on classifiers present the drawback of being trained with loss functions aimed at classification rather than quantification. Other recent attempts to address this issue suffer certain limitations regarding reliability, measured in terms of classification abilities. This paper presents a learning method that optimizes an alternative metric that combines simultaneously quantification and classification performance. Our proposal offers a new framework that allows the construction of binary quantifiers that are able to accurately estimate the proportion of positives, based on models with reliable classification abilities.

1. Introduction

Any data scientist who had tackled real-world problems knows that there exist classification domains that are inherently complex, it being very difficult to obtain accurate predictions when focusing on each specific example; i.e., to achieve high classification accuracy. However, it is not so strange to require estimations about the characteristics of the overall sample instead, mainly with respect to data distribution. Tentative application scopes include opinion mining [1], network-behavior analysis [2], remote sensing [3], quality control [4], word-sense disambiguation [5], monitoring of support-call logs [6], credit scoring [7] and adaptive fraud-detection [8], among others.

For instance, in order to measure the success of a new product, there is an increasing demand for methods for tracking overall consumer opinion, superseding classical approaches aimed at individual perceptions. To answer questions like *how many clients are satisfied with our new product?* we need effective algorithms focused on estimating the distribution of classes from a sample. This has emerging relevance when dealing with the tracking of trends over time [9], such as early detection of epidemics and endangered species, risk prevalence, market and ecosystem evolution, or any other kind of distribution change in general.

In many business, scientific and medical applications, it is sufficient, and sometimes even more relevant, to obtain estimations at an aggregated level in order to properly plan strategies. Companies could obtain greater returns on investment if they are able to accurately estimate the proportion of *events* that will involve higher costs or benefits. This will avoid wasting resources in guessing the class of each specific event; a task that usually reveals itself as complex, expensive and error-prone. For example, the estimation of the proportion of policy holders that will be involved in accidents during the next year, or the estimation of overall consumer satisfaction with respect to any specific product, service or brand.

In machine learning, the task of quantification is *to accurately estimate the number of positive cases* (*or class distribution*) *in a test set, using a training set that may have a substantially different distribution* [10]. Despite having many potential applications, this problem has barely been addressed within the community, and has yet to be properly standardized in terms of error measurement, experimental setup and methodology in general. Unfortunately, quantification has attracted little attention due to the mistaken belief of it being somewhat trivial. The key problem is that it is not as simple as classifying and counting the examples of each class, seeing as different distributions of train and test data can have a huge impact on the performance of state-of-the-art classifiers. The general assumption made by classification methods is that the samples are representative [11], which implies that the within-class probability densities, $Pr(\boldsymbol{x}|y)$, and the a priori class distribution, $Pr(y)$, do not vary.

The influence of different changing environments on classification and the performance of knowledge-based systems has been analyzed in several studies (see, for instance, [7,12,13]), suggesting that addressing distribution drifts is a complex and critical problem. Moreover, many papers focus on addressing distribution

* Corresponding author. Tel.: +34 985182501; fax: +34 985182125.
E-mail addresses: barranquero@aic.uniovi.es (J. Barranquero), jdiez@aic.uniovi.es (J. Díez), juanjo@aic.uniovi.es (J. José del Coz).

http://dx.doi.org/10.1016/j.patcog.2014.07.032
0031-3203/© 2014 Elsevier Ltd. All rights reserved.

changes for classification, offering different views of what is subject to change and what is assumed to be constant. As in previous quantification-related papers, we focus only on studying changes in the a priori class distribution, while maintaining within-class probability densities constant. Domains of this kind are identified as $Y \rightarrow X$ problems by Fawcett and Flach [14]. Provided that we use stratified sampling [15], an example of situations where $Pr(\boldsymbol{x}|y)$ does not change is when the number of examples of one or both classes is conditioned by the costs associated with obtaining and labeling them [16]. The explicit study of other types of distribution shifts, as well as $X \rightarrow Y$ domains, fall outside the scope of this paper (for further reading, we refer the reader to [17–20]).

Receiver Operating Characteristic (ROC) analysis is quite a popular technique for the graphical analysis of classification models [21]. A classifier may be trained for one particular operating condition, defined by one class distribution and cost proportion, but might then be deployed on a different condition. ROC curves visualize how the true positive rate (TPR) and the false positive rate (FPR) evolve for the same classifier for a range of thresholds. The threshold is the element to adapt a classifier to a given operating condition. ROC-based methods [8,22] and cost curves [23] have been successfully applied to adjust the classification threshold, given that new class priors are known in advance. However, as already stated by Forman [10], these approaches are not useful for estimating class distributions from test sets. Similarly, if these new priors are unknown, two main approaches have been followed in the literature. On the one hand, most published papers focus on adapting the deployed models to the new conditions [24–28]. On the other hand, the alternative view is mainly concerned with enhancing *robustness* in order to learn models that are more resilient to changes in class distribution [29]. Whatever the case may be, the aim of these methods, although related, is quite different from that of quantification, as adapting a classifier for improving individual classification performance does not imply obtaining better quantification predictions, as we shall discuss later. Moreover, there exists a natural connection with imbalance-tolerant methods, mainly those based on preprocessing of data [30]. Actually, quantification was originally designed to deal with highly imbalanced datasets [10]; however, these preprocessing techniques are not directly applicable in changing environments.

