Learning Personalized Attribute Preference via Multi-Task AUC Optimization

Zhiyong Yang,[1,2] **Qianqian Xu,**[3] **Xiaochun Cao,**[1,2] **Qingming Huang**[3,4,5*]

[1]SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[4]University of Chinese Academy of Sciences, Beijing, China
[5]Key Laboratory of Big Data Mining and Knowledge Management, CAS, Beijing, China
{yangzhiyong, caoxiaochun}@iie.ac.cn, xuqianqian@ict.ac.cn, qmhuang@ucas.ac.cn

Abstract

Traditionally, most of the existing attribute learning methods are trained based on the consensus of annotations aggregated from a limited number of annotators. However, the consensus might fail in settings, especially when a wide spectrum of annotators with different interests and comprehension about the attribute words are involved. In this paper, we develop a novel multi-task method to understand and predict personalized attribute annotations. Regarding the attribute preference learning for each annotator as a specific task, we first propose a multi-level task parameter decomposition to capture the evolution from a highly popular opinion of the mass to highly personalized choices that are special for each person. Meanwhile, for personalized learning methods, ranking prediction is much more important than accurate classification. This motivates us to employ an Area Under ROC Curve (AUC) based loss function to improve our model. On top of the AUC-based loss, we propose an efficient method to evaluate the loss and gradients. Theoretically, we propose a novel closed-form solution for one of our non-convex subproblem, which leads to provable convergence behaviors. Furthermore, we also provide a generalization bound to guarantee a reasonable performance. Finally, empirical analysis consistently speaks to the efficacy of our proposed method.

Introduction

Visual attributes are semantic cues describing visual properties such as texture, color, mood, and *etc*. Typical instances are *comfortable* or *high heeled* for shoes, and *smiling* or *crying* for human faces. During the past decade, attribute learning has emerged as a powerful building block for a wide range of applications (Song, Tan, and Chen 2014; Su et al. 2017; Wang et al. 2017; Yang et al. 2018).

The status quo of the attribute learning methods are mostly based on the global labels aggregated from few annotators (Farhadi et al. 2009; Sadovnik et al. 2013; Luo et al. 2018). Recently, the rise of online crowd-sourcing platforms (like Amazon Mechanical Turk) makes collecting attribute annotations from a broad variety of annotators possible (Kovashka and Grauman 2015), which offers us a chance to revisit the attribute learning. For consensus attribute learning, the underlying assumption is that the user decisions perturb

slightly at random around the common opinion. However, different annotators might very well have distinct comprehension regarding the meaning of the attributes *(typical ones like "open", "fashionable")*. This suggests that the gap between the personalized decisions and the common opinions could not be simply interpreted as random noises. What's worse, one might even come to find conflicting results from different users. In such a case, there is a need to learn the consensus effects as well as personalized effects simultaneously, especially when personalized annotations are available. There are two crucial issues that should be noticed in this problem.

The grouping effects, lying in between consensus and personalized ones, also play an essential role in understanding the user-specific attribute annotations. As pointed out in the previous literature (Kovashka and Grauman 2015), when understanding semantic attributes, humans often form the "school of thoughts" in terms of their cultural backgrounds and the way they interpret the semantic words. Though personalized effect might lead to conflicting results from different schools or groups, the users within the same group might very likely provide similar decisions. Moreover, different groups of people may probably favor distinct visual cues, as there is significant diversity among user groups. In other words, each user group should be assigned with a distinct feature subset. Accordingly, visual features and users should be simultaneously grouped to guarantee better performance. Furthermore, seeing that we cannot obtain the user-feature groups in advance, this constraint should be implicitly and automatically reflected via the structure of model parameters.

Unlike the consensus-based attribute predictions, preference learning is more important than label prediction for personalized attribute learning, be it recommendation and image searching. Under such circumstances, when the attribute words are used as keywords or tags, it should be guaranteed that the positive labeled instances are ranked higher than the negative ones. It is well-known that the Area Under the ROC curve (AUC) metric exactly meets this requirement (Yang et al. 2017), and thus a better objective for our task.

Our goal in this paper is to learn personalized attributes based on the two mentioned issues. More precisely, we regard attribute preference learning for a specific user as a task. On top of this, we propose a multi-task model for the

---

*The corresponding author.

problems we tackle. Our main contributions are listed as follows: a) In the multi-task model, we propose a three-level decomposition of the task parameters which includes a consensus factor, a user-feature co-clustering factor and a personalized factor. b) The proximal gradient descent method is adopted to solve the model parameters. Regarding our contribution here, we derive a novel closed-form solution for the proximal operator of the group factor and provide an efficient AUC-based evaluation method. c) Systematic theoretical analyses are carried out on the convergence behaviors and generalization bounds of our proposed method, while empirical studies are carried out for a simulation dataset and two real-world attribute annotation datasets. Both theoretical and empirical results suggest the superiority of our proposed algorithm. **The codes are now available** [1]

Related Work

Attribute learning Attribute learning has long been playing a central role in many machine learning and computer vision problems. Along this line of research, there are some previous studies that investigate the personalization of attribute learning. (Kovashka and Grauman 2013) learns user-specific attributes with an adaption process. More precisely, a general model is first trained based on a large pool of data. Then a small user-specific dataset is employed to adapt the trained model to user-specific predictors. (Kovashka and Grauman 2015) argues that one attribute might have different interpretations for different groups of persons. Correspondingly, a shade discovery method is proposed therein to leverage group-wise user-specific attributes. Both works adopt two-stage or multi-stage models and even extra dataset. For the group modeling , (Kovashka and Grauman 2015) only focus on user-level grouping while ignores the grouping effect of the features. Furthermore, the merit of AUC is also neglected. In contrast, we propose a fully automatic AUC-based attribute preference learning model where the user-feature coclustering effect is considered.

