Neurocomputing 277 (2018) 149–160

Contents lists available at ScienceDirect

Neurocomputing

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

Two birds with one stone: Classifying positive and unlabeled examples on uncertain data streams

Donghong Han[a,c], Shuoru Li[a], Fulin Wei[a,*], Yuying Tang[a], Feida Zhu[b], Guoren Wang[a,c]

[a] College of Computer Science and Engineering, Northeastern University, Shenyang, China
[b] School of Information Systems, Singapore Management University, Singapore
[c] Key Laboratory of Medical Image Computing (NEU), Ministry of Education, Shenyang, China

ARTICLE INFO

Article history:
Received 25 September 2016
Revised 28 February 2017
Accepted 5 March 2017
Available online 24 August 2017

Keywords:
Uncertain Data Streams
PU learning
Concept drift
Ensemble classifier
Cluster
ELM

ABSTRACT

An important feature characteristic of the data streams in many of today's big data applications is the intrinsic uncertainty, which could happen for both item occurrence and attribute value. While this has already posed great challenges for fundamental data mining tasks such as classification, things are made even more complicated by the fact that completely-labeled examples are usually unavailable in such settings, leaving researchers the only option to learn classifiers on partially-labeled examples on uncertain data streams. Furthermore, there will be concept drift on evolving data streams. To address these challenges, this paper therefore focuses on the study of learning from positive and unlabeled examples (PU) on uncertain data streams. To the best of our knowledge, this paper is the first work to address the uncertainty issue in both item occurrence (occurrence level) and attribute value (attribute level) for the problem of PU learning over streaming data. Firstly, we propose an algorithm to classify positive and unlabeled examples with both uncertainties (PUU). The algorithm extracts reliable positive and negative examples by clustering-based method, and then trains the classifier with Weighted Extreme Learning Machine (Weighted ELM). Secondly, we propose an algorithm of PU learning over uncertain data streams (PUUS). It adopts ensemble model and trains base classification by PUU. In order to detect concept drift, PUUS uses cluster set similarities between the current data block and history data block. We propose different update strategies for different concept drift to adapt to evolving uncertain data streams. Experimental results show that PUUS can effectively classify uncertain data streams with just positive and unlabeled data, while achieving in the meantime good performance in detecting and handling concept drifts.

1. Introduction

Recent years have witnessed the power of big data in driving and fueling intelligence and innovation in a wide range of industries to an unprecedented level [1]. In particular, as a typical example of big data, data streams are widely used in many fields, such as Web traffic statistics, financial analysis, Web application services, etc. Due to transmission error, measurement inaccuracy, sensor malfunction and so on, uncertainty is an intrinsic nature of the data streams in many applications including WSN (Wireless Sensor Networks) [2], and RFID (Radio Frequency Identification) [3,4]. Note that in the context of uncertain data, uncertainty could be observed for both data item occurrence and data attribute values (we would call *occurrence level* for the former and *attribute level* for the latter), compounding the difficulty of streaming data classification with an extra degree. Furthermore, different from traditional static data sets, streaming data distribution is subject to continuous change. It is essential to detect concept drift and update learning model.

In reality, users often focus on only one target category of their interest on uncertain data streams and in general are not concerned with others. For example, for fraud detection or intrusion detection, only the confirmed fraud and the successful intrusion data are to be identified and handled. In these cases, we just need to label a few examples of target category and learn model to predict class label for mass data. Tasks like such – classifying uncertain data streams with positive and unlabeled samples (but not negative examples) – are called PU learning for uncertain streaming data.

However, while uncertainty in both occurrence level and attribute level are often simultaneously observed, there has been till this day little research to consider the two in a unified model for

* Corresponding author.
E-mail addresses: handonghong@cse.neu.edu.cn (D. Han), 1151844742@qq.com (F. Wei).

http://dx.doi.org/10.1016/j.neucom.2017.03.094
0925-2312/

Table 1
Positive and unlabeled examples with uncertainty.

ID	Temperature	Soil fertile ability	Probability	Category	Time
1	$f_{1,2}(x)$, [17,18]	(fFertile, poor)(0.9, 0.1)	0.9	1	t_1
2	$f_{2,2}(x)$, [16,19]	(fFertile, poor)(0.8, 0.2)	0.5	?	t_2
3	$f_{3,2}(x)$, [8,17]	(fFertile, poor)(0.6, 0.4)	0.6	1	t_3
4	$f_{4,2}(x)$, [16,18]	(fFertile, poor)(0.1, 0.9)	0.2	1	t_4
5	$f_{5,2}(x)$, [9,10]	(fFertile, poor)(0.3, 0.7)	0.6	?	t_5

PU learning. Table 1 shows some samples of positive and unlabeled examples with both uncertainties. Columns 2 and 3 represent temperature attribute and soil fertile ability of each sample. The former is uncertain continuous attributes, $f_{i,j}(x)$ representing the probability density function(pdf) of the ith sample at the jth column. The latter is uncertain discrete attribute. There is attribute level uncertainty in both columns. In the first sample, temperature ranges from 17 to 18 °C, and obeys the distribution probability density with $f_{1,2}(x)$. Soil fertile ability values with the probability of fertile or sterile is 0.9 or 0.1, respectively. Column 4 represents occurrence probability of the sample, that is, the occurrence level uncertainty of the sample. Column 5 is the class label of the sample. For a sample, "1" means that it belongs to labeled positive class, "?" represents that the sample's category is unlabeled. Column 6 is the sample arrival time, which can be used to distinguish the data block that the sample belongs to.

