Neurocomputing 189 (2016) 179–188

Contents lists available at ScienceDirect

Neurocomputing

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

Flexible multi-task learning with latent task grouping

Shi Zhong [a], Jian Pu [b], Yu-Gang Jiang [a], Rui Feng [a], Xiangyang Xue [a]

[a] Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China
[b] Institute of Neuroscience, Chinese Academy of Sciences, Shanghai, China

ARTICLE INFO

Article history:
Received 20 October 2015
Received in revised form
21 December 2015
Accepted 27 December 2015
Communicated by Peng Cui
Available online 15 January 2016

Keywords:
Multi-task learning
Group structure
Regularization

ABSTRACT

In multi-task learning, using task grouping structure has been shown to be effective in preventing inappropriate knowledge transfer among unrelated tasks. However, the group structure often has to be predetermined using prior knowledge or heuristics, which has no theoretical guarantee and could lead to unsatisfactory learning performance. In this paper, we present a *flexible* multi-task learning framework to identify *latent* grouping structures under agnostic settings, where the prior of the latent subspace is unknown to the learner. In particular, we relax the latent subspace to be full rank, while imposing sparsity and orthogonality on the representation coefficients of target models. As a result, the target models still lie on a low dimensional subspace spanned by the selected basis tasks, and the structure of the latent task subspace is fully determined by the data. The final learning process is formulated as a joint optimization procedure over both the latent space and the target models. Besides providing proofs of theoretical guarantee on learning performance, we also conduct empirical evaluations on both synthetic and real data. Experimental results and comparisons with competing approaches corroborate the effectiveness of the proposed method.

1. Introduction

Multi-task learning (MTL) aims to train prediction models for a series of tasks jointly and simultaneously. Through exploiting the commonality among multiple tasks, MTL has been shown to be more effective than independently training each single task. In an ideal scenario where all the tasks are related, the identified commonality can normally help improve the generalization performance significantly. However, in realistic applications, the hidden structures among multiple learning tasks can be very complicated. For instance, the learning tasks could consist of several disjoint or partially overlapped task groups as well as some outlier tasks. In the case that some unrelated tasks are mixed together, simply sharing commonality among all the tasks will certainly decrease the learning performance, and such a phenomenon is thus called negative transfer [1].

One way to avoid the negative transfer is to organize related tasks in clusters, namely *task grouping* [2], and knowledge transfer is performed only within each group. Briefly speaking, there exist two levels of task grouping. The first level groups the learning tasks in an explicit manner, where one often assumes that the prediction models of the tasks within the same group share certain commonalities, such as similar structures, parameters, or priors [3–7]. However, such an explicit task grouping strategy tends to be over rigorous, as it often results in disjoint sharing of commonality among all the tasks. On the contrary, the second level, implicit task grouping, was proposed as a valuable option since it can reveal the hidden relationships among the learning tasks [8–12]. For example, Kang et al. [11] proposed a subspace based regularization framework to identify disjoint task groups, where the tasks within each group are assumed to lie in a low-dimensional space. Realizing the limitation of the disjoint task grouping, Kumar and Daumé III [12] further proposed a subspace based task grouping strategy that allows tasks from different groups to overlap by having common basis tasks, namely Grouping and Overlap in MTL (GO-MTL). However, the determination of the number of hidden basis tasks remains unsolved in principle, and prior works often rely on heuristics or empirical validation, which holds no theoretical guarantee.

Motivated by Kumar and Daumé III [12], in this paper we propose a flexible MTL (FMTL) paradigm to identify the task grouping and overlap without imposing any specific structure assumptions, e.g., the number of latent basis tasks. Similar to [12], we assume that the model parameters $\{\mathbf{w}_l\}_{l=1}^{L}$ of L learning tasks reside in a latent subspace spanned by a set of unknown basis tasks $\mathbf{M} = \mathbf{m}_1, \cdots, \mathbf{m}_k, \cdots, \mathbf{m}_L$, where $\mathbf{m}_k \in \mathbb{R}^d$ is the model parameter for the k-th basis task and d is the feature dimension. More specifically, we use a latent factor model to factorize the target model into the latent subspace and the corresponding representation as

