Distributed Transfer Linear Support Vector Machines

Rui Zhang
Department of Electrical and Computer Engineering
New York University
Brooklyn, NY, 11201
Email: rz885@nyu.edu

Quanyan Zhu
Department of Electrical and Computer Engineering
New York University
Brooklyn, NY, 11201
Email: qz494@nyu.edu

Abstract—Transfer learning has been developed to improve the performances of different but related tasks in machine learning. However, such processes become less efficient with the increase of the size of training data and the number of tasks. Moreover, privacy can be violated as some tasks may contain sensitive and private data, which are communicated between nodes and tasks. We propose a consensus-based distributed transfer learning framework, where several tasks aim to find the best linear support vector machine (SVM) classifiers in a distributed network. With alternating direction method of multipliers, tasks can achieve better classification accuracies more efficiently and privately, as each node and each task train with their own data, and only decision variables are transferred between different tasks and nodes. Numerical experiments on MNIST datasets show that the knowledge transferred from the source tasks can be used to decrease the risks of the target tasks that lack training data or have unbalanced training labels. We also show that the target tasks can enter and leave in real-time without rerunning the whole algorithm.

Index Terms—Transfer Learning, Multi-Task Learning, Distributed Learning, Support Vector Machines

I. INTRODUCTION

Machine learning algorithms are largely used nowadays in various areas, e.g., face detection [1] and search engines [2]. Traditionally, machine learning makes predictions or classifications based on the assumption that the training and the testing data come from the same source or distribution [3]. However, this assumption may not hold in many real applications [4]; for example, the training data can be outdated, or insufficient to build a good classifier. In such cases, it is difficult to find the classifier using traditional machine learning frameworks.

Recent researches on transfer learning provide a solution to address such problems. It has been shown that machine learning tasks can benefit from other similar tasks by knowledge transfer [3], [4]. For instance, web-page data can become outdated easily as the web content changes frequently, and new training data are expensive to acquire as the labeling of the data is costly. Since parts of the outdated data still contain useful information, knowledge can be transferred from them to train a classifier together with the new data [5].

Although the knowledge transfer can improve the performance of machine learning, the training process using a large amount of data is often not efficient. For traditional transfer learning, training data are communicated between tasks [6]. The direct data sharing is not possible when the volume of the data is huge and they contain private information. For example, training data may come from different nodes of a wireless sensor network (WSN), and their communication with a fusion center can be either costly or restricted due to scalability, privacy or power limitations [7].

This paper aims to address this issue by extending transfer learning into a distributed framework in the context of support vector machines (SVMs) illustrated in Fig. 1. The framework trains different but related tasks together with linear SVMs at each node in a fully distributed network. The decision variables to classify testing data are found by minimizing the regularized errors of training data of each task. One set of consensus constraints is introduced to force all the tasks to share the same terms of decision variables at each node while another set of consensus constraints is used to force all the nodes to share the same decision variables of each task. With alternating direction method of multipliers (ADMoM) [8], the centralized problem can be solved in a fully distributed way. Each task at a node shares its decision variables with the same task in the neighboring nodes and other tasks in the same node. As a result, the classification accuracy of each task in each node can be improved without sharing local and private data.

The consensus-based distributed framework provides a way to address distributed transfer learning problems in connected
networks. Since each task at a node makes decisions using its local data, the training process becomes more efficient and scalable. Allowing tasks and nodes to communicate their decision variables with others, we can achieve more accurate classifications without sharing private data between different tasks and different nodes, which effectively reduces the communication overhead and maintains privacy at the same time.

Note that the problem of transfer learning between tasks in one node can be viewed as a transfer learning problem studied in [6]. Besides, the problem of distributed machine learning with a single task is a distributed support vector machines (DSVM) problem recently studied in [7].

The proposed framework is a generalization of both centralized transfer learning scheme and distributed machine learning. It provides a large-scale transfer learning framework where each task transfers knowledge to other tasks and each node transfers knowledge to its neighboring nodes. Performances of all the tasks in each node are illustrated in terms of their training efficiency and data privacy.

**Related Work.** Our research is closely related to multi-task learning [6, 9], transfer learning [10], and distributed machine learning [7, 11, 12]. There is a strong connection between multi-task learning and transfer learning. In multi-task learning, both the source and target tasks are learned simultaneously and perfectly, while transfer learning only aims at improving the performance of the target tasks.

