Information Fusion 34 (2017) 87–100

Contents lists available at ScienceDirect

Information Fusion

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

Full Length Article

Using ensembles for problems with characterizable changes in data distribution: A case study on quantification

Pablo Pérez-Gállego, José Ramón Quevedo, Juan José del Coz*

Artificial Intelligence Center, University of Oviedo, Gijón, Spain

ARTICLE INFO

Article history:
Received 4 January 2016
Revised 25 June 2016
Accepted 2 July 2016
Available online 4 July 2016

Keywords:
Distribution changes
Ensembles
Quantification

ABSTRACT

Ensemble methods are widely applied to supervised learning tasks. Based on a simple strategy they often achieve good performance, especially when the single models comprising the ensemble are diverse. Diversity can be introduced into the ensemble by creating different training samples for each model. In that case, each model is trained with a data distribution that may be different from the original training set distribution. Following that idea, this paper analyzes the hypothesis that ensembles can be especially appropriate in problems that: (i) suffer from distribution changes, (ii) it is possible to characterize those changes beforehand. The idea consists in generating different training samples based on the expected distribution changes, and to train one model with each of them. As a case study, we shall focus on binary quantification problems, introducing ensembles versions for two well-known quantification algorithms. Experimental results show that these ensemble adaptations outperform the original counterpart algorithms, even when trivial aggregation rules are used.

1. Introduction

Ensemble learning consists in constructing a meta-model that results from the combination of a set of individual models, using a particular aggregation rule. Ensembles get benefited from the existent diversity in the model set, producing a solution that implicitly represents some sort of agreement between the individual models. From a practical point of view, they generally perform better than a single-model solution [1–4], although this cannot be guaranteed [5]. An ensemble limits the risk of obtaining a particular bad response from a single model; formally this is due to the fact that the ensemble tends to reduce the variance of its base classifier. Intuitively, the same idea is highly present in the human decision making processes; a set of opinions is more rich than an isolated opinion, especially when there exist a high degree of diversity within the opinions.

Let us introduce some notation for ensembles under the framework of supervised learning. Let $\mathcal{X}$ be an input space and $\mathcal{Y}$ an output space. There exist a training set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn from an unknown distribution $\mathbf{P}(\mathbf{X}, \mathbf{Y})$ from the product $\mathcal{X} \times \mathcal{Y}$. Usually each example, x_i, is represented by an attribute vector $(x_{i,1}, x_{i,2} \ldots, x_{i,d})$ and a target class y_i that may belong to a discrete set in classification problems or to $\mathbb{R}$ in the case of a regression problem. The objective is to approximate an unknown function $f : \mathcal{X} \to \mathcal{Y}$ by generating a function h, called hypothesis, defined into some hypothesis space $\mathcal{H}$. To do so, many algorithms search for the best single hypothesis h that approximates f taking into account the training set D, the selected hypothesis space and a target loss function. In contrast, an ensemble produces a hypothesis h resulting from the combination of a set of m (weak) hypothesis $\{h_1, h_2, \ldots, h_m\}$, in which each model h_j is usually learned using a subsample D_j generated from D. The combination of the set of hypothesis or models is performed by means of a particular aggregation strategy.

Supervised learning makes the assumption that the unknown distribution $\mathbf{P}(\mathbf{X}, \mathbf{Y})$, from where the examples are drawn independently and identically distributed (i.i.d. assumption), does not change between the training and testing or production phases. Or state it differently, it is assumed that the training set truly reflects the probability distribution of the problem. However, in practice, this assumption gets often violated in real-world applications [6,7]. This situation is referred to as *dataset shift* in the research community, and it takes place when $\mathbf{P}(\mathbf{X}, \mathbf{Y})$ changes from training to testing data [8,9]. Characterizing those changes in the distribution sometimes depend on the target application, and even though it can be challenging in some cases, it is surprisingly simple, even trivial, in other problems.

Our intention in this paper is to present a new scenario in which the application of ensemble algorithms results appropriate

* Corresponding author.
E-mail addresses: pablopg@aic.uniovi.es (P. Pérez-Gállego), quevedo@aic.uniovi.es (J.R. Quevedo), juanjo@aic.uniovi.es, juanjo@uniovi.es (J.J. del Coz).

http://dx.doi.org/10.1016/j.inffus.2016.07.001
1566-2535/

and effective. We refer to problems verifying the next properties: (i) are known to suffer from distribution changes between training and testing phases, (ii) we are able to characterize the distributional changes beforehand (i.e. we can define the conditions that make $\mathbf{P}(\mathbf{X}, \mathbf{Y})$ to change). The objective is to take advantage of this a priori knowledge and to use it during the training stage. From an ensemble point of view, we can make the most of this knowledge in order to introduce *diversity* into the weak models, a desirable property to boost its efficacy [10–14]. The central idea of this paper is to generate different training samples, with each one representing an specific and expected distribution change. This approach is different to other propositions that have been suggested to tackle tasks that present some sort of drift in the distribution [15], especially *concept drift* problems [16]. Those methods are based on removing, modifying or adding new models to the ensemble, mainly because the concept $\mathbf{P}(y|x)$ changes throughout time and it is not possible to exploit any prior knowledge. Our approach is different in the sense that, as we know the characteristics of the expected changes, we can use that knowledge to build an enriched ensemble from the beginning without the need of subsequent modifications.