Multi-task Learning Multi-task learning aims at improving the generalization performance by sharing information among multiple tasks. Many efforts have been made to improve multi-task learning (Chen, Zhou, and Ye 2011; Yu, Tresp, and Yu 2007; Han and Zhang 2016; Zhao et al. 2018; Massias et al. 2018), etc. Recently, there is also a wave to investigate the clustering and grouping based multi-task learning (Zhou, Chen, and Ye 2011; Kumar and Daumé 2012; Xu et al. 2015). Among this works, (Xu et al. 2015) is the most relevant work compared with our work since we adopt the co-clustering regularization proposed therein. However, our work differs significantly in that 1) our model is specially designed for the attribute preference learning problem; 2) we propose a novel hierarchical decomposition scheme for the model parameters and; 3) we propose a novel closed-form solution for the proximal mapping of the co-clustering penalty; 4) we focus on efficient AUC optimization instead of regression or classification.

Model Formulation

In this section, we propose a novel AUC-based multi-task model. Specifically, we first introduce the notations used in this paper, followed by problem settings in our model and a multi-level parameter decomposition. After that, we systematically elaborate two building blocks of our model: the AUC-based loss and evaluation, and the regularization scheme.

Notations

$\langle \cdot, \cdot \rangle$ denotes the inner product for two matrices or two vectors. The singular values of a matrix $\boldsymbol{A}$ are denoted as $\sigma_1(\boldsymbol{A}), \cdots, \sigma_m(\boldsymbol{A})$ such that $\sigma_1(\boldsymbol{A}) \geq \sigma_2(\boldsymbol{A}) \geq \cdots, \geq \sigma_m(\boldsymbol{A}) \geq 0$. $\boldsymbol{I}$ is the identity matrix, $I[\mathcal{A}]$ is the indicator function of the set $\mathcal{A}$, $\boldsymbol{1}$ denotes the all-one vector or matrix. $\mathcal{U}(a, b)$ denotes the uniform distribution and $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution. $\otimes$ denotes the Cartesian product.
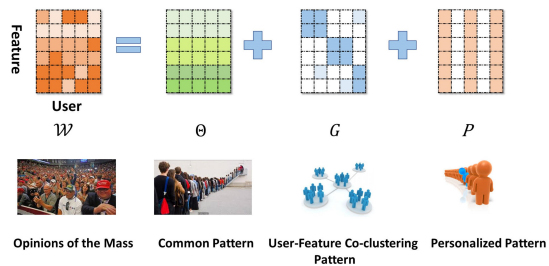
Problem Settings

Figure 1: An illustration of the Multi-level Decomposition of the model parameters.

For a given attribute[2], assume that we have U users who have annotated a given set of images. Further, we assume that the ith user labeled n_i images with $n_{+,i}$ positive labels and $n_{-,i}$ negative labels. $\mathcal{S}_{+,i} = \{k \mid y_k^{(i)} = 1\}$ and $\mathcal{S}_{-,i} = \{k \mid y_k^{(i)} = -1\}$. We denote the training data as: $\mathcal{S} = \left\{ (\boldsymbol{X}^{(1)}, \boldsymbol{y}^{(1)}), \cdots, (\boldsymbol{X}^{(U)}, \boldsymbol{y}^{(U)}) \right\}$. For $\mathcal{S}$, $\boldsymbol{X}^{(i)} \in \mathbb{R}^{n_i \times d}$ is the image feature inputs for the images that the ith user labeled. Each row of $\boldsymbol{X}^{(i)}$ represents the extracted features for a corresponding image. $\boldsymbol{y}^{(i)} \in \{-1, 1\}^{n_i}$ is the corresponding label vector. If $y_k^{(i)} = 1$, then the user thinks that the kth image bears the given attribute, otherwise we have $y_k^{(i)} = -1$.

Taking advantage of the multi-task learning paradigm, the attribute preference learning for a user is regarded as a specific task. Our goal is then to learn all the task models $\boldsymbol{f}^i(x)$, where, for each task, a linear model is learned as the scoring function, i.e $\boldsymbol{f}^{(i)}(\boldsymbol{x}) = \boldsymbol{W}^{(i)\top} \boldsymbol{x}$.

As shown in the introduction, it is natural to observe diversity in personalized scores. However, this diversity could

by no means goes arbitrary large. In fact, we could interpret such limited diversity in a consensus-to-personalization manner. A common pattern is shared among the mass that captures the popular opinion. Different people might have different bias and preference, which drives them away from a consensus. Users sharing similar biases tend to form groups. The users within a group share similar biases towards the popular opinion based on a similar subset of the features of the object. Finally, a highly personalized user in a group tends to adopt an extra bias toward the group opinion. Mathematically, this interpretation induces a multi-level decomposition of the model weights : $\boldsymbol{W}^{(i)} = \boldsymbol{\theta} + \boldsymbol{G}^{(i)} + \boldsymbol{P}^{(i)}$. $\boldsymbol{\theta} \in \mathbb{R}^{d \times 1}$ is the common factor that captures the popular global preference. $\boldsymbol{G}^{(i)} \in \mathbb{R}^{d \times 1}$ is the grouping factor for the i-th task. For mathematical convenience, we denote $\boldsymbol{G} = [\boldsymbol{G}^{(1)}, \cdots, \boldsymbol{G}^{(U)}]$, and we have $\boldsymbol{G} \in \mathbb{R}^{d \times U}$. $\boldsymbol{P}^{(i)}$ is the user-specific factor mentioned above. Similarly, we define $\boldsymbol{P} = [\boldsymbol{P}^{(1)}, \cdots, \boldsymbol{P}^{(U)}]$, and $\boldsymbol{P} \in \mathbb{R}^{d \times U}$. An illustration of this decomposition is shown in Figure 1.