In this paper, we consider multi-task learning as a scenario of transfer learning, since it can be easily extended to transfer learning by assigning a larger weight to the loss function of the target tasks [4].

Current research on transfer learning has focused on solving centralized problems. Both [6] and [9] have focused on multi-task learning, and they have shown that multi-task learning outperforms single-task learning. Raina et al. have presented in [10] a self-taught learning framework where the knowledge from unlabeled data is used to improve the performance of a given classification task. However, these centralized transfer learning approaches require global collection and processing of all the training data, which not only takes more time and space for training, but also makes sensitive data easily accessible by unwanted tasks.

Distributed machine learning aims at large-scale problems and networked problems for traditional machine learning. Both [11] and [7] have presented approaches on DSVMs where information, such as support vectors and decision variables, is communicated between nodes to achieve a better global performance more efficiently.

**Main Contributions.** In this work, we extend transfer learning into a distributed framework to maintain the performance and efficiency of training different but related tasks in a network. Our contributions can be summarized as follows:

- We design a SVM-based multi-task transfer learning to force each task sharing information with other tasks. We then extend it into a distributed framework using consensus constraints in a networked system where several tasks are trained together.
- We develop a distributed transfer support vector machine (DTSVM) algorithm with Alternating Direction Method of Multipliers (ADMoM) [8]. In this algorithm, each task in a node processes their own data using information from other tasks in the same node and the same task in neighboring nodes. Notice that training and testing data are not shared between tasks and nodes, which not only reduces the communication overhead but also maintains privacy.
- We demonstrate with numerical experiments that the DTSVM algorithm can improve the performance of the target tasks that contain limited training data or unbalanced training labels. We also show that our algorithm can improve the performance of the nodes without using the data from the source tasks by using information sent from the nodes that contain the data of the source tasks. We further demonstrate that our algorithm is suitable for online learning where the target tasks can freely enter or leave the training of the source tasks in real time.

The rest of this paper is organized as follows. Section 2 presents a consensus-based centralized transfer learning approach on SVMs. Section 3 outlines the extended distributed transfer support vector machines (DTSVM). Section 4 and 5 present numerical results and concluding remarks, respectively.

**Notations.** Boldface letters represent matrices (column vectors); \((\cdot)^T\) denotes matrix and vector transposition; \(\|\cdot\|\) denotes the norm of the matrix or vector; \(\text{diag}(y)\) denotes the diagonal matrix with \(y\) on its main diagonal; \(\mathcal{V}\) denotes the set of nodes in a network; \(\mathcal{F}_c\) denotes the set of neighboring nodes of node \(v\); \(\mathcal{F}\) denotes the set of tasks.

### II. CENTRALIZED TRANSFER LEARNING

In this section, we present a centralized transfer learning approach on SVMs. Consider \(T\) learning tasks with \(\mathcal{F} = \{1,...,T\}\) denotes the set of tasks. We assume that each task \(t\) has a labeled training set \(\mathcal{D}_t = \{(x_n,y_n) | x_n \in \mathcal{X}_t, y_n \in \{-1,+1\}\}_{n=1}^N\), where \(\mathcal{X}_t \subset \mathbb{R}^p\) represents the input space of task \(t\). Note that \(\mathcal{F}_t\) is different for each task, but has the same dimension \(p\). For each task, a linear SVM aims to find a maximum-margin discriminant function \(g_t(x) = \text{sign}\left(\mathbf{w}_t^T \mathbf{x} + \mathbf{b}_t^*\right)\), which gives input testing data \(x\) a label \(-1\) or \(+1\). Decision variables \(\{\mathbf{w}_t^*, \mathbf{b}_t^*\}\) can be found by solving the following minimization problem \([13]\):

\[
\begin{align*}
\min_{\mathbf{w}_t, \mathbf{b}_t|\xi_{tn}} & \quad \frac{1}{2} \|\mathbf{w}_t\|^2 + C \sum_{n=1}^{N_t} \xi_{tn} \\
\text{s.t.} & \quad y_{tn}(\mathbf{w}_t^T \mathbf{x}_n + \mathbf{b}_t) \geq 1 - \xi_{tn}; \\
& \quad \xi_{tn} \geq 0.
\end{align*}
\]