In order to prove the validity of our idea, we have applied it to *binary quantification* problems. Quantification is defined as the task of estimating the number of examples belonging to each class (class distribution) in a test set, using a training set that may have been drawn from a different distribution [17]. In the case of binary quantification the set of class values is restricted to two and the objective is to correctly estimate the number of positive examples (prevalence). Quantification tasks fit perfectly our requirements, since, by definition, the class probabilities may change, and we are also able to characterize and to restrict ourselves to a certain types of changes, as we shall see later.

There are many real-world problems that can be solved using quantification algorithms. Tentative application scopes include opinion mining [18], network-behavior analysis [19], quality control [20], monitoring of support-call logs [21] and credit scoring [22], among others. For instance, there is an increasing demand for automatic methods to track overall consumer opinions [23]. The goal is to answer questions like *how many consumers are satisfied with our new product?*. This task requires effective algorithms focused on estimating the distribution of classes from a sample. Notice that the goal is not to label individual examples (solved using traditional classification algorithms), but to obtain estimations at aggregated level. This kind of problems are related with those aimed at tracking trends over time [24], such as early detection of epidemics and endangered species, risk prevalence and ecosystems evolution.

In the experimental results section we empirically demonstrate that the estimates provided by a single quantifier can be improved by using its ensemble version. We have compared the performance of our ensemble quantifier approach with a baseline quantifier, *CC* (*Classify and Count*)[25], and two state-of-the-art quantifiers, *AC* (*Adjusted Count*)[25] and *HDy*[26]. Nevertheless, we think that the significance of this article goes beyond that fact, since the proposed approach can be applied to different distribution change problems, as long as it is possible to characterize those changes beforehand. Interestingly, several studies characterizing some of these problems have been published recently [8,9,27,28].

The rest of this paper is organized as follows: Section 2 briefly describes distribution changes and how to characterize them and Section 3 introduces the binary quantification problem and the quantification algorithms used in this paper. In Section 4 the details of our ensemble quantification approach are presented. The experimental setup and empirical results are shown in Section 5. Section 6 summarizes the main conclusions.

2. Characterizing problems with changes in data distribution

A categorization and a discussion about problems presenting distribution changes, or using the current terminology, problems suffering from *dataset shift*, can be found for instance in [8,9]. Supervised learning problems are defined by a set of covariates, x, a class variable, y, and the examples are drawn at random from the joint probability distribution of both. To better understand dataset shift it is important to realize how the data is generated according to the causal relationship between covariates and the class variable, since it determines the kind of changes in the distribution that a problem may suffer from. In this sense, a taxonomy proposed in [29] identifies two types of problems: $\mathcal{X} \rightarrow \mathcal{Y}$ in which the class value y is causally determined by the covariate values x, and problems $\mathcal{Y} \rightarrow \mathcal{X}$ where the covariates x causally depend on the class label y. Spam detection constitutes an example of the first type of problems, the mail content determines whether it is spam or not. On the other hand, medical diagnosis problems are a typical example of the second; suffering from a determined disease y cause a series of symptoms x to appear.

Given an instance x and a class value y, their joint probability $\mathbf{P}(x, y)$ can be written as $\mathbf{P}(y|x)\mathbf{P}(x)$ in $\mathcal{X} \rightarrow \mathcal{Y}$ problems and $\mathbf{P}(x|y)\mathbf{P}(y)$ in the case of $\mathcal{Y} \rightarrow \mathcal{X}$ problems. Dataset shift arises when any of these elements change between training and test, that is to say $\mathbf{P}_{tr}(x, y) \neq \mathbf{P}_{tst}(x, y)$. Thus, several types of dataset shift problems can be identified depending on the elements that change:

- *Covariate shift*: $\mathbf{P}(x)$ changes but $\mathbf{P}(y|x)$ remains constant
- *Prior probability shift*: $\mathbf{P}(y)$ changes but $\mathbf{P}(x|y)$ does not
- *Concept shift* (o *drift*): $\mathbf{P}(y|x)$ changes but $\mathbf{P}(x)$ does not ($\mathcal{X} \rightarrow \mathcal{Y}$ problems), or $\mathbf{P}(x|y)$ changes but $\mathbf{P}(y)$ remains constant ($\mathcal{Y} \rightarrow \mathcal{X}$ problems)

Supervised learning methods generally assume that the joint probability distribution remains unaltered between training and test. However, in practice, there are many important applications suffering from changes to a greater or lesser extent. These kind of problems are interesting from an ensemble learning point of view because some of the aforementioned changes can be easily characterized. This is especially true in the case of *prior probability shift* problems, that are also referred to as *quantification* problems in the literature. A typical application of quantification learning is to estimate the prevalence of positive and negative opinions. Imagine that we want to track the opinions about a product in Twitter during a period of time and give just an estimate on how many are positives (and negatives), without predicting individual opinions (this would be a classification task). In such a problem, when the class distribution $\mathbf{P}(y)$ changes (e.g. the number of positive opinions increases), the opinions maintain the same distribution when the class is fixed. The way in which users express their opinions does not change from one day to another, there would be very good opinions using strong words expressing that felling, moderately positive comments and so on. When can assume in that case that $\mathbf{P}(x|y)$ is constant.

As we state in the previous section, ensembles have been applied before for problems that present a shift in the distribution, mainly in concept drift tasks [30]. The main idea is to build an ensemble with models created at different moments in time and the goal is to have at least one model representing each distinct concept. The ensemble maintains a *memory* of models representing past concepts because some of them may become useful again in the future [31]. Different strategies for training such ensembles can be employed, for instance, to divide historical sequential data into non overlapping blocks [32–34] or using different sized training windows [35–37]. After the set of models is trained, adaptivity is achieved by defining a combination or fusion rule. Basically, the