Proceedings of the 36th Chinese Control Conference
July 26-28, 2017, Dalian, China

Distributed Stochastic Subgrandient-based Design for Support Vector Machines

Yinghui Wang, Peng Lin, Yiguang Hong
1. Key Lab of Systems and Control, Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing 100190, P. R. China,
E-mail: wangyinghui@amss.ac.cn; penglin@amss.ac.cn;yghong@iss.ac.cn

Abstract: This paper develops a distributed stochastic subgrandient-based support vector machine algorithm when training data to train support vector machines are distributed in the network. In this situation, all the data are decentralized stored and unavailable to all agents and each agent has to make its own update based on its computation and communication with neighbors. With mild connectivity conditions, we show the convergence of the proposed algorithm even though the network topology is time-varying. Convergent rate is also given for the proposed algorithm. Moreover, we provide numerical simulations on a real classification training set to illustrate the effectiveness of the fully distributed algorithm.

Key Words: Distributed algorithm, support vector machine, stochastic subgradient, switching topology.

1 Introduction

In recent years, supervised machine learning can address many applications in knowledge discovery, pattern recognition, and data mining [1, 2], due to recent advances in machine learning. The supervised classification problem is of vital importance for machine learning. The support vector machine (SVM), with each applicability, data sparsity, and global optimality, is one of the most popular classification algorithms in many practical fields [3, 4].

Most designs of SVM algorithms are basically centralized. However, the increasing challenges following the large size of networks and decentralized storage makes people tend to investigate SVM training methods quite independent of some centers when each agent only has limited data. One of the effort to deal with the complicated situation is the parallel designs of SVMs [5, 6], where there is a center basically for the job assignment. When the training data set is extremely large, partial SVMs are obtained using small training subsets and combined at a central unit. However, because of the dependence of some centers, these algorithms still have some weak points.

To further solve the problem, distributed support vector machines (DSVMs) are needed to get rid of any centers or central units for the achievement of the tasks. In fact, in the distributed design, the communication bandwidth and energy may ask us to pay more attention to the local computation of each agent and reduce, if possible, the communication workload in large-scale networks, which is not an easy job. There are some efforts for the design distributed SVMs. For example, [7] adopted the alternating direction method of multipliers [8]. The algorithm was proposed for the fixed topologies, which may fail when some link failures occur. Also, [9] and [10] relied on gossip-based stochastic support vectors obtained from local training data sets in the network. From the point of information privacy, it is not so good to exchange rare support vectors.

The motivation of this paper is to study a efficient distributed SVM algorithm in a time-varying communication network. Therefore, we propose and then analyze distributed stochastic subgrandient-based support vector machine algorithm. In this paper, we consider that each agent in the network can have access to local training subsets and can share information with its one-hop neighbours. To be specific, the technical contribution of the paper includes:

- Distributed design for time-varying networks: Different from [7], we propose a distributed stochastic subgrandient-based support vector machine algorithm based on jointly-connected networks to reduce communication workload/cost among agents. Still, compared with [9] and [10], we build our algorithm based on subgradient of the hyperplane rather than support vectors for information privacy.
- Convergence analysis: we prove the convergence of the distributed stochastic subgrandient-based support vector machine algorithm. Convergence rate is also given to the proposed algorithm.

The organization of this paper is as follows. Preliminaries and problem formulation are given in Section 2. Distributed stochastic subgrandient-based support vector machines algorithm is described in Section 3, with scalability and cinvergence analysis. Then the simulation experiment is shown in Section 4. Finally, the conclusions are presented in Section 5.

2 Preliminaries and Problem Formulation

2.1 Preliminaries

In this subsection, some preliminary knowledge related to convex analysis [11] and graph theory [12] is addressed to formulate the distributed support vector machine problem.

For a convex function f over a convex set Ω, whose subgradient at a point x is denoted by $\triangledown f(x)$, the following inequality holds:

$$f(x) \geq f(y) + \langle \triangledown f(y), x - y \rangle, \ \ \forall x, y \in \Omega..$$

Moreover, over the convex set Ω if

$$f(x) \geq f(y) + \langle \triangledown f(y), x - y \rangle + \frac{\sigma}{2}||x - y||^2, \ \ \forall x, y \in \Omega,$$

we say f is σ-strongly convex.