With all the above-mentioned settings, we adopt a general objective function in the form:

$$\min_{\boldsymbol{\theta}, \boldsymbol{G}, \boldsymbol{P}} \sum_{i=1}^{U} \ell_i(\boldsymbol{f}^{(i)}, \boldsymbol{y}^{(i)}) + \lambda_1 \mathcal{R}_1(\boldsymbol{\theta}) + \lambda_2 \mathcal{R}_2(\boldsymbol{G}) + \lambda_3 \mathcal{R}_3(\boldsymbol{P}).$$
(1)

Given (1), there are two crucial building blocks to be determined:

- The empirical loss function for a specific user $i : \ell_i(\cdot, \cdot)$ which directly induces AUC optimization;

- Regularization terms $\mathcal{R}_1(\boldsymbol{\theta})$, $\mathcal{R}_2(\boldsymbol{G})$, and $\mathcal{R}_3(\boldsymbol{P})$ which are defined by prior constraints on $\boldsymbol{W}$.

In what follows, we will elaborate the formulation of two building blocks, respectively.

Regularization

For the common factor $\boldsymbol{\theta}$, we simply adopt the most widely-used ℓ_2 regularization $\mathcal{R}_1(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$ to reduce the model complexity. For $\boldsymbol{G}$, as mentioned in the previous parts, what we pursue here is a user-feature co-clustering effect. A previous work in (Xu et al. 2015) shows that one way to simultaneously cluster the rows and columns of a matrix in $\boldsymbol{R}^{m \times n}$ into κ groups is to penalize the sum of squares of the bottom $min\{n, m\} - \kappa$ singular values. This motivates us to adopt a regularizer on $\boldsymbol{G}$ in the following form $: \mathcal{R}_2(\boldsymbol{G}) = \sum_{\kappa+1}^{min\{d, U\}} \sigma_i^2(\boldsymbol{G})$. For any user i, a non-zero column $\boldsymbol{P}^{(i)}$ is favorable only when she/he has a significant disagreement with the common-level and the group-level results. This inspires us to define $\mathcal{R}_3(\boldsymbol{P}) = \|\boldsymbol{P}\|_{1,2}$ norm to induce column-wise sparsity.

Empirical Loss and Its Evaluation

Since the empirical loss is evaluated separately for each user, without loss of generality, the following discussion only focus on a given user i.

Empirical Loss AUC is defined as the probability that a randomly sampled positive instance has a higher predicted score than a randomly sampled negative instance. Since we need to minimize our objective function, we focus on the loss version of AUC, i.e. the mis-ranking probability. Though the data distribution is unknown, given each user u_i and $\mathcal{S}_{+,i}$, $\mathcal{S}_{-,i}$ defined in the problem setting, we could attain an finite sample-based estimation of the loss version of AUC:

$$\ell_{AUC}^{(i)} = \sum_{x_p \in \mathcal{S}_{+,i}} \sum_{x_q \in \mathcal{S}_{-,i}} \frac{I(\boldsymbol{x}_p, \boldsymbol{x}_q)}{n_{+,i} n_{-,i}},$$

where $I(\boldsymbol{x}_p, \boldsymbol{x}_q)$ is a discrete mis-ranking punishment in the form: $I(\boldsymbol{x}_p, \boldsymbol{x}_q) = I_{[f^{(i)}(\boldsymbol{x}_p) > f^{(i)}(\boldsymbol{x}_q)]} + \frac{1}{2} I_{[f^{(i)}(\boldsymbol{x}_p) = f^{(i)}(\boldsymbol{x}_q)]}$. It is easy to see that $\ell_{AUC}^{(i)}$ is exactly the mis-ranking frequency for user i on the given dataset. Unfortunately, optimizing this metric directly is an $\mathcal{NP}$ hard problem. To address this issue, we adopt the squared surrogate loss $s(t) = (1 - t)^2$ (Gao et al. 2016). Accordingly, the empirical loss $\ell_i(\boldsymbol{f}^{(i)}, \boldsymbol{y}^{(i)})$ could be defined as:

$$\ell_i(\boldsymbol{f}^{(i)}, \boldsymbol{y}^{(i)}) = \sum_{x_p \in \mathcal{S}_{+,i}} \sum_{x_q \in \mathcal{S}_{-,i}} \frac{s\left(\boldsymbol{f}^{(i)}(\boldsymbol{x}_p) - \boldsymbol{f}^{(i)}(\boldsymbol{x}_q)\right)}{n_{+,i} n_{-,i}}.$$
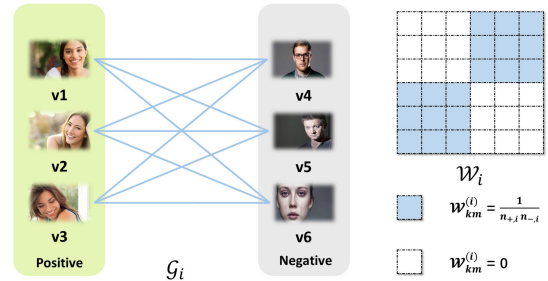

Figure 2: An illustration of the AUC graph, taking the annotation for attribute *smile* as an example.