Note that, \(\xi_{tn}\) is the slack variable, which accounts for non-separable case. Problem \([1]\) is a traditional SVM problem for single task learning. With the assumption that different tasks are related to each other on the basis of similarity between distributions of samples \(\mathcal{F}_t [4]\), the decision variables \(\mathbf{w}_t, \mathbf{b}_t\) can be divided into: \(\mathbf{w}_t = \mathbf{w}_0 + \mathbf{w}_t; \mathbf{b}_t = \mathbf{b}_0 + \mathbf{b}_t\), where \(\mathbf{w}_0\) and...
\[ b_i \text{ are common terms over all tasks, while } w_i \text{ and } b_i \text{ are task specific terms} [4], [6]. \] We further write the decision variables as:
\[ \hat{w}_t = w_{0t} + w_t; \hat{b}_t = b_{0t} + b_t, \tag{2} \]
with \( w_{01} = \ldots = w_{0T} \) and \( b_{01} = \ldots = b_{0T} \) forcing all common terms to agree with each other among all tasks. Thus, a consensus-based centralized approach of multi-task transfer learning can be formulated as the following problem:
\[ \min_{\{w_t, b_{0t}, \hat{w}_t, \hat{b}_t; \xi_{tn}\}} \frac{1}{2} \sum_{t \in T} \| w_{0t} \|^2 \]
\[ + \frac{\gamma}{2} \sum_{t \in T} \| w_t \|^2 + TC \sum_{t \in T} \sum_{n=1}^N \xi_{tn} \]
\[ \text{s.t.} \]
\[ y_{tn}(\hat{w}_t^T x_n + \hat{b}_t) \geq 1 - \xi_{tn}, \quad \forall t \in T; \tag{3i} \]
\[ \xi_{tn} \geq 0, \quad \forall t \in T; \tag{3ii} \]
\[ w_{0t} = w_{0t} + b_{0t} = 0, \quad \forall t, s \in T, s \neq t. \tag{3iii} \]

Note that, consensus constraints \( \tag{3i} \) are used to restrict common terms. \( \xi_1 \) and \( \xi_2 \) are positive regularization parameters, which determine how much \( \hat{w}_t \) differs in each task by controlling the size of \( w_{0t} \) and \( w_t \). When \( \xi_1 \) is large, \( w_{0t} \) tends to be equal to 0, which makes all tasks unrelated. On the other hand, when \( \xi_2 \) is small, \( w_t \) tends to be equal to 0, which makes all tasks find the same classifier.

By solving Problem \( \tag{3} \), we can find the decision variables \( \hat{w}_t \) and \( \hat{b}_t \) simultaneously with information transferred through consensus constraints \( \tag{3i} \) and common terms \( w_{0t} \) and \( b_{0t} \). Problem \( \tag{3} \) provides a centralized framework to transfer learning. In the following section, we further extend it to a distributed network.

### III. Distributed Transfer Learning

Consider a network with \( \mathcal{V} = \{1, \ldots, V\} \) representing the set of nodes. Node \( v \in \mathcal{V} \) only communicates with its neighboring nodes \( \mathcal{B}_v \subseteq \mathcal{V} \). Without loss of generality, we assume that any two nodes in this network are connected by a path, i.e., there is no isolated node in this network. At each node \( v \), \( T \) labeled training sets \( \mathcal{D}_v = \{(x_{vn}, y_{vn})|x_{vn} \in \mathcal{R}, \forall n \in \{-1, +1\}\}^N_{n=1} \) of size \( N_v \) are available for each task \( t \in T \) (e.g., see Fig. 1).

The maximum-margin linear discriminant function at every node \( v \in \mathcal{V} \) for each task \( t \in T \) can be described as \( g_{vt}(x_v) = x_v^T \hat{w}_v + \hat{b}_v \), where decision variables \( \hat{w}_v = w_{0vt} + w_{vt} \) and \( \hat{b}_v = b_{0vt} + b_{vt} \). Note that there are two sets of consensus constraints, \( w_{0vt} = \ldots = w_{0VT} = \ldots = w_{0V} = 0 \) and \( b_{0vt} = \ldots = b_{0VT} = \ldots = b_{0V} = 0 \) are used to force all common terms of decision variables to agree with each other among all the nodes and all the tasks, while \( w_{vt} = \ldots = w_{vT} \) and \( b_{vt} = \ldots = b_{vT} \) are used to force all decision variables \( \{w_{vt}, b_{vt}\}_{v \in \mathcal{V}} \) of task \( t \) to agree with each other among all the nodes. This approach enables each task \( t \) at each node \( v \) to classify any new input \( x_v \) to one of the two classes \( \{+1, -1\} \) without communicating \( \mathcal{D}_v \) to other nodes \( v' \neq v \).