Consider a graph $G = (\mathcal{M}, \mathcal{E}(t))$ as the information sharing topology among M agents in the network. $\mathcal{M} :=$

$\{1, ..., M\}$ represents the agents sets and $\mathcal{E}(t)$ describes the active commnuicating links among agents at time t. If agent q is agent i's one-hop neighbour at time t, which means that agent i can receive information from agent q directly at time t, then there exists a directed edge from q to i and denoted by $(q, i) \in \mathcal{E}$. Denote the one-hop neighbours of agent i at time t as $\mathcal{N}_i(t) = \{q|(q, i) \in \mathcal{E}(t)\}$. Still, $A(t) \in \mathbf{R}^{M\times M}$ is used to describe the communication pattern among the agents at time t, whose elements are defined as follows:

(a) $a_i^i(t) > 0$;
(b) $a_i^q(t) > 0$ for any $(q, i) \in \mathcal{E}(t)$;
(c) $a_i^q(t) = 0$ for any agents q that are not neighbors of agent i.

Moreover, we make the following assumptions on the communication pattern of the network.

Assumption 1. *The graph $G = (\mathcal{M}, \mathcal{E}(t))$ and the weight matrix $A(t)$ satisfy:*

(a) $A(t)$ is doubly stochastic.
(b) For all $i \in \mathcal{M}$, $a_i^i(t) \geq v$ and $a_i^q(t) \geq v$ if $(q, i) \in \mathcal{E}(t)$, where v is a positive scalar.
(c) The graph $(\mathcal{M}, \mathcal{E}(t) \cup \mathcal{E}(t+1) \cup \cdots \cup \mathcal{E}(t+B-1)$ is strongly connected for all $t \geq 0$ and some positive integer B.

Assumption 1 provides a quite general connectivity condition for the distributed multi-agent system, which has been widely used in [13, 14].

2.2 Formulation

Describe the structure of the centralized primal non-separable SVM formulation [15, 16] briefly. Consider the training sets $S_i := \{(x_{ij}, y_{ij}) : j = 1, ..., N_i\}$, where $x_{ij} \in X$ is a $p \times 1$ data vector belonging to the input space $X \subseteq \mathbf{R}^p$ and $y_{ij} \in Y := \{-1, 1\}$ denotes its corresponding class label. Given the local variables w_i^* and b_i^*, the centralized maximum-margin linear discriminant function $h_i^*(x)$ can be described as $h_i^*(x) = w_i^{*\top} x + b_i^*$. Given $\lambda > 0$, we can state the primal sub-SVM problem for agent i as follows:

$$\min_{w_i\in\mathbf{R}^d, b_i\in\mathbf{R}} \frac{1}{2} w_i^\top w_i + \frac{\lambda}{N_i}\sum_{j=1}^{N_i} \max\{1 - c_{ij} - y_{ij}(w_i^\top x_{ij} + b_i), 0\} \qquad (1)$$

where the slack variables c_{ij} account for non-linearly separable training sets.

Every agent $i \in M$ has access to a labeled training set $S_i := \{(x_{ij}, y_{ij}) : j = 1, ..., N\}$ of size N_i. In the distributed fasion, the goal is to find a global maximum-margin linear discriminant function $h(x)$, which enables each agent i to classify any new input vector x to one of the two labels $\{-1, 1\}$ without sending N_i samples to other agents in the network.

To this end, consider adding consensus constraints to force w_i^* and b_i^* to agree across neighboring agents. We present a formulation of the primal sub-SVM problem in (1) in a distributed fasion:

$$\min_{w_i\in\mathbf{R}^d, b_i\in\mathbf{R}} \frac{1}{2}\sum_{i=1}^{M} w_i^\top w_i + \sum_{i=1}^{M}\frac{\lambda}{N_i}\sum_{j=1}^{N_i}\max\{1 - c_{ij} - y_{ij}(w_i^\top x_{ij} + b_i), 0\}$$
$$s.\,t.\ w_i = w_q,\ b_i = b_q,\ i \in \mathcal{M},\ q \in \mathcal{M}. \qquad (2)$$

Problem (2) can be solved in a distributed fashion. Every agent i has ability to optimize (1) and also meet the consensus constraints $w_i = w_q, b_i = b_q$, by communicate only with its one-hop neighbour q. Moreover, Assumption 1 guarantees consensus in neighborhoods N_i for every agent i enables network-wide consensus, whose mathmatical demonstration is given in Theorem 1 .