Efficient AUC-based Evaluation At the first glance, the pair-wise AUC loss induces much heavier computation burdens than the instance-wise losses. It is interesting to note that, after carefully reformulating ℓ_i , the computational burden coming from the pair-wise formulation could be perfectly eliminated. To see this, let us define a graph as $\mathcal{G}^{(i)} = (\mathcal{V}^{(i)}, \mathcal{E}^{(i)}, \mathcal{W}^{(i)})$. The vertex set $\mathcal{V}^{(i)}$ is the set of all the instances in $(\boldsymbol{X}^{(i)}, \boldsymbol{y}^{(i)})$. There exists an edge $(k, m) \in \mathcal{E}^{(i)}$ with weight $\mathcal{W}_{km}^{(i)} = \frac{1}{n_{+,i} n_{-,i}}$ if and only if $y_k^{(i)} \neq y_m^{(i)}$. This graph is further illustrated in Figure 2. Given $\mathcal{W}^{(i)}$, the Laplacian matrix $\boldsymbol{L}^{(i)}$ of $\mathcal{G}_i$ could be expressed as: $\boldsymbol{L}^{(i)} = diag(\mathcal{W}^{(i)} \mathbf{1}) - \mathcal{W}^{(i)}$. The empirical loss could be reformulated as a quadratic form defined by $\boldsymbol{L}^{(i)}$: $\ell_i(\boldsymbol{f}^{(i)}, \boldsymbol{y}^{(i)}) = \frac{1}{2}(\tilde{\boldsymbol{y}}^{(i)} - \boldsymbol{f}^{(i)})^{\top} \boldsymbol{L}^{(i)} (\tilde{\boldsymbol{y}}^{(i)} - \boldsymbol{f}^{(i)})$, where $\tilde{\boldsymbol{y}}^{(i)} = \frac{\boldsymbol{y}^{(i)} + 1}{2}$. We see the AUC loss evaluation involves computing a quadratic form of $\boldsymbol{L}^{(i)}$. The following

proposition gives a general result which suggests an efficient method to compute $\boldsymbol{A}^\top \boldsymbol{L}^{(i)} \boldsymbol{B}$, and $\boldsymbol{A}^\top \boldsymbol{L}^{(i)}$.

Proposition 1. *For any $\boldsymbol{A} \in \mathbb{R}^{n_i \times a}$ and $\boldsymbol{B} \in \mathbb{R}^{n_i \times b}$, where a and b are positive integers. $\boldsymbol{A}^\top \boldsymbol{L}^{(i)} \boldsymbol{B}$, and $\boldsymbol{A}^\top \boldsymbol{L}^{(i)}$ could be finished within $\mathcal{O}\big(n_i(a + b + ab)\big) = \mathcal{O}(abn_i)$ and $\mathcal{O}(an_i)$, respectively.*

Remark 1. *According to this proposition, the complexity of $\boldsymbol{A}^\top \boldsymbol{L}^{(i)} \boldsymbol{B}$ could be reduced from $O(abn_{+,i}n_{-,i})$ to $O(ab(n_{+,i} + n_{-,i}))$, whereas the complexity of $\boldsymbol{A}^\top \boldsymbol{L}$ could be reduced from $O(an_{+,i}n_{-,i})$ to $O(a(n_{+,i} + n_{-,i}))$.*

To end this section, we summarize our final objective function as:

$$(P^*) \min_{\boldsymbol{\theta},\boldsymbol{G},\boldsymbol{P}} \underbrace{\sum_i \sum_{\boldsymbol{x}_p \in \mathcal{S}_{+,i}} \sum_{\boldsymbol{x}_q \in \mathcal{S}_{-,i}} \frac{s\big(\boldsymbol{W}^{(i)^\top}(\boldsymbol{x}_p - \boldsymbol{x}_q)\big)}{n_{+,i}n_{-,i}}}_{\mathcal{L}(\boldsymbol{W})}$$

$$+ \lambda_1 \underbrace{\|\boldsymbol{\theta}\|_2^2}_{\mathcal{R}_1(\boldsymbol{\theta})} + \lambda_2 \underbrace{\sum_{\kappa+1}^{min\{d,U\}} \sigma_i^2(\boldsymbol{G})}_{\mathcal{R}_2(\boldsymbol{G})} + \lambda_3 \underbrace{\|\boldsymbol{P}\|_{1,2}}_{\mathcal{R}_3(\boldsymbol{P})}$$

$$s.t \quad \boldsymbol{W}^{(i)} = \boldsymbol{\theta} + \boldsymbol{G}^{(i)} + \boldsymbol{P}^{(i)}$$

For the sake of simplicity, we denote the empirical loss as $\mathcal{L}(\boldsymbol{W})$, and we denote the objective function of (P^*) as $\mathcal{F}(\boldsymbol{\theta},\boldsymbol{G},\boldsymbol{P})$. Note that $\mathcal{L}(\boldsymbol{W})$ should be a function of $\boldsymbol{\theta}$, $\boldsymbol{G}$, and $\boldsymbol{P}$.

Optimization

We adopt the proximal gradient method as the optimizer for our problem. In this section, we introduce the outline of the optimization method and provide a novel closed-form solution for the proximal operator of $\mathcal{R}_2(\boldsymbol{G})$.

For each iteration step k, giving a reference point $\boldsymbol{W}^{ref_k} = (\boldsymbol{\theta}^{ref_k}, \boldsymbol{G}^{ref_k}, \boldsymbol{P}^{ref_k})$, then the proximal gradient method updates the variables as :

$$\boldsymbol{\theta}^k := argmin_{\boldsymbol{\theta}} \frac{1}{2} \left\| \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}^k \right\|_2^2 + \frac{\lambda_1}{\rho_k}\|\boldsymbol{\theta}\|_2^2 \tag{2}$$

$$\boldsymbol{G}^k := argmin_{\boldsymbol{G}} \frac{1}{2} \left\| \boldsymbol{G} - \tilde{\boldsymbol{G}}^k \right\|_F^2 + \frac{\lambda_2}{\rho_k} \sum_{\kappa+1}^{min\{d,U\}} \sigma_i^2(\boldsymbol{G}) \tag{3}$$