The discriminant function \( g_{vt}(x_v) \) can be obtained by solving the following optimization problem:
\[ \min_{\{w_{0vt}, w_{vt}, b_{0vt}, \xi_{tvn}\}} \frac{1}{2} \sum_{v \in \mathcal{V}} \sum_{t \in T} \| w_{0vt} \|^2 \]
\[ + \frac{\gamma}{2} \sum_{v \in \mathcal{V}} \sum_{t \in T} \| w_{vt} \|^2 + VTC \sum_{v \in \mathcal{V}} \sum_{t \in T} \sum_{n=1}^N \xi_{tvn} \]
\[ \text{s.t.} \]
\[ y_{tvn}(w_{0vt}^T x_v + b_{0vt}) \geq 1 - \xi_{tvn}, \quad \forall v \in \mathcal{V}, t \in T; \tag{4i} \]
\[ \xi_{tvn} \geq 0, \quad \forall v \in \mathcal{V}, t \in T; \tag{4ii} \]
\[ w_{0vt} = w_{0vt} + b_{0vt} = 0, \quad \forall v \in \mathcal{V}, t \in T, u \in \mathcal{B}_v. \tag{4iii} \]

In the above problem, the third and the fourth constraints impose the consensus on the common terms \( w_{0vt} \) and \( b_{0vt} \) at every node \( v \) for each task \( t \), while the fourth and the fifth constraints impose the consensus on decision variables \( \hat{w}_v := w_{0vt} + w_{vt} \) and \( b_v := b_{0vt} + b_{vt} \) across neighboring nodes for each task \( t \).

To solve Problem \( \tag{4} \), we first define the vector of decision variables \( \mathbf{r}_v := [w_{0vt}, w_{vt}, b_{0vt}, b_{vt}]^T \), the augmented matrix \( \mathbf{X}_v := [(x_{1v}, \ldots, x_{Nv}), \mathbf{1}_v] \), the diagonal matrix \( \mathbf{Y}_v := \text{diag}[(y_{1v}, \ldots, y_{Nv})] \), and the vector of slack variables \( \mathbf{\xi}_v := [\xi_{t1v}, \ldots, \xi_{tNv}]^T \). With these definitions, it follows readily that \( w_{0vt} = \mathbf{1}_v \mathbf{r}_v \), and \( \hat{w}_v = \mathbf{1}_v \mathbf{r}_v \), where \( \mathbf{1}_v := [1, \ldots, 0]^T \), \( \mathbf{0}_v := [0, \ldots, 1]^T \), \( \mathbf{1}_v := [1, \ldots, 1]^T \), \( \mathbf{0}_v := [0, \ldots, 0]^T \), and \( \mathbf{M}_v := \mathbf{1}_v \mathbf{1}_v^T \).

Problem \( \tag{4} \) can be rewritten as
\[ \min_{\{\mathbf{r}_v, \mathbf{\xi}_v, \mathbf{\Phi}_v, \mathbf{\omega}_v\}} \frac{1}{2} \sum_{v \in \mathcal{V}} \sum_{t \in T} r_{vt}^T \mathbf{M}_v \mathbf{r}_vt \]
\[ + \frac{\gamma}{2} \sum_{v \in \mathcal{V}} \sum_{t \in T} r_{vt}^T \mathbf{M}_v \mathbf{r}_vt + VTC \sum_{v \in \mathcal{V}} \sum_{t \in T} \xi_{tvn} \]
\[ \text{s.t.} \]
\[ Y_v \mathbf{X}_v [\mathbf{1}_v \mathbf{r}_v] \geq 1 - \mathbf{\xi}_v, \quad \forall v \in \mathcal{V}, t \in T; \tag{5i} \]
\[ \mathbf{\xi}_v \geq \mathbf{0}, \quad \forall v \in \mathcal{V}, t \in T; \tag{5ii} \]
\[ \mathbf{1}_v \mathbf{r}_v = \mathbf{\Phi}_v \mathbf{\Phi}_v, \quad \mathbf{\Phi}_v = [\mathbf{1}_v, \mathbf{0}_v], \quad \forall v \in \mathcal{V}, t \in T, s \neq t; \tag{5iii} \]
\[ \mathbf{r}_v = \mathbf{\omega}_v \mathbf{u}_v, \quad \mathbf{\omega}_v = \mathbf{u}_v, \quad \forall v \in \mathcal{V}, t \in T, u \in \mathcal{B}_v. \tag{5iv} \]

Problem \( \tag{5} \) can be solved iteratively in a distributed way with ADMoM [8], which is shown as the following proposition.

