Pattern Recognition 43 (2010) 3151–3161

Contents lists available at ScienceDirect

Pattern Recognition

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

Transfer estimation of evolving class priors in data stream classification

Zhihao Zhang, Jie Zhou *

Tsinghua National Laboratory for Information Science and Technology (TNList), State Key Laboratory on Intelligent Technology and Systems, Department of Automation, Tsinghua University, Beijing 100084, PR China

ARTICLE INFO

Article history:
Received 29 June 2009
Received in revised form
15 February 2010
Accepted 25 March 2010

Keywords:
Concept drift
Transfer learning
Prior estimation

ABSTRACT

Data stream classification is a hot topic in data mining research. The great challenge is that the class priors may evolve along the data sequence. Algorithms have been proposed to estimate the dynamic class priors and adjust the classifier accordingly. However, the existing algorithms do not perform well on prior estimation due to the lack of samples from the target distribution. Sample size has great effects in parameter estimation and small-sample effects greatly contaminate the estimation performance. In this paper, we propose a novel parameter estimation method called *transfer estimation*. Transfer estimation makes use of samples not only from the target distribution but also from similar distributions. We apply this new estimation method to the existing algorithms and obtain an improved algorithm. Experiments on both synthetic and real data sets show that the improved algorithm outperforms the existing algorithms on both class prior estimation and classification.

1. Introduction

Data stream classification is a hot topic in data mining research [1,2]. It has gained a high attraction due to the important applications in lots of fields, including financial applications, vehicle monitoring, sensor networks, etc. A great challenge for data stream classification is the varying class priors along the data sequence, i.e. the class priors may not be constant. A typical example is vehicle classification. We want to classify vehicles into different types, for example cars and trucks. However, the ratios of vehicles evolve along the time (for instance usually more cars in the daytime and more trucks during night). Classifiers may perform badly when the "assumed" priors of training data are different from the "true" priors of test data. This phenomenon on static data set is called biased class distribution or class imbalance [3–6]. For data streams, the class prior evolution is called virtual concept drift [7], which is mentioned as "concept drift" for convenience in the following text.

Several algorithms have been proposed to deal with concept drift problems [8–13]. Most of the existing algorithms worked in an inductive way. They first train an initial classifier on the training data and then use it to predict the labels of test samples. When the performance deteriorates, they need new labeled samples to retrain the classifier. In real-world applications, labeled samples are often difficult, expensive, or time consuming to obtain, as they require the efforts of experienced human annotators [14]. This problem can be solved by transductive learning. Transductive learning was proposed in Ref. [15], and some new interpretations include [16,17]. Different from inductive learning, transductive learning aims to generate an adjusted classifier for the given test samples. The most straightforward way of adjusting the classifier is to estimate the class priors of test samples and adjust the classifier according to Bayes' rule. Under this framework, the key issue is how to estimate the class priors.

Class prior estimation is an important problem in the fields of information theory and statistics, which is called prior probabilities estimation (PPE) [18–20]. Confusion matrix method is one solution for PPE, which is based on the computation of confusion matrix [21]. It first estimates the confusion matrix on the training set from cross-tabulated classification frequencies provided by the classifier. Then the confusion matrix is used to infer the priori probabilities by solving a system of n linear equations with respect to the estimated class priors. In a recent study, Forman [22] focused on estimating the class priors using an improved form of the commonly used confusion matrix technique. A better prior estimation method is proposed in Refs. [23,24]. This work can be seen as a semi-supervised transductive algorithm. It uses the batch EM(BEM) algorithm to iteratively estimate the class priors on the test set. The converged estimation is used to adjust the classifier. It was proved that the EM learning in BEM algorithm is globally convergent and the (unique) maximum likelihood estimate can be found [25]. It was shown that BEM algorithm is more effective than confusion matrix method [23,24]. These prior estimation methods were all designed for static classification problems and are not suitable for data streams with concept drift, since class priors evolve along the data sequence and should be estimated dynamically. To our best knowledge, the only

* Corresponding author.
E-mail addresses: zhangzh06@mails.tsinghua.edu.cn (Z. Zhang), jzhou@tsinghua.edu.cn (J. Zhou).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.03.021

transductive algorithm that handles concept drift is the OEM algorithm proposed in Ref. [26].

OEM algorithm is an extension of BEM algorithm. When a new test sample comes, OEM algorithm gives an estimate of current class priors based on EM algorithm. The class priors are updated under exponential forgetting rule. Then the classifier is adjusted according to Bayes' law and used to predict the label. The OEM algorithm showed a better performance than BEM algorithm when class priors evolve along data sequences [26]. However, the prior estimation capability of OEM algorithm is insufficient. The reason is that few samples are available from the target distribution due to the prior evolution. Since sample size has great effects in parameter estimation, small-sample effects greatly degrade the performance of systems [27]. In concept drift, the prior evolution leads to the continuous change of distributions. Strictly speaking, only one sample is available from the target distribution, which makes the prior estimation in concept drift a difficult small sample estimation problem. In OEM algorithm, only the current sample is used to estimate the class priors, which leads to the poor performance of prior estimation.

In this paper, we propose a novel parameter estimation method called *transfer estimation*. It is inspired by transfer learning, a novel approach in machine learning community. Transfer learning aims to deal with small sample learning problems by transferring knowledge from related domain to the target domain [28]. Similar with transfer learning, transfer estimation makes uses of samples from not only the target distribution but also related distributions. Thus more information can be used to help improve the estimation performance. We analyze the property of this new estimation method. The theoretical condition is given which ensures transfer estimation a better performance. We apply transfer estimation method to OEM algorithm and get a new transductive algorithm. Experiments on both synthetic data sets and real-world data sets show that the new algorithm outperforms the OEM algorithm.

