Information Sciences 466 (2018) 236–257

Contents lists available at ScienceDirect

Information Sciences

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

Distributed online semi-supervised support vector machine

Ying Liu*, Zhen Xu, Chunguang Li

College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, PR China

ARTICLE INFO

Article history:
Received 25 December 2017
Revised 16 July 2018
Accepted 24 July 2018
Available online 26 July 2018

Keywords:
Support vector machine
Semi-supervised learning
Distributed learning
Privacy-preserving
Classification
Manifold regularization
Random approximation
Sparse feature map

ABSTRACT

Recently, the research on semi-supervised support vector machine (S^3VM) has received much attention, and many S^3VM algorithms have been proposed. Existing studies have shown that S^3VM is effective especially in the situations where labeled data is scarce. Nevertheless, most of existing S^3VM algorithms belong to centralized learning, that is, all the data is stored and processed at a fusion center. In many real-world applications, data may be horizontally or vertically distributed over multiple nodes (parties). Besides, from the concerns of privacy and security, each node would not like to share its original data with the others. On the other hand, considering that the data is usually sequentially generated, online processing is preferred. In this paper, we propose two online distributed S^3VM (dS^3VM) algorithms, which are respectively used for horizontally and vertically partitioned data classification. In these two algorithms, to get a fully decentralized implementation, we propose a new form of manifold regularization defined on some anchor points that are adaptively selected by an online strategy. Besides, we use the sparse random feature map to approximate the kernel feature map. In this manner, the model parameters can be collaboratively estimated without transmitting the original data between neighbors. The convergence performances of the proposed algorithms are analyzed. Simulations on several data sets are performed. Results show that the proposed dS^3VM algorithms achieve good classification performance even when there is only a small portion of labeled data.

© 2018 Published by Elsevier Inc.

1. Introduction

Recently, semi-supervised support vector machine (S^3VM) has become one of the most popular machine learning methods, and widely applied to text and image classification [5,30]. Extensive studies have shown that the S^3VM algorithms achieve better performance than the corresponding supervised learning algorithms especially when the labeled data is scarce while the unlabeled data is abundant and can provide valuable information, such as the density function and the manifold regularization of the input space [1,8,15,24]. Nevertheless, most of the existing S^3VM algorithms belong to centralized learning, in which all the training data must be stored and processed at a fusion center.

However, in many practical applications, data may be distributed over different nodes [3,7,13,18]. Generally speaking, according to the distributions of the data, there are two main cases, *horizontally partitioned data* and *vertically partitioned data*, see Fig. 1 [28,31]. For the former case, each node collects/stores a subset of data with the entire attributes. For the latter case, each node collects/stores all the data with partial attributes. In these distributed situations, because of the limited communication resources, centralizing the data to the fusion center to perform the task of classification may be impractical.

* Corresponding author.
E-mail address: yingliu@zju.edu.cn (Y. Liu).

https://doi.org/10.1016/j.ins.2018.07.045
0020-0255/© 2018 Published by Elsevier Inc.

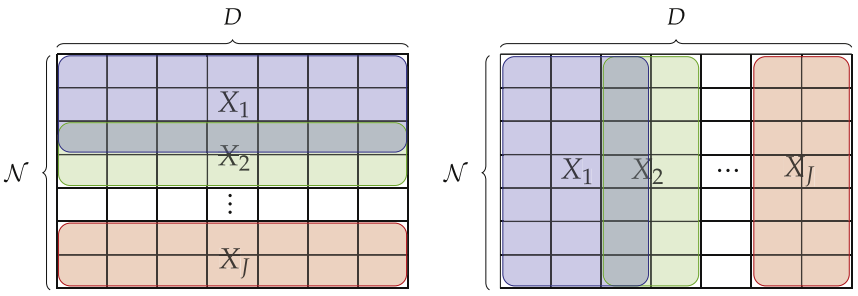

Fig. 1. Pictorial representation of horizontally partitioned data (left) and vertically partitioned data (right). Note that here D and $\mathcal{N}$ denote the total number of attributes and data samples, respectively.

So, there is a great demand for decentralized implementation of S^{3}VM algorithm, in which the global classification problem can be solved at each node distributively based on its local data and the information exchanged from its neighbors.

