Applied Soft Computing 47 (2016) 389–394

Contents lists available at ScienceDirect

Applied Soft Computing

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

Incremental Semi-Supervised classification of data streams via self-representative selection

Zhixi Feng, Min Wang*, Shuyuan Yang, Licheng Jiao

Key Lab of National Radar Signal Processing, Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, International Collaboration Joint Lab in Intelligent Perception and Computation, Xidian University, Xi'an, Shaanxi 710071, China

ARTICLE INFO

Article history:
Received 20 April 2015
Received in revised form 29 January 2016
Accepted 12 February 2016
Available online 13 April 2016

Keywords:
Incremental learning
Semi-supervised classification
Self-representative selection
Data streams
Big data

ABSTRACT

Incremental learning has been developed for supervised classification, where knowledge is accumulated incrementally and represented in the learning process. However, labeling sufficient samples in each data chunk is of high cost, and incremental technologies are seldom discussed in the semi-supervised paradigm. In this paper we advance an Incremental Semi-Supervised classification approach via Self-Representative Selection (IS^{3}RS) for data streams classification, by exploring both the labeled and unlabeled dynamic samples. An incremental self-representative data selection strategy is proposed to find the most representative exemplars from the sequential data chunk. These exemplars are incrementally labeled to expand the training set, and accumulate knowledge over time to benefit future prediction. Extensive experimental evaluations on some benchmarks have demonstrated the effectiveness of the proposed framework.

1. Introduction

Data today is more deeply woven into the fabric of our daily lives than ever before due to the rapid improvement of digital technology of storage and information processing. Very recent few years have witnessed an explosive growth of data, where continuously collected data streams accounts for a large and important part [1,2]. From the perspectives of computation and machine intelligence, one should establish a data-driven machine that is capable of incrementally analyzing large-scale dynamic data stream, and accumulating knowledge incrementally over time to benefit future learning and decision-making process [3–11]. Consequently, a machine learning paradigm, Incremental Learning (InLe), is developed where the learning process takes place according to the newly emerged examples [12–21]. Compared with traditional supervised learning, InLe is capable of learning new information from sequential examples to facility the decision-making process. It is very suitable for applications where examples do not always arrive simultaneously, and the newly arriving data may bring a new perspective, may even change the statistical distribution of data. Moreover, from the biological viewpoint, InLe is more consistent with human learning where human beings already use possessed knowledge along with the experiences for learning and decision making.

Nowadays many incremental learning architectures [22,23] and algorithms [12–15,20,21,33] have been developed to deal with data streams, which can be categorized as Absolute Incremental Learning (AInLe) and Selective Incremental Learning (SInLe). In AInLe, new data are analyzed separately, and new features are formed and combined with the existing ones. In SInLe, the selected training set based on the proximity and impact of new data and new information are retrained in light of new information. Most of available InLe approaches are SInLe, which do not assume the availability of a sufficient labeled dataset before the learning, but the training examples appear over time. However, in real-life scenarios, new examples are not always labeled timely. In practical, massive amounts of data are collected dynamically in very rapid mode, resulting in the difficulty of offering labeled samples over time. For example, labeling examples from surveillance and mobile sensor network data streams is infeasible both in time and resource. On the other hand, preparing a sufficiently large number of labeled training samples at the very beginning is practically impossible, for the changing environment where new characteristic of samples or even new kind of samples are generated over time. Consequently, it is necessary to automatically update an existing training set in an incremental fashion to accommodate new information, by adding newly emerged samples to the training set.

Although the classification of data streams are characteristics of scarce labeled examples, enormous number of sequentially incoming samples are available. Because learning from labeled as well as unlabeled data is very useful for incremental learning,

* Corresponding author. Tel.: +86 88202312.
E-mail address: wangmin@xidian.edu.cn (M. Wang).

http://dx.doi.org/10.1016/j.asoc.2016.02.023
1568-4946/© 2016 Published by Elsevier B.V.

semi-supervised learning technologies can be developed by exploiting unlabeled data to modify and refine the classifier or discriminate criteria to improve classification accuracy [24–26]. Different with AInLe and SInLe, Semisupervised Incremental Learning (SSInLe) first builds knowledge base incrementally from the available labeled data. Then with the unlabeled data, SSInLe updates and restructures the knowledge incrementally. Finally it makes decisions about the new instance on the basis of the knowledge base and update the training set.

SSInLe is very important from various real-time learning perspectives, but few works have done on it. In order to explore both the labeled and dynamic unlabeled samples for a more accurate prediction of data streams, in this paper we advance an Incremental Semi-Supervised classification approach via Self-Representative Selection (IS^{3}RS), for data streams classification. In the SSInLe, an important issue is to identify relevant unlabeled data that can be added to the existing training set. In our method, an incremental self-representative data selection strategy is proposed to find the representative exemplars from the sequential data chunk. These exemplars are incrementally labeled to expand the training set, to accumulate knowledge over time to benefit future prediction. Inspired by the representation learning theory [27], we aim to find a subset of data that efficiently describe the entire data set. It assumes that each data in a dataset can be represented as a linear combination of a limited number of exemplars, which is regarded as a compact representation of data set. By adding some initial exemplars to the labeled set, a new training set can be obtained. Then we can acquire the labels of exemplars by co-training technique [28] via self-representation of each data chunk. The most confidently recovered testing data is added into training set to facilitate the learning.