$$\boldsymbol{P}^k := argmin_{\boldsymbol{P}} \frac{1}{2} \left\| \boldsymbol{P} - \tilde{\boldsymbol{P}}^k \right\|_F^2 + \frac{\lambda_3}{\rho_k}\|\boldsymbol{P}\|_{1,2} \tag{4}$$

where: $\tilde{\boldsymbol{\theta}}^k = \boldsymbol{\theta}^{ref_k} - \frac{1}{\rho_k}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{W}^{ref_k})$, $\tilde{\boldsymbol{G}}^k = \boldsymbol{G}^{ref_k} - \frac{1}{\rho_k}\nabla_{\boldsymbol{G}}\mathcal{L}(\boldsymbol{W}^{ref_k})$, $\tilde{\boldsymbol{P}}^k = \boldsymbol{P}^{ref_k} - \frac{1}{\rho_k}\nabla_P\mathcal{L}(\boldsymbol{W}^{ref_k})$ and ρ_k is chosen with a line-search strategy, where we keep on updating $\rho_k = \alpha\rho_k, \alpha > 1$ until it satisfies:

$$\mathcal{L}(\boldsymbol{W}) < \mathcal{L}(\boldsymbol{W}^{ref_k}) + \Psi_{\rho_k}(D\boldsymbol{\theta}) + \Psi_{\rho_k}(D\boldsymbol{G}) + \Psi_{\rho_k}(D\boldsymbol{P}). \tag{5}$$

Here $D\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\theta}^{ref_k}, D\boldsymbol{G} = \boldsymbol{G} - \boldsymbol{G}^{ref_k}$ and $D\boldsymbol{P} = \boldsymbol{P} - \boldsymbol{P}^{ref_k}$, $\Psi_{\rho_k}(D\boldsymbol{A}) = \left\langle \nabla_{\boldsymbol{A}}\mathcal{L}(\boldsymbol{W}^{ref_k}), D\boldsymbol{A} \right\rangle + \frac{\rho_k}{2}\left\langle D\boldsymbol{A}, D\boldsymbol{A} \right\rangle$.

Remark 2. *The existence of such ρ_k is guaranteed by the Lipschitz continuity of $\nabla\mathcal{L}(\boldsymbol{W})$. Note that, we choose the last historical update of the parameter as the reference point i.e. $\boldsymbol{W}^{ref_k} = \boldsymbol{W}^{k-1}$.*

The solution to Eq.(2) and Eq.(4) directly follows the proximal operator of the ℓ_2 norm and the $\ell_{1,2}$ norm (Sra, Nowozin, and Wright 2012). For Eq.(3), we provide a novel closed-form solution in the following subsection.

A closed-form solution for $\boldsymbol{G}$ subproblem Note that $\sum_{\kappa+1}^{min\{d,U\}} \sigma_i(\boldsymbol{G})^2$ is not convex, solving the $\boldsymbol{G}$ subproblem is challenging. Conventionally, this problem is solved in an alternative manner (Xu et al. 2015) which is inefficient and lack of theoretical guarantees. Thanks to the general singular value thresholding framework (Lu et al. 2015; Lin et al. 2017), we could obtain a closed-form optimal solution according to the following proposition.

Proposition 2. *An Optimal Solution of (3) is:*

$$\boldsymbol{G}^* = \boldsymbol{U}\mathcal{T}_{\kappa,\frac{\lambda_3}{\rho_k}}(\boldsymbol{\Sigma})\boldsymbol{V}^\top \tag{6}$$

where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is a SVD decomposition of $\tilde{\boldsymbol{G}}^k$, $\mathcal{T}_{\kappa,c}$ maps $\boldsymbol{\Sigma} = diag(\sigma_1, \cdots, \cdots, \sigma_{min\{d,U\}})$ to a diagonal matrix having the same size with $\mathcal{T}_{\kappa,c}(\boldsymbol{\Sigma})_{ii} = (\frac{1}{2c+1})^{I[i>\kappa]}\sigma_i$.

Theoretical Analysis
Lipschitz Continuity of the Gradients of $\mathcal{L}(\boldsymbol{W})$
In the preceding section, we have pointed out that the Lipschitz Continuity of the Gradients of $\mathcal{L}(\boldsymbol{W})$ is a necessary condition for the success of the line search process to find ρ_k. Now in the following theorem, we formally prove this property as theoretical support for the optimization method.

Theorem 1 (Lipschitz Continuous Gradient). *Suppose that the data is bounded in the sense that:*

$$\forall i, \|\boldsymbol{X}^{(i)}\|_2 = \sigma_{X_i} < \infty, \; n_{+,i} \geq 1, \; n_{-,i} \geq 1.$$

Given two arbitrary distinct parameters $\boldsymbol{W}, \boldsymbol{W}'$, we have:

$$\|\nabla\mathcal{L}(vec(\boldsymbol{W})) - \nabla\mathcal{L}(vec(\boldsymbol{W}'))\| \leq \gamma\Delta\boldsymbol{W}$$

where: $\gamma = 3U\sqrt{(2U+1)}\max_i\left\{\frac{n_i\sigma_{X_i}^2}{n_{+,i}n_{-,i}}\right\}$, $vec(\boldsymbol{W}) = [\boldsymbol{\theta}, vec(\boldsymbol{G}), vec(\boldsymbol{P})]$, $\Delta\boldsymbol{W} = \|vec(\boldsymbol{W}) - vec(\boldsymbol{W}')\|$.

Convergence Analysis

Since the regularization term $\sum_{i=\kappa+1}^{min\{d,U\}} \sigma_i^2(\boldsymbol{G})$ is nonconvex, the traditional sub-differential is not fully available anymore. In this paper, we adopt the generalized sub-differential defined in (Rockafellar and Wets 2009; Liu et al. 2018). To guarantee a nonempty sub-differential set, the objective function must be lower semi-continuous. In our problem, it is obvious that $\mathcal{L}(\boldsymbol{W}), \mathcal{R}_1(\boldsymbol{\theta}), \mathcal{R}_2(\boldsymbol{P})$ are lower semi-continuous functions by their continuity. For the nonconvex term $\sum_{i=\kappa+1}^{min\{d,U\}} \sigma_i^2(\boldsymbol{G})$, the following lemma shows that is also continuous and thus lower semi-continuous.