Lately, two new distributed S^{3}VM (dS^{3}VM) algorithms have been proposed in [18]. In these algorithms, the hinge loss over the unlabeled data has been replaced by an approximation function, and then two distributed gradient descent algorithms have been developed. Since these algorithms use a linear discriminant function, they may be inapplicable to some cases when the boundaries between different classes are complicated. Considering that the real-world data sets are not always linearly separable, the kernel-based nonlinear S^{3}VM is preferred here.

On the other hand, due to the concern of privacy, each node would not like to share its own original data with the others. Such a requirement significantly increases the difficulties in decentralizing kernel-based S^{3}VM. As we all know, the kernel-trick depends on data pairs. If a pair of data comes from two different nodes, it is difficult to compute the kernel function and the manifold regularization without transmitting the original data (attributes) between neighbors.

To achieve secure kernel-based classification over network, some distributed algorithms have been proposed [12,23,27]. For horizontally partitioned data, the original data can be quantized to binary feature [27] or transformed to a linear combination of monomial feature mapped vectors [12]. Then, the kernel matrix can be securely computed using a secure dot product operation. For vertically partitioned data, the global kernel matrix can be merged by the gram matrix of vertically partitioned data through several secure addition/multiplication operations [23]. Nevertheless, these algorithms are supervised learning, and thus they are inefficient to the cases when there is only a small portion of labeled data.

To tackle this problem, a distributed semi-supervised learning (SSL) algorithm targeted for horizontally partitioned data classification has been proposed lately [6]. In this algorithm, a large amount of data patterns among nodes are exchanged in the initial stage, and then the global Euclidean distance matrix is computed using distributed matrix completion. Nevertheless, this algorithm belongs to batch learning. That is, the whole data set must be loaded into the memory in advance and processed in each iteration. So, these algorithms are unsuitable for large-scale data classification due to the high computation cost and memory capacity. Besides, since the current data is sequentially generated even at an unprecedented rate, the classifier must be retrained from scratch once a new data is arrived. This is time and space consuming. In this case, *online* processing that can adaptively deal with streaming data without storing a large number of historical data, is preferred.

In this paper, we consider the problem of distributed classification for streaming data over a network using semi-supervised learning. Besides, to protect the data, the original data of each node should not be disclosed to the others. To tackle this problem, we propose two online dS^{3}VM algorithms, which are respectively used for horizontally and vertically partitioned data classification. The main contributions of this paper are summarized as follows.

- We propose a new form of manifold regularization defined on some anchor points that are adaptively selected using an online strategy. Based on the proposed manifold regularization, we decentralize the global optimization problem, and prove that the distributed formulation is equivalent to the centralized one.
- Two online dS^{3}VM algorithms are developed, which are respectively used for horizontally and vertically partitioned data classification. In these algorithms, we use the sparse random feature map to approximate the kernel feature map such that the model parameters can be adaptively estimated without transmitting the original data. The boundness of the approximation error using sparse random feature map is analyzed theoretically.
- The convergence performances of the proposed two online dS^{3}VMs algorithms are analyzed.

The rest of this paper is organized as follows. In Section 2, some preliminaries are given. In Section 3, a distributed S^{3}VM algorithm for horizontally partitioned data (dS^{3}VM$_h$) is developed, followed by its performance analysis. In Section 4, a distributed S^{3}VM algorithm for vertically partitioned data (dS^{3}VM$_v$) is proposed, and then its performance is analyzed. Experimental results are presented in Section 5. Finally, some conclusions are drawn in Section 6.

Notation: In this paper, the notation $\{\cdot\}$ denotes a set, and $\sharp(A)$ stands for the cardinality of set A. The notation $||\boldsymbol{x}||_p$ stands for the p-norm of a vector $\boldsymbol{x}$. The notation $\lfloor\cdot\rfloor$ denotes the operation to round an element to its nearest integer towards minus infinity. Besides, $[\cdot]_n$ denotes the nth entry of a vector and $[\cdot]_{nk}$ denotes the nkth entry of a matrix. The

Download English Version:

https://daneshyari.com/en/article/6856165

Download Persian Version:

https://daneshyari.com/article/6856165

Daneshyari.com