The main approach that has been studied in the literature for learning an explicit binary-quantification model is based on standard classifiers, following a two-step training procedure. The first step is to train a classifier optimizing a classification metric, usually accuracy. The next step is then to study some relevant properties of this classifier. The aim of this second step is to correct the quantification prediction obtained from aggregating classifier estimates [10,31].

An open question is whether it may be more effective to learn a classifier optimizing a quantification metric, instead of a classification performance measure. Conceptually, this alternative strategy is more formal, because the learning process takes into account the target performance measure. The main contribution of this paper is to explore this approach in detail.

The idea of optimizing a pure quantification metric during learning was introduced by Esuli and Sebastiani [1], although these authors neither implement nor evaluate it. Their proposal is based on learning a binary classifier with optimum quantification performance. We argue that this method has a pitfall. The key problem that arises when optimizing a pure quantification measure is that the resulting hypothesis space contains several global optimums. In practice, however, these optimum hypotheses are not equally good due to the fact that they differ in terms of the quality of their future quantification predictions. This paper claims that the robustness of a quantifier based on an underlying classifier is directly related to the reliability of such classifier. For instance, given several models showing equivalent quantification performance during training, the learning method should prefer the best one in terms of its potential for generalization. As we shall analyze later, this factor is closely related to their classification abilities.

This lead us to further explore Esuli and Sebastiani's approach with the aim of building a learning method able to induce more robust quantifiers based on classifiers that are as reliable as possible. In order to accomplish this goal, we introduce a new metric that combines both factors. That is, a metric that combines classification performance with quantification performance, resulting in better quantification models.

As occurs with any other quantification metric, our proposal measures performance from an aggregated perspective, taking into account the whole sample. The difficulty involved in optimizing such functions is that they are not decomposable as a linear combination of the individual errors. Hence, not all binary learners are capable of optimizing them directly, requiring a more advanced learning machine. In this paper we adapt Joachim's multivariate SVMs [32] to implement our proposal and the idea presented by Esuli and Sebastiani. In order to validate these two approaches, another key contribution is to perform an exhaustive study in which we compare them, along with several state-of-the-art quantifiers, by means of benchmark datasets from the UCI Machine Learning repository [33].

The paper is organized as follows. Section 2 introduces binary quantification as a learning task. Core concepts, notation and performance metrics for binary quantification are presented first. Then, a brief review of available quantification methods is provided, including those approaches based on adjusted classification (Section 2.2.2) and threshold selection policies (Section 2.2.3). Quantification-oriented learning is analyzed in depth in Section 3. First, we describe the idea proposed by Esuli and Sebastiani. Then, we discuss a possible pitfall in their approach. Finally, we introduce our method (Section 3.3), based on a new quantification measure called *Q-measure*. For a better understanding of our proposal, we describe *Q-measure*, both conceptually and graphically, in comparison with other performance measures. Section 4 reports the experiments performed, including the experimental setup, datasets, algorithms and statistical tests employed. The results are discussed in terms of different quantification measures. The paper ends by drawing some conclusions in Section 5.

2. Binary quantification

From a statistical point of view, the aim of a binary quantification task is to estimate the prevalence of an event or property within a sample. During the learning stage, we have a training set with examples labeled as positives or negatives; formally, $D = \{(\boldsymbol{x}_i, y_i) : i = 1 \ldots S\}$, in which $\boldsymbol{x}_\mathrm{i}$ is an object of the input space $\mathcal{X}$ and $y_i \in \mathcal{Y} = \{-1, +1\}$. This dataset shows a specific distribution that can be summarized with the actual proportion of positives or prevalence. The learning goal is to obtain a model able to predict the prevalence (p) of another sample, usually identified as the test set, that may show a markedly different distribution of classes. Thus, the input data is equivalent to that of traditional classification problems, but the focus is on the estimated prevalence (p') of the sample, rather than on the class assigned to each individual example. Notice that we use p and p' to identify the actual and estimated prevalences of any sample; these variables are not tied to training or test sets in any way.

Table 1 summarizes the notation that we shall employ throughout the paper. First, an algorithm is applied over the training set in