Lemma 1. *The function* $\sum_{i=\kappa+1}^{min\{d,U\}} \sigma_i^2(\boldsymbol{G})$ *is continuous with respect to* $\boldsymbol{G}$.

Then the convergence properties of the proposed method could be summarized in the following theorem. Here we define $\Delta(\boldsymbol{\theta}^k) = \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k$, $\Delta(\boldsymbol{G}^k) = \boldsymbol{G}^{k+1} - \boldsymbol{G}^k$, $\Delta(\boldsymbol{P}^k) = \boldsymbol{P}^{k+1} - \boldsymbol{P}^k$.

Theorem 2. *Assume that the initial solutions* $\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0$ *are bounded, with the line-search strategy defined in (5), the following properties hold :*

- *1) The sequence* $\{\mathcal{F}(\boldsymbol{\theta}^k, \boldsymbol{G}^k, \boldsymbol{P}^k)\}$ *is non-increasing in the sense that :* $\forall k, \exists C_{k+1} > 0$,

$$
\begin{aligned}
\mathcal{F}(\boldsymbol{\theta}^{k+1}, \boldsymbol{G}^{k+1}, \boldsymbol{P}^{k+1}) \leq \mathcal{F}(\boldsymbol{\theta}^k, \boldsymbol{G}^k, \boldsymbol{P}^k) - \\
C_{k+1}(\|\Delta(\boldsymbol{\theta}^k)\|_2^2 + \|\Delta(\boldsymbol{G}^k)\|_F^2 + \|\Delta(\boldsymbol{P}^k)\|_F^2)
\end{aligned} \tag{7}
$$

- *2)* $\lim_{k\to\infty} \boldsymbol{\theta}^k - \boldsymbol{\theta}^{k+1} = 0$, $\lim_{k\to\infty} \boldsymbol{G}^k - \boldsymbol{G}^{k+1} = 0$, $\lim_{k\to\infty} \boldsymbol{P}^k - \boldsymbol{P}^{k+1} = 0$.
- *3) The parameter sequences* $\{\boldsymbol{\theta}^k\}_k$, $\{\boldsymbol{G}^k\}_k$, $\{\boldsymbol{P}^k\}_k$ *are bounded*
- *4) Every limit point of* $\{\boldsymbol{\theta}^k, \boldsymbol{G}^k, \boldsymbol{P}^k\}_k$ *is a critical point of the problem.*
- *5)* $\forall \, T \geq 1, \exists \, C_T > 0$:

$$
\min_{0 \leq k < T} \left(\|\Delta(\boldsymbol{\theta}^k)\|_2^2 \right) \leq \frac{C_T}{T}, \; \min_{0 \leq k < T} \left(\|\Delta(\boldsymbol{G}^k)\|_F^2 \right) \leq \frac{C_T}{T},
$$
$$
\min_{0 \leq k < T} \left(\|\Delta(\boldsymbol{P}^k)\|_F^2 \right) \leq \frac{C_T}{T}.
$$

Generalization Bound

Define the parameter set Θ as :

$$
\Theta = \big\{ (\boldsymbol{\theta}, \boldsymbol{G}, \boldsymbol{P}) : \sqrt{\mathcal{R}_1(\boldsymbol{\theta})} \leq \psi_1, \mathcal{R}_2(\boldsymbol{G}) \leq \psi_2,
$$
$$
\|\boldsymbol{G}\|_2 \leq \sigma_{max} < \infty, \mathcal{R}_3(\boldsymbol{P}) \leq \psi_3 \big\}
$$

We have the following uniform bound.

Theorem 3. *Assume that* $\exists \Delta_\chi > 0$, *all the instances are sampled such that,* $\|x\| \leq \Delta_\chi$ *.Define* $C = (\psi_1 + \sqrt{\psi_2 + \kappa \cdot \sigma_{max}^2} + \psi_3) \zeta$ *as* $\zeta = \Delta_\chi C$, *we have, for all* $\delta \in (0, 1)$, *for all* $(\boldsymbol{\theta}, \boldsymbol{G}, \boldsymbol{P}) \in \Theta$:

$$
\begin{aligned}
\mathbb{E}_\mathcal{D}(\sum_i \ell_{AUC}^{(i)}) \leq & \mathcal{L}(\boldsymbol{W}) + \sum_{i=1}^U \frac{B_1}{\sqrt{(n_i \chi_i (1 - \chi_i))}} \\
& + B_2 \sqrt{\frac{ln(\frac{2}{\delta})}{\sum_{i=1}^U n_i \chi_i (1 - \chi_i)}}
\end{aligned}
$$

holds with probability at least $1 - \delta$, *where* $B_1 = 8\sqrt{2}C\Delta_\chi(1 + \zeta)$, $B_2 = 10\sqrt{2}(1+\zeta)\zeta$, $\chi_i = \frac{n_{+,i}}{n_i}$. *The distribution* $\mathcal{D} = \otimes_{i=1}^U (\mathcal{D}_{+,i} \otimes \mathcal{D}_{-,i})$, *where for user* i, $\mathcal{D}_{+,i}$, $\mathcal{D}_{-,i}$ *are conditional distributions for positive and negative instances, respectively.*

Remark 3. *According to Theorem 2, the loss function is non-increasing. For the solution of our method* $(\boldsymbol{\theta}^*, \boldsymbol{G}^*, \boldsymbol{P}^*)$, *we then have:* $\sqrt{\mathcal{R}_1(\boldsymbol{\theta}^*)} \leq \sqrt{\frac{\mathcal{F}(\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0)}{\lambda_1}}$,