**Proposition 1.** With \( \lambda_v^{(0)} = 0_{(p+1) \times 1} \) and \( \beta_v^{(0)} = 0_{(2p+2) \times 1} \), Problem \( \tag{5} \) can be solved by the following iterations:
\[ \lambda_{v_{(k+1)}} \in \mathcal{D} \max_{0_{\lambda_v} \leq \lambda_v \leq VTC \lambda_v} -\frac{1}{2} \lambda_v^T Y_v \mathbf{X}_v [\mathbf{1}_v \mathbf{u}_v \mathbf{1}_v]^T Y_v^T \mathbf{X}_v \lambda_v + (\mathbf{1}_v + Y_v \mathbf{X}_v [\mathbf{1}_v \mathbf{u}_v \mathbf{1}_v]^T)^T_{\lambda_v} \]
\[ \mathbf{r}_{v_{(k+1)}} = U_v^{-1} \left( \mathbf{1}_v \mathbf{1}_v^T Y_v \mathbf{X}_v \lambda_{v_{(k+1)}} - \mathbf{r}_{v_{(k)}} \right). \]
In this section, we present numerical experiments of DTSVM. We use the MNIST database of handwritten digits to evaluate the distributed transfer learning algorithm. The MNIST database contains images of digit "0" to "9", here we set classifying "3" and "9" as Task 1, classifying "7" and "9" as Task 2, and classifying "8" and "9" as Task 3. Note that, when the data is trained using DTSVM, the performance of transfer learning improves the performances of the tasks.

IV. NUMERICAL EXPERIMENTS

Fig. 6 shows the results when the training data of the target task is limited and has unbalanced labels. We can see that transfer learning can also improve the classification accuracy of these cases. Note that when there are only 2 classes in the target task, Task 1 is limited and has unbalanced labels. We can see that the classification accuracy of these cases is lower than the risks of both DSV and DSVM.

Proposition 1 illustrates the iterations of distributed transfer learning (DTSVM). It is a fully distributed support vector machines (DSVM) and distributed support vector machines (CSVM) and centralized support vector machines (CSVM). The algorithm of CSVM can also improve the information from the nodes with DSV.

\[ g(x) = \sum_{i=1}^{n} \alpha_i y_i \langle x, s_i \rangle + \sum_{i=1}^{n} \beta_i y_i \langle x, v_i \rangle \]

Fig. 7 and Fig. 8 show the results of Task 1 and Task 2. The left figure and the right figure show the results of Task 1 and Task 3. Both Task 1 and Task 3 have 1800 training samples, and Task 3 has 300 training samples. The left figure shows the results when the data is trained using DTSVM and DSVM, while the right figure shows the results when the data is trained using DTSVM and CSVM.
from the source task will train with DTSVM, while nodes who lack that will train with DSVM. We can see that nodes with DTSVM have lower risks. Moreover, nodes with DSVM also have lower risks as they receive information from nodes with DTSVM. This experiment shows that the performances of the nodes who lack training data from the source tasks can be improved with the knowledge transferred from the nodes who contain that data. Fig. 7 shows the results of online transfer learning. Task 1 and Task 2 are the target tasks whose risks we aim to reduce, while Task 3 is the source task that can be used to improve the performances of the target tasks. At different stages, Task 1 and Task 2 will enter or leave the DTSVM algorithm with Task 3. Both Task 1 and Task 2 have better performances after training with Task 3. This experiment shows that our DTSVM algorithm can work online without rerunning the whole system.

V. CONCLUSION

In this paper, we have extended a centralized SVM-based transfer learning into a distributed framework. By using ADMoM, we have developed a fully distributed algorithm (DTSVM) where each task in each node operates their own data without transferring training data to other tasks and neighboring nodes. Numerical experiments have shown that our DTSVM algorithm can improve the performances of the target tasks that lack training data or have unbalanced training labels. We have also shown that our algorithm can improve the
performances of the nodes who lack the data from the source tasks, by sending information from the nodes who contain the data from the source tasks. We have demonstrated that our algorithm is suitable for online learning where the target tasks can freely enter or leave the training of the source tasks in real-time. One direction of future works is to extend the current framework to nonlinear algorithms and other machine learning algorithms.