E-mail addresses: zhongshi@fudan.edu.cn (S. Zhong), jianpu@fudan.edu.cn (J. Pu), ygj@fudan.edu.cn (Y.-G. Jiang), fengrui@fudan.edu.cn (R. Feng), xyxue@fudan.edu.cn (X. Xue).

http://dx.doi.org/10.1016/j.neucom.2015.12.092
0925-2312/

$\mathbf{w}_l = \mathbf{M}\mathbf{s}_l$. Instead of predetermining the size of latent basis tasks and constraining the subspace to be low rank [12], we use a full rank subspace and introduce two regularization terms to the corresponding representation matrix $\mathbf{S} = \mathbf{s}_1, \cdots, \mathbf{s}_l, \cdots, \mathbf{s}_L$ of the learning tasks. The first regularization term enforces $\mathbf{S}$ to be row sparse that encourages the related tasks to share a subset of basis tasks. The second column-orthogonality regularization term supplies as a complement of the row-sparsity term, which prohibits unrelated tasks to share basis tasks. Finally, we formulate the learning procedure as an optimization problem over two variables, i.e., the latent basis tasks $\mathbf{M}$ and the target model $\{\mathbf{w}_l\}_{l=1}^{L}$. Since the optimization over the latent tasks can be solved analytically, the original problem can be reformed as a *convex* minimization problem over the transformed target model that can be efficiently solved using the accelerated proximal gradient method. We show that our proposed FMTL method holds theoretical guarantee of the performance bound. Extensive experiments are conducted to validate the effectiveness of our method in both regression and classification problems, and results demonstrate that our method outperforms several recent MTL methods.

The remainder of the paper is organized as follows. Section 2 gives a brief review of related works. Section 3 introduces our new formulation of FMTL with latent task grouping. In Section 4, we elaborate the optimization strategy with detailed analysis. We discuss theoretical performance bound in Section 3, and provide empirical studies and comparisons with representative MTL algorithms in Section 6. Finally, Section 7 concludes this paper.

2. Related work

Due to practical needs in many applications, significant efforts have been paid to the design of MTL algorithms. *Model commonality* has been regarded as one of the key ingredients for joint model training. Many works focused on exploiting structure commonality of multiple learning tasks, such as low rank subspace sharing [3,13] and feature set sharing [8,14–19]. In addition, parameter commonality aims to identify the shared parameters across different tasks. Depending on the form of the used models, the shared parameters can be the hidden units in neural networks [4], the priors in hierarchical Bayesian models [5,20–22], the parameters in Gaussian process covariance [23], the feature mapping matrices [24], graph induced structures [25,26] and even the similarity metrics [7,6]. However, these methods solely considering model commonality may suffer from unsatisfactory learning performance since they neglected the fact that some tasks may be unrelated, which is often true in real applications.

To avoid the adverse effect incurred by unrelated tasks, one effective solution is to organize the tasks into groups, namely *task grouping* where the commonality is mainly shared within each group. Thrun and O'Sullivan [2] proposed to mutually measure the relatedness of tasks and select sharing information, which is regarded as one of the pioneer works for task grouping. Jacob et al. [9] developed a similar idea by imposing a cluster norm penalty and formulating the learning procedure as a convex optimization problem. Kang et al. [11] presented a disjoint task grouping method, where they assumed that the commonality sharing only occurs within each task group. By imposing a sparse inducing penalty, Kumar and Daumé III [12] further proposed to group the tasks in a low dimensional subspace using the latent factor model, where the tasks from different groups can partially share a subset of basis tasks. However, the number of latent basis tasks has to be determined by empirical validation or heuristics. Realizing the importance of inferring the "right" number of latent tasks, Passos et al. [27] and Gupta et al. [28] employed nonparametric Bayesian methods to infer the number of latent tasks. However, these Bayesian inference based methods have no guarantee of convergence rate and could suffer from a local optimum.

Finally, identifying *outlier tasks* has also been investigated in some recent works, where one assumes that the major task groups are peppered with some irrelevant outlier tasks. Hence, extracting those outlier tasks can help further alleviate the impact from the negative transfer. A decomposition scheme was utilized to partition the model into a group task component and an outlier task component [29–32].