$\mathcal{R}_2(\boldsymbol{G}^*) \leq \frac{\mathcal{F}(\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0)}{\lambda_2}$, $\mathcal{R}_3(\boldsymbol{P}^*) \leq \frac{\mathcal{F}(\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0)}{\lambda_3}$. *Meanwhile, it could be derived from Theorem 2 that* $\boldsymbol{G}^*$ *is bounded. By choosing* $\psi_1 = \sqrt{\frac{\mathcal{F}(\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0)}{\lambda_1}}, \psi_2 = \frac{\mathcal{F}(\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0)}{\lambda_2}, \psi_3 = \frac{\mathcal{F}(\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0)}{\lambda_3}$, *all solutions chosen by our algorithm belongs to* Θ. *Then, with high probability, all these solutions could reach a reasonable generation gap between the expected 0-1 AUC loss metric* $\mathbb{E}_\mathcal{D}(\sum_i \ell_{AUC}^{(i)})$ *and the estimated surrogate loss on the training data* $\mathcal{L}(\boldsymbol{W})$, *with an order* $\mathcal{O}(\sum_{i=1}^U \frac{1}{\sqrt{(n_i \chi_i (1 - \chi_i))}})$.

Empirical Study

Experiment Settings

For all the experiments, hyper-parameters are tuned based on the training and validation set(account for 85% of the total instances), and the result on the test set are recorded.

Competitors

In this paper, we compare our model with the following competitors: **Robust Multi-Task Learning (RMTL)** (Chen, Zhou, and Ye 2011): RMTL aims at identifying irrelevant tasks when learning from multiple tasks. To this end, the model parameter is decomposed into a low-rank structure and group sparse structure. **Robust Multi-Task Feature Learning (rMTFL)** (Yu, Tresp, and Yu 2007): rMTFL assumes that the model W can be decomposed into two components: a shared feature structure $\boldsymbol{P}(\ell_{1,2}$ norm penalty) and a group-sparse structure $\boldsymbol{Q}$ ($\ell_{1,2}$ norm penalty on its transpose) that detects outliers. **Lasso**: The the ℓ_1-norm regularized multi-task least squares method. **Joint Feature Learning (JFL)**(Nie et al. 2010): In JFL all the models are expected to share a common set of features. To this end, the group sparsity constraint is imposed on the models via the $\ell_{1,2}$ norm. **The Clustered Multi-Task Learning Method (CMTL)**: (Zhou, Chen, and Ye 2011): CMTL assumes that the tasks could be clustered into k groups. Then a k-means based regularizer is adopted to leverage such a structure. **The task-feature coclusters based multi-task method (COMT)** (Xu et al. 2015): COMT assumes that the task-specific components bear a feature-task coclustering structure. **Reduced Rank Multi-Stage multi-task learning (RAMU)** (Han and Zhang 2016): RAMU adopts a capped trace norm regularizer to minimize only the singular values smaller than an adaptively tuned threshold.

Note that since (Kovashka and Grauman 2013) adopts an extra data pool and (Kovashka and Grauman 2015) includes extra initialization algorithms based on (Kovashka and Grauman 2013), our method is not compared with them for the sake of fairness.

Simulated Dataset

In this subsection, we will generate a simulated annotation dataset with 100 simulated users, where the features and AUC scores are produced according to our proposed model. For each user, 5000 samples are generated as $\boldsymbol{X}^{(i)} \in \mathbb{R}^{5000 \times 80}$ and $\boldsymbol{x}_k^{(i)} \sim \mathcal{N}(0, \boldsymbol{I}_{80})$. This results in a volume

Table 1: AUC Comparison on Simulation Dataset

Alg	RMTL	rMTFL	LASSO	JFL
mean	83.48	83.45	83.57	83.49
Alg	CMTL	COMT	RAMU	Ours
mean	83.47	83.44	83.50	**99.65**

Table 2: Running Time Comparison (seconds): Original stands for the original AUC evaluation, wheres ours stands for our acceleration scheme.

ratio	20%	40%	60%	80%	100%
Orginal	18.57	74.22	151.86	268.55	nan
Ours	**3.06**	**5.50**	**8.65**	**12.46**	**15.82**

of *500,000 overall annotations*. To capture the global information, we set $\boldsymbol{\theta}$ as $\boldsymbol{\theta} \sim \mathcal{U}(0,5) + \mathcal{N}(0, 0.5^2)$. In terms of the co-cluster nature, $\boldsymbol{G}$ is produced with a block-wise grouping structure for feature-user co-cluster. Specifically, we create 5 blocks for $\boldsymbol{G}$, namely: $\boldsymbol{G}(1:20, 1:20)$, $\boldsymbol{G}(21:40, 21:40)$, $\boldsymbol{G}(41:50, 41:60)$, $\boldsymbol{G}(51:70, 61:80)$ and $\boldsymbol{G}(71:80, 81:100)$. For each of the block, the elements are generated from the distribution $\mathcal{N}(C_i, 2.5^2)$ (generated via element-wise sampling) where $C_i \sim \mathcal{U}(0, 10)$ is the centroid for the corresponding cluster and thus is shared among a specific cluster. For the elements that do not belong to the 5 chosen blocks are set as 0. For $\boldsymbol{P} \in \mathbb{R}^{d \times U}$, we set $\boldsymbol{P}(:, 1:5)$, $\boldsymbol{P}(:, 10:15)$, $\boldsymbol{P}(:, 20:25)$ randomly with the distribution $\mathcal{U}(0, 10)$, while the remaining entries are set as 0. For each user, the scoring function are generated as $\boldsymbol{s}^{(i)} = \boldsymbol{X}^{(i)}(\boldsymbol{\theta} + \boldsymbol{G}^{(i)} + \boldsymbol{P}^{(i)}) + \epsilon^{(i)}$, where $\epsilon^{(i)} \in \mathbb{R}^{5000 \times 1}$, and $\epsilon^{(i)} \sim \mathcal{N}(0, 0.01^2 \boldsymbol{I}_{5000})$. To generate the labels $\boldsymbol{Y}^{(i)}$ for each i, the top 100 instances with highest scores are labeled as 1, while the remaining instances are labeled as -1.