**APPENDIX A**

Problem 5 can be solved in a distributed way with AD-MoM [8], which solves the following problem:

$$\min_{\{r, \omega\}} F_1(r) + F_2(\omega)$$

subject to:

$$Mr = \omega$$

with the following iterations:

$$r^{(k+1)} = \arg \min_{r} F_1(r) + \alpha^{(k)}T \|Mr - \omega^{(k)}\|^2 + \frac{\eta}{2} \|Mr - \omega^{(k)}\|^2,$$

$$\omega^{(k+1)} = \arg \min_{\omega} F_2(\omega) - \alpha^{(k)}T \|Mr - \omega^{(k)}\|^2 + \frac{\eta}{2} \|Mr - \omega^{(k)}\|^2,$$

where \(\alpha\) denotes the Lagrange multiplier corresponding to the constraint \(Mr = \omega\).

We follow a similar step in [7], by setting

$$r = [r_{ij}; \ldots; r_{ij}; r_{ij}; \ldots; r_{ij}]$$

and

$$\omega = \{\{\phi_{st}\}_{t \in \mathcal{F}, s \in \mathcal{F}, s \neq t}; \ldots; \{\phi_{st}\}_{t \in \mathcal{F}, s \in \mathcal{F}, s \neq t}; \{\phi_{st}\}_{t \in \mathcal{F}, s \in \mathcal{F}, s \neq t}; \ldots; \{\phi_{st}\}_{t \in \mathcal{F}, s \in \mathcal{F}, s \neq t}\}.$$

Problem 5 can be transformed into the form of (13), and thus be solved by Iterations 14-16. By splitting each iteration into sub-problems and further simplifications, distributed iterations of solving problem 5 can be summarized into the following lemma.

**Lemma 1.** Problem 5 can be solved by the following iterations:

$$\{r_{vt}^{(k+1)}, \xi_{vy}^{(k+1)}\} \in \arg \min_{r_{vt}, \xi_{vy}} \mathcal{L}(r_{vt}, \xi_{vy}, \phi_{st}, \omega_{vt}, \alpha_{st,d}, \beta_{vt})$$

$$\phi_{st}^{(k+1)} \in \arg \min_{\phi_{st}} \mathcal{L}(r_{vt}^{(k)}, \xi_{vy}^{(k)}, \phi_{st}, \omega_{vt}, \alpha_{st,d}, \beta_{vt})$$

$$\omega_{st}^{(k+1)} \in \arg \min_{\omega_{st}} \mathcal{L}(r_{vt}^{(k)}, \xi_{vy}^{(k)}, \phi_{st}, \omega_{vt}, \alpha_{st,d}, \beta_{vt})$$

$$\alpha_{st,1}^{(k+1)} = \alpha_{st,1}^{(k)} + \eta_{1}(\{1, 0\}r_{vt}^{(k+1)} - \phi_{st}^{(k+1)}),$$

$$\alpha_{st,2}^{(k+1)} = \alpha_{st,2}^{(k)} + \eta_{2}(\{1, 0\}r_{vt}^{(k+1)} - \phi_{st}^{(k+1)}),$$

$$\beta_{vt,1}^{(k+1)} = \beta_{vt,1}^{(k)} + \eta_{2}(r_{vt}^{(k+1)} - \omega_{vt}^{(k+1)}),$$

$$\beta_{vt,2}^{(k+1)} = \beta_{vt,2}^{(k)} + \eta_{2}(\omega_{vt}^{(k+1)} - r_{vt}^{(k+1)}),$$

where

$$\mathcal{L}(r_{vt}, \xi_{vy}, \phi_{st}, \omega_{vt}, \alpha_{st,d}, \beta_{vt}) = \frac{\eta}{2} \sum_{v \in \mathcal{V}} \sum_{T \in \mathcal{F}} r_{vt}^2M_1r_{vt}$$

$$+ \frac{\eta}{2} \sum_{v \in \mathcal{V}} \sum_{T \in \mathcal{F}} r_{vt}^2M_2r_{vt} + VT C \sum_{v \in \mathcal{V}} \sum_{T \in \mathcal{F}} \xi_{vy}$$