Based on the above challenges, this paper focuses on the research of classifying uncertain data streams with models learned from positive and unlabeled examples. To our knowledge, this paper is the first to study PU learning problems on uncertain data streams, handling both occurrence level and attribute level uncertainty at the same time. We summarize our main contributions as follows:

- Based on Weighted ELM [39], we propose a classification algorithm PUU, which learns from positive and unlabeled examples with two kinds of uncertainty. In the proposed algorithm, in order to reduce dimension, the attribute mean and mean deviation are used to represent approximately the attribute level uncertainty. Meanwhile, occurrence probability is used to indicate the occurrence level uncertainty. And we adopt the method based on clustering to extract credible positive and negative examples.
- We then extend PUU into PUUS algorithm to classify uncertain data streams with positive and unlabeled examples. In an ensemble model, we calculate the cluster similarities between current blocks and historical blocks to detect concept drifts. Only when the cluster similarity is greater than the threshold, which is set in accordance with different concept drifts, it adopts different classifier update strategy.
- Experimental results show that the proposed algorithms can deal with the PU learning problem with both uncertainties considered simultaneously, and can effectively detect concept drifts on uncertain data streams.

The remainder of the paper is organized as follows. In Section 2, we introduce related work and ELM theory. In Section 3, our proposed algorithms are presented in detail. The experimental results and the evaluation of the performance are discussed in Section 4. At last, Section 5 concludes the paper.

2. Preliminaries

Broadly, most research work use possible world model as an uncertain data model [5]. If there is the independence assumption, the occurrence level uncertainty means the occurrence probability of the tuple, and when it increases, the number of instances in possible world will increase exponentially. Attribute level uncertainty describes the uncertainty information of each dimension for a tuple. In this paper, we focus on the data model with both uncertainties.

Firstly, we review the previous works on classification algorithms over both precise and uncertain data streams, and PU learning. And then, we present a brief overview of ELM.

2.1. Related work

2.1.1. Classification algorithms over data streams

The data distribution of evolving data streams may change with time. This phenomenon, dubbed concept drift, is the important challenge obviously different from traditional static data model. There are two kinds of concept drift: burst concept drift and progressive concept drift. In order to deal with concept drift, the existing algorithms are divided into two categories including single classifier approaches and ensemble models.

The single classifier approaches mainly use sliding window mechanism to select tuples which are suitable for the current concept to train classifier. In 2000, the very fast decision tree (VFDT) algorithm was proposed to classify data streams [6]. In 2001, the concept-adapting very fast decision tree (CVFDT) presented in [7] aimed at dealing with time-varying concepts. Based on sliding window, CVFDT was updated by removing the low accuracy tree nodes and adding a new sub-tree. In 2004, [8] proposed Ultra-Fast forest tree system (UFFT), in which multiple binary trees compose decision forest and each binary tree corresponds to only two categories. The final prediction result is decided by the votes of all the members. Fan et al. proposed random decision tree (RDT) [9] to learn classifiers without training sets.

The ensemble model is composed of multiple independent base classifiers. As the streaming data arrives continuously, for different data streams, the amount of data adapting to the current window concept is also different. Therefore, the single classifier strategy has some limitations. As one of the classic ensemble-based classification methods, the streaming ensemble algorithm (SEA) was presented in [10]. SEA divides the data stream into different segments, in which each segment trains base classifier respectively. Wang et al. put forward the weighted ensemble classifier (WCE) for data streams classification [11]. When the accuracy of base classifier is lower than the threshold, it will be updated with a new one to adapt to the current data. For training the classifier, not all features are important in high-dimensional data sets. Therefore Nguyen et al. [12] proposed HEFT-Stream algorithm, in which feature extraction was incorporated into a heterogeneous ensemble model to deal with different types of concept drifts. In [13], an incremental weighted function was put forward to evaluate the performance of base classifiers. An on-line weighted ensemble (OWE) regressive model was proposed in [14], which can incrementally learn from a lot of changing tuples and simultaneously retain information appearing in the scene. Sun et al. proposed a class-based ensemble to solve category evolution [15].

The uncertainty of streaming data increases the complexity of classification algorithms. So far, only a few literatures are available. Based on CVFDT [7], Liang et al. proposed the