3. Formulation

Assuming that we are given L tasks associated with training data $\{(\mathbf{X}_1, \{\mathbf{y}_1), \cdots, (\mathbf{X}_L, \mathbf{y}_L)\}$, where $\mathbf{X}_l \in \mathbb{R}^{d \times n_l}$ and $\mathbf{y}_l \in \mathbb{R}^{n_l}$ are the input and output of n_l training instances for the l-th task. For a typical linear regression or classification problem, the prediction function for the l-th task is usually expressed by $f(\mathbf{X}, \mathbf{w}_l) = f(\mathbf{X}^\top \mathbf{w}_l)$, where $\mathbf{w}_l$ is the parameter vector of the target model. We stack all the parameter vectors of the L tasks to obtain a target parameter matrix $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_L] \in \mathbb{R}^{d \times L}$.

Following Kumar and Daumé III [12], we use a latent model to factorize the target matrix into two matrices as:

$$\mathbf{W} = \mathbf{M}\mathbf{S}, \tag{1}$$

where each column of $\mathbf{M}$ represents a latent task, and each column of $\mathbf{S} = \mathbf{s}_1, \cdots, \mathbf{s}_L$ is the representation of each target model using the latent tasks: $\mathbf{w}_l = \mathbf{M}\mathbf{s}_l$. In the latent factor model proposed by [12], the latent task subspace is set to be low rank, i.e., $\mathbf{M} \in \mathbb{R}^{d \times k}$ with $k < \min(d, L)$. Hence, the rank of $\mathbf{W}$ is less than or equal to k, which reflects the hidden grouping structure of the tasks. However, as mentioned earlier, the rank of $\mathbf{M}$, i.e., the number of basis tasks has to be predetermined empirically based on prior information or heuristics, which has no theoretical guarantee.

As the objective is to obtain a low rank parameter matrix $\mathbf{W}$ to reveal the grouping structure of the learning tasks, we enforce the representation matrix $\mathbf{S}$ to exhibit the low rank structure, while relaxing the latent subspace $\mathbf{M}$ to be a full rank matrix, i.e., $\mathbf{M} \in \mathbb{R}^{d \times d}$. In particular, we impose two structure regularization terms of the representation matrix $\mathbf{S}$. The first is a ℓ_{21}-norm regularization term, which introduces *row-sparsity* on $\mathbf{S}$ matrix that encourages related tasks to share a subset of basis tasks. The second term is *column-orthogonality* that prevents unrelated tasks from sharing common basis. Formally, we formulate our FMTL objective as:

$$\begin{aligned} &\min_{\mathbf{M},\mathbf{S}} \sum_{l=1}^{L} \mathcal{L}(f(\mathbf{X}^\top \mathbf{M}\mathbf{s}_l), \mathbf{y}_l) + \alpha \|\mathbf{S}\|_{2,1} + \beta \|\mathbf{S}^\top \mathbf{S}\|_F^2, \\ &\text{subject to : } \mathbf{M}^\top \mathbf{M} = \mathbf{I}_{d \times d}. \end{aligned} \tag{2}$$

The first component $\mathcal{L}(f(\mathbf{X}^\top \mathbf{M}\mathbf{s}_l)$ is a preselected loss function on the training set. For regression and classification problems, the squared loss and logistic loss are typically used, respectively:

$$\mathcal{L}(f(\mathbf{X}_l^\top \mathbf{M}\mathbf{s}_l), \mathbf{y}_l) = (\mathbf{X}_l^\top \mathbf{M}\mathbf{s}_l - \mathbf{y}_l)^2$$
$$\mathcal{L}(f(\mathbf{X}_l^\top \mathbf{M}\mathbf{s}_l), \mathbf{y}_l) = \log\left(1 + \exp(-\mathbf{y}_l \mathbf{X}_l^\top \mathbf{M}\mathbf{s}_l)\right).$$

The second and third components represent the two types of structure regularization terms on the representation matrix of the target model in the latent subspace, where the coefficients α and β weigh the contribution from each term. The constraint $\mathbf{M}^\top \mathbf{M} = \mathbf{I}_{d \times d}$ is used to ensure that the latent basis tasks $\mathbf{M}$ are orthogonal and form a subspace in $\mathbb{R}^d$.

Note that the $\ell_{2,1}$ norm as a structural penalty forces $\mathbf{S}$ to be a row sparse matrix, which is equivalent to selecting a subset of basis tasks to represent the target model $\mathbf{W}$. The column-orthogonal regularization term is employed to penalize the