The remained of this paper is organized as follows: In Section 2, the incrementally semi-supervised framework and self-representation are detailed. In Section 3, some experiments are taken on several datasets to validate the efficiency of our proposed method. The configurations, results and discussions of experiments are given. Conclusions and discusses are presented in Section 4.

2. Incremental semi-supervised learning via Self-Representative Selection (IS^{3}RS)

The proposed IS^{3}RS approach is illustrated in Fig. 1, which consists of three phases: self-representative selection, co-training, and finial decision. First each data chunk is self-represented to determine its exemplars. Under the framework of co-training, labels of these exemplars are predicted by the K-nearest neighbor (KNN) classifier. Then the training set is expanded by adding the most confident exemplars together with their predicted labels. Finally, the final classification is performed based on the expanded training set. In the following we describe each step in detail.

2.1. Self-Representative Selection of exemplars

As described in [27,34], the representative training data plays a key role in deciding the performance of learning algorithm. Therefore, learning representative data from vast amount of data is of great importance when building effective classifier or other prediction for data streams. In the data chunk classification, a key factor is whether the learning machine can take advantage of the representative testing data to construct a compact training set. Among various kinds of representative selection methods, sparsity inspired representation learning attracts a lot of interests because of its simple principle and feasibility. Moreover, it does not need to cast any distribution prior on data and present convincing performance. In this paper, we learn exemplars by a self-representation of data, under the assumption that there exist some exemplars, and each data in the dataset can be described as a linear combination of those exemplars. Mathematically, given a data set $\mathbf{X} \in \Re^{D\times N}$ with some D-dimensional data $\mathbf{x}_i$, where D is the dimensionality of data and N is the number of samples in the data set. We would like to select an informative data subset that can represent the whole dataset. Selecting exemplars can be reduced to the following optimization problem,

$$\begin{cases} \min\limits_{S} \left\|\mathbf{X}-\mathbf{XS}\right\|_F^2 \\ s.t. \left\|\mathbf{S}\right\|_{row,0} \le k \end{cases} \tag{1}$$

where $\mathbf{S} \in \Re^{N\times N}$ is the coefficient matrix and $\left\|\mathbf{S}\right\|_{row,0}$ counts the number of nonzero rows of $\mathbf{S}$. In other words, we expect to select at most $k(k \ll N)$ samples in $\mathbf{X}$ that can best represent $\mathbf{X}$. These k informative samples are called as exemplars. This is a self-representation model, where the dictionary is the data set itself. The property makes the obtained exemplars coincide with the actual data point which can be well revealed the whole data set. By minimizing the reconstruction error of each data point as a linear combination of the examples in the dataset and enforcing $\left\|\mathbf{S}\right\|_{0,q} \le k$, $(\|\bullet\|_{0,q}$ norm is defined as $\left\|\mathbf{S}\right\|_{0,q} = \sum_{i=1}^{N} I(\left\|s^i\right\|_q > 0)$,

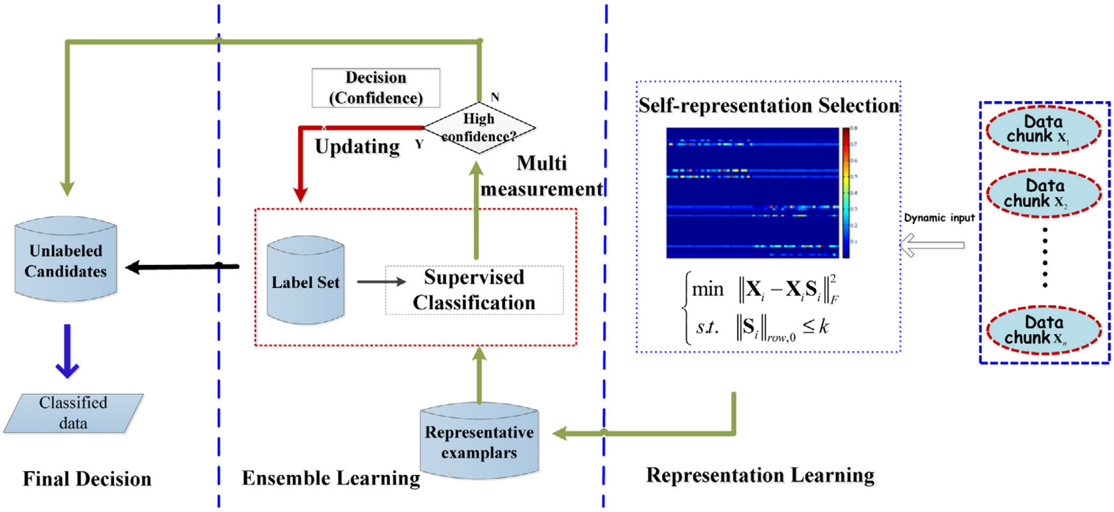

Fig. 1. An illustration of the proposed IS^{3}RS approach.