The rest of this paper is organized as follows. In Section 2, we introduce and analyze the proposed transfer estimation method. This new estimation method is applied to OEM algorithm and we get a new transductive algorithm. In Section 3, experimental results and analysis are reported. Finally, conclusions are given in Section 4.

2. Transfer estimation of class priors

2.1. Problem definition

We begin our discussion by introducing some notations used in this paper. Given an m-class classification problem, $\Omega=\{\omega_1,\ldots,\omega_m\}$ is the set of class labels. The data stream is an ordered set of samples $S=\{(x_1,y_1),\ldots,(x_t,y_t),\ldots\}$ as shown in Fig. 1. x_t is the attribute vector of the t th sample and $y_t\in\Omega$ is the class label, t can be from 1 to infinite. $P_t(x,\omega_i)=P(x|\omega_i)P_t(\omega_i)$ is the joint distribution with $\omega_i\in\Omega$, where $P(x|\omega_i)$ are the within-class probability densities and $P_t(\omega_i)$ are class priors. We assume the within-class probability densities remain unchanged and class priors evolve along data sequence. In the training sequence S_{tr}, y is known. And in the test sequence S_{ts}, y is unknown. The task is to predict the class label y_t of sample $(x_t,y_t)\in S_{ts}$ under the maximum a posteriori (MAP) rule:

$$\hat{y}_t=\arg\max_{\omega_i\in\Omega}P(\omega_i|x_t), \tag{1}$$

where $P(\omega_i|x_t)$ is the a posteriori probability of class ω_i of sample x_t, and $\hat{y}_t$ is the predicted label. Generative classifiers and discriminative classifiers can both give the a posteriori

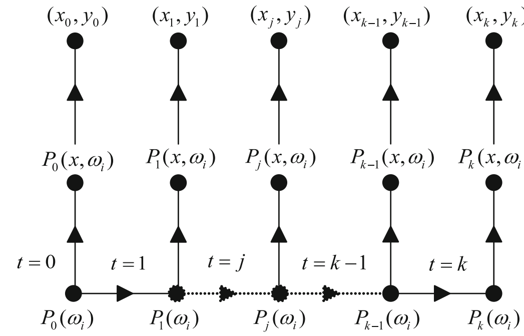

Fig. 1. Data stream with evolving class priors.

probability [29]. Discriminative classifiers directly compute the a posteriori probability. However, generative classifiers first estimate the within-class conditional probabilities and calculate the a posteriori probability using Bayes' Law as

$$P(\omega_i|x)=\frac{P(x|\omega_i)P(\omega_i)}{\sum_{\omega_j\in\Omega}P(x|\omega_j)P_t(\omega_j)}. \tag{2}$$

It is commonly accepted that discriminative classifiers are better than generative classifiers [29]. In concept drift, however, generative classifiers are more suitable since the within-class distributions remain unchanged, which is experimentally validated in Ref. [26]. In this paper, we use generative classifiers as the basic classifiers. Naive Bayes classifier is a typical example of generative classifiers, which makes parametric assumption about the within-class conditional probabilities, for example a normal distribution assumption [30].

To calculate the a posterior probability shown by Eq. (2), the key issue is how to estimate the current class priors $P_t(\omega_i)$, which is a parameter estimation problem. x_t can be seen as drawn from the mixed distribution:

$$P_t(x)=\sum_{\omega_i\in\Omega}P(x|\omega_i)P_t(\omega_i), \tag{3}$$

where class priors $P_t(\omega_i)$ are the unknown parameters to estimate.

2.2. OEM algorithm

The state-of-art work of class prior estimation for static data is the BEM algorithm [23]. Yang et al. [26] proposed the OEM algorithm for evolving class prior estimation, which is an extension of BEM algorithm. The class prior estimation of OEM algorithm can be represented as follows:

$$\begin{cases}\hat{P}_0(\omega_i)=P_{ini}(\omega_i),\\ \hat{P}(\omega_i|x_k)=\dfrac{P(x_k|\omega_i)\hat{P}_{k-1}(\omega_i)}{\sum_{\omega_j\in\Omega}P(x_k|\omega_j)\hat{P}_{k-1}(\omega_j)},\\ \hat{P}_k(\omega_i)=(1-\alpha)\hat{P}_{k-1}(\omega_i)+\alpha\hat{P}(\omega_i|x_k),\end{cases} \tag{4}$$

where $\hat{P}(\omega_i|x_k)$ is the estimated a posteriori probability of k th sample, $\hat{P}_k(\omega_i)$ is the updated class priors and α is the forgetting factor which is set empirically.

At the beginning $t=0$, class priors on the training set are initialized as $\hat{P}_0(\omega_i)=P_{ini}(\omega_i)$. And the initial classifier is trained on the independent training set with $\hat{P}_0(\omega_i)$. Based on EM algorithm, Yang et al. proved that the a posterior probability of class ω_i can be used as the estimator of class prior $P(\omega_i)$. At time $t=k-1$, we have the prior estimation $\hat{P}_{k-1}(\omega_i)$ of $P_{k-1}(\omega_i)$. At time $t=k$ with test sample (x_k,y_k), the a posterior probability estimation $\hat{P}(\omega_i|x)$ for generative classifier is calculated as in the second formula of Eqs. (4). Class priors at time $t=k$ are adjusted under the