Define for notation brevity:

$$\begin{cases} \xi_i & = [w_i^\top, b_i]^\top, \\ X_{ij} & = [x_{ij}^\top, 1]^\top, \\ Y_i & = diag([Y_{i1}, \ldots, Y_{iN}]). \end{cases} \qquad (3)$$

With (3), it follows that $w_i = (I_{p+1} - \Pi_{p+1})\xi_i$, where Π_{p+1} is a $(p+1)\times(p+1)$ matrix with zeros everywhere except for the $(p+1, p+1)$-th entry, given by $[\Pi_{p+1}]_{(p+1)(p+1)} = 1$. Thus, problem (2) can be rewritten as

$$\min_{\xi_i\in\mathbf{R}^{p+1}} \frac{1}{2}\sum_{i=1}^{M}\xi_i^\top(I_{p+1} - \Pi_{p+1})\xi_i + \sum_{i=1}^{M}\frac{\lambda}{N_i}\sum_{j=1}^{N_i}\max\{1 - c_{ij} - y_{ij}\xi_i^\top X_{ij}, 0\}$$
$$s.\,t.\ \xi_i = \xi_q,\ i \in \mathcal{M},\ q \in \mathcal{M}. \qquad (4)$$

3 Distributed Stochastic Subgrandient-based Support Vector Machines Algorithm

In order to solve the distributed primal sub-SVM formulation (4), we consider a distributed stochastic subgrandient-based support vector machines algorithm. Denote $\Omega = \{||\xi||_2 \leq R\}$ as a bounded closed convex set in $\mathbf{R}^{p+1}$. Consider the following optimization problem

$$min_{\xi\in\Omega}\ \ F(\xi) = \sum_{i=1}^{M} F_i(\xi) = \sum_{i=1}^{M}(f_i(\xi) + \frac{\lambda}{N_i}\sum_{j=1}^{N_i} g_{ij}(\xi))$$
$$where\quad f_i(\xi) = \frac{1}{2}\xi^\top(I_{p+1} - \Pi_{p+1})\xi,$$
$$g_{ij}(\xi) = \max\{1 - c_{ij} - y_{ij}\xi_i^\top X_{ij}, 0\}. \qquad (5)$$

Remark 1. *We assume $F_i(\xi)$ to be strongly convex over $\Omega = \{||\xi||_2 \leq R\}$ according to [4]. Therefore, we have that for all $\xi \in \Omega$, $||\nabla f_i(\xi)|| \leq M_f$ and $||\nabla g_i(\xi)|| \leq M_g$. Still, for function $g_{ij}(\xi)$, we assume that there exists an $\iota > 0$ such that* $\min_{g_i(\xi)=0} ||\nabla g(x)|| \geq \iota$.

Note that each F_i is available to agent i only. Suppose $\nabla F_{ij}(\xi)$ is the subgradient of $\frac{1}{2}\xi^\top(I_{p+1} - \Pi_{p+1})\xi + \lambda\max\{1 - c_{ij} - y_{ij}X_{ij}\xi, 0\}$. We propose a distributed stochastic subgradient-based algorithm [17] to solve problem (5). This leads to the construction and also the convergent performance of Algorithm 1.

Denote $\tilde{\xi}_i(T) = P_\Omega(\hat{\xi}_i(T))$, where $\hat{\xi}_i(T) = \frac{1}{T}\sum_{t=1}^{\top}\xi_i(t)$. Next, we' give the convergent property

Download English Version:

https://daneshyari.com/en/article/5505072

Download Persian Version:

https://daneshyari.com/article/5505072

Daneshyari.com