$$+ \sum_{v \in \mathcal{V}} \sum_{T \in \mathcal{F}} \sum_{s \in \mathcal{F}, s \neq T} \left\{\alpha_{st,1}^{(k)}[\{1, 0\}r_{vt} - \phi_{st}]ight\}$$

$$+ \sum_{v \in \mathcal{V}} \sum_{T \in \mathcal{F}} \sum_{s \in \mathcal{F}, s \neq T} \left\{\alpha_{st,2}^{(k)}(\phi_{st} - \{1, 0\}r_{vt})\right\}$$

$$+ \sum_{v \in \mathcal{V}} \sum_{T \in \mathcal{F}} \sum_{s \in \mathcal{F}, s \neq T} \left\{\beta_{vt,1}^{(k)}(r_{vt} - \omega_{vt}) + \beta_{vt,2}^{(k)}(\omega_{vt} - r_{vt})\right\}$$

$$+ \frac{\eta}{2} \sum_{v \in \mathcal{V}} \sum_{w \in \mathcal{B} \in \mathcal{V}} \sum_{u \in \mathcal{U}, u \neq \emptyset} \left\{\|\{0, 1\}r_{vt} - \phi_{st}\|_2^2 + \|\phi_{st} - \{0, 1\}r_{vt}\|_2^2\right\}.$$

Setting initial conditions \(\alpha_{st,1}^{(0)} = \alpha_{st,2}^{(0)} = 0_{(p+1) \times 1}\) and \(\beta_{vt,1}^{(0)} = \beta_{vt,2}^{(0)} = 0_{(2p+2) \times 1}\), we have \(\alpha_{st,1}^{(k)} = \alpha_{st,2}^{(k)}\) and \(\beta_{vt,1}^{(k)} = \beta_{vt,2}^{(k)}\) for \(k \geq 0\). We further define \(\alpha_{st} = \sum_{s \in \mathcal{F}, s \neq t} \alpha_{st}\) and \(\beta_{tv} = \sum_{v \in \mathcal{B} \in \mathcal{V}} \beta_{vt}\). Note that, \(\phi_{st} = \frac{1}{2}\{0, 1\}(r_{vt} + r_{vs})\), and \(\omega_{vt} = \frac{1}{2}(r_{vt} + r_{vs})\), which can be solved directly from (18) and (19). With further simplification, iterations 17-23 can be simplified as the following lemma.

**Lemma 2.** With \(\alpha_{st}^{(0)} = 0_{(p+1) \times 1}\) and \(\beta_{vt}^{(0)} = 0_{(2p+2) \times 1}\), iterations 17-23 can be reduced into the following iterations:

$$\{r_{vt}^{(k+1)}, \xi_{vy}^{(k+1)}\} \in \arg \min_{r_{vt}, \xi_{vy}} \mathcal{L}^{\prime}(r_{vt}, \xi_{vy}, \alpha_{st}, \beta_{vt})$$

$$\alpha_{st}^{(k+1)} = \alpha_{st}^{(k)} + \frac{\eta}{2} \sum_{s \in \mathcal{F}, s \neq t} \left\{r_{vt}^{(k+1)} - r_{vt}^{(k+1)}\right\},$$

$$\beta_{vt}^{(k+1)} = \beta_{vt}^{(k)} + \frac{\eta}{2} \sum_{w \in \mathcal{B} \in \mathcal{V}} \sum_{u \in \mathcal{U}, u \neq \emptyset} \left\{r_{vt}^{(k)} + r_{vt}^{(k)}\right\}.$$

Introducing unused constraints \(Y_{vt}X_{vt}^T[1, 0]r_{vt} \geq 1_{vt} - \xi_{vt}\) and \(\xi_{vt} \geq 0_{vt}\) with Lagrangian multipliers \(\lambda_{vt}\) and \(\gamma_{vt}\) into (25), by KKT conditions, we can achieve:

$$r_{vt} = U_{vt}^{-1}\left([1, 0]^TX_{vt}^T Y_{vt}\lambda_{vt} - 2[1, 0]^T \alpha_{st} - 2\beta_{vt}\right)$$

$$+ \eta_1 \sum_{s \in \mathcal{F}, s \neq t} [1, 0]^T [1, 0](r_{vt} + r_{vt}) + \eta_2 \sum_{w \in \mathcal{B} \in \mathcal{V}} (r_{vt} + r_{vt}^T).$$

$$VT C \lambda_{vt} - \gamma_{vt} = 0_{vt}.$$
Letting
\[ f_{st} = 2[1, 0]^T \alpha_{st} + 2\beta_{st} - \eta_1 \sum_{x \in T \not= t} [1, 0]^T (r_{st}^{(k)} + r_{st}^{(k)}) - \eta_2 \sum_{x \in S \not= t} (r_{st}^{(k)} + r_{st}^{(k)}), \]

we can also achieve:
\[ \lambda_{st} \in \arg\max_{\lambda_{st}} - \frac{1}{2} \lambda_{st}^T Y_{st} X_{st} [I, I] U_{st}^{-1} [I, I]^T X_{st}^T Y_{st} \lambda_{st} \]
\[ + (1_{st} + Y_{st} X_{st} [I, I] U_{st}^{-1} f_{st})^T \lambda_{st}. \]

Thus, iterations of solving Problem (5) can be summarized as Proposition 1.

Note that, since \( M_1 \) and \( M_2 \) are semi-positive matrices, the objective function of Problem (5) can be shown that it is closed, proper, and convex. Moreover, it is easy to see that the unaugmented Lagrangian \( \mathcal{L} \) in (24) has a saddle point which satisfies
\[ \mathcal{L}(r_{st}^{(k)}, \xi_{st}^{(k)}, \phi_{st}^{(k)}, \alpha_{uts}^{(k)}, \alpha_{sts}^{(k)}, \beta_{uts}^{(k)}, \beta_{sts}^{(k)}) \]
\[ \leq \mathcal{L}(r_{st}^{(k)}, \xi_{st}^{(k)}, \phi_{st}^{(k)}, \alpha_{uts}^{(k)}, \alpha_{sts}^{(k)}, \beta_{uts}^{(k)}, \beta_{sts}^{(k)}). \]

Thus, iterations in Proposition 1 converge to the solution of Problem (5) based on Section 3.2 and Appendix A in [8].

REFERENCES

[1] E. Osuna, R. Freund, and F. Girosi, “Training support vector machines: an application to face detection,” in Computer vision and pattern recognition, 1997. Proceedings., 1997 IEEE computer society conference on, pp. 130–136, IEEE, 1997.

[2] A. McCallum, K. Nigam, J. Rennie, and K. Seymore, “A machine learning approach to building domain-specific search engines,” in IJCAI, vol. 99, pp. 662–667, Citeseer, 1999.

[3] L. Shao, F. Zhu, and X. Li, “Transfer learning for visual categorization: A survey,” Neural Networks and Learning Systems, IEEE Transactions on, vol. 26, no. 5, pp. 1019–1034, 2015.

[4] S. J. Pan and Q. Yang, “A survey on transfer learning,” Knowledge and Data Engineering, IEEE Transactions on, vol. 22, no. 10, pp. 1345–1359, 2010.

[5] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, “Boosting for transfer learning,” in Proceedings of the 24th international conference on Machine learning, pp. 193–200, ACM, 2007.

[6] T. Evgeniou and M. Pontil, “Regularized multi–task learning,” in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 109–117, ACM, 2004.

[7] P. A. Forero, A. Cano, and G. B. Giannakis, “Convexity-based distributed support vector machines,” The Journal of Machine Learning Research, vol. 11, pp. 1663–1707, 2010.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” Foundations and Trends® in Machine Learning, vol. 3, no. 1, pp. 1–122, 2011.

[9] G. Skolidis and G. Sanguinetti, “Semisupervised multitask learning with gaussian processes,” Neural Networks and Learning Systems, IEEE Transactions on, vol. 24, no. 12, pp. 2101–2112, 2013.

[10] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, “Self-taught learning: transfer learning from unlabeled data,” in Proceedings of the 24th international conference on Machine learning, pp. 759–766, ACM, 2007.

[11] A. Navia-Váquez and E. Parrado-Hernandez, “Distributed support vector machines,” Neural Networks, IEEE Transactions on, vol. 17, no. 4, pp. 1091–1097, 2006.

[12] R. Bekkerman, M. Bilenko, and J. Langford, Scaling up machine learning: Parallel and distributed approaches. Cambridge University Press, 2011.

[13] V. Vapnik, The nature of statistical learning theory. Springer Science & Business Media, 2013.