The performance of all the involved algorithms on the simulated dataset is recorded in Table 1. The corresponding results show that our proposed algorithm consistently outperforms other competitors. Specifically, our algorithm reaches an AUC score of 99.65, where the second best algorithm only attain a score of 83.57.

Besides the generalized performance, we could also verify empirically the ability of our algorithm to recover the expected structures on the parameters $\boldsymbol{W}$. With the same simulated dataset, we compare the parameter $\boldsymbol{W}$ learned from our proposed and the Ground Truth parameters in Figure 3. The results show that our proposed methods could roughly recover the expected group-based structure.

In Theorem 2, we have proved the convergence behavior of the proposed algorithm. To verify theoretical findings, we plot the loss and parameter evolution against the number of iteration in Figure 4. In Figure 4-(a), we see that the loss function constantly decreases as the iteration proceeds, whereas in Figure 4-(b), it is easy to find that the parameter difference $log(\|\boldsymbol{W}^{t+1} - \boldsymbol{W}^t\|)$ also keeps decreasing. All these empirical observations coincide with our theoreti-
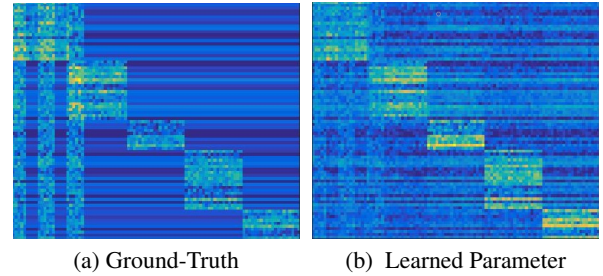

| (a) Ground-Truth | (b) Learned Parameter |

Figure 3: The Potential of our proposed method to Recover the Expected Structure of the Parameters
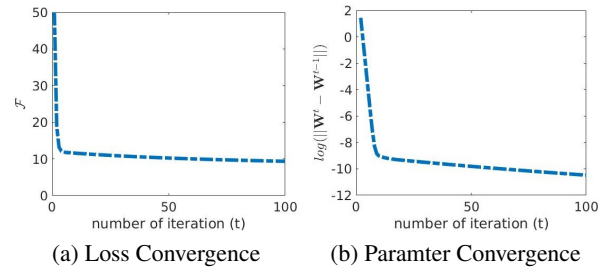

| (a) Loss Convergence | (b) Paramter Convergence |

Figure 4: The Convergence Behavior On Simulation Dataset: a)shows the loss convergence, whereas b) exhibits the convergence property in terms of the parameters.

cal results.

To verify the efficiency of the proposed AUC evaluation scheme, we evaluate the running time of our algorithm with and without the AUC evaluation scheme. The resulting comparison is recorded in Table 2 when different ratios of the dataset are adopted as the training set. As what exhibited here, the original algorithm without the efficient AUC evaluation scheme gets slowly sharply when the training sample increases. When 100% samples are included in the training set, our server couldn't finish the program within 1h due to the memory limit (24GB). We denote as nan correspondingly. In contrast, we can see an up to 20 times speed-up with the help of the proposed scheme in proposition 1.

Shoes Dataset

The Shoes Dataset is collected from (Kovashka and Grauman 2015) which contains 14,658 online shopping images. In this dataset, 7 attributes are annotated by users with a wide spectrum of interests and backgrounds. For each attribute, there are at least 190 users who take part in the annotation, and each user is assigned with 50 images. Overall, 90,000 annotations are collected in this dataset. We concatenate the GIST and color histograms provided by the original dataset as the features. To remove the redundant input features, Principal Component Analysis (PCA) is performed before training, and only the components that are capable of interpreting the first 99% of the total data variance are preserved. Meanwhile, we also need to eliminate the effect of the users who extremely prefer to provide merely the positive (negative)

Table 3: Performance Comparison based on the AUC metric

| | Attributes | | | | | | | | | | | |
| | Shoes | | | | | | | Sun | | | | |
Alg	BR	CM	FA	FM	OP	ON	PT	CL	MO	OP	RU	SO
RMTL	79.31	84.99	66.90	85.08	75.67	67.22	75.14	69.36	62.71	75.28	67.91	69.23
rMTFL	70.90	83.78	67.27	85.91	73.71	65.21	77.11	69.27	62.15	75.80	68.16	68.76
LASSO	68.46	80.48	65.90	84.01	71.47	64.60	75.08	67.64	61.83	75.39	68.57	69.13
JFL	72.00	83.10	67.26	85.93	73.02	65.39	77.09	68.63	61.94	75.00	67.17	68.78
CMTL	74.54	85.16	68.21	85.32	75.06	68.17	77.62	72.55	66.61	79.78	72.34	72.82
COMT	84.24	88.68	69.66	89.19	80.93	72.99	80.62	70.69	63.72	76.93	69.43	70.44
RAMU	78.33	84.58	65.78	84.68	75.25	66.72	73.50	72.95	69.25	79.81	74.39	72.50
Ours	**92.95**	**90.92**	**73.24**	**92.65**	**87.95**	**81.07**	**86.22**	**79.31**	**78.19**	**86.50**	**81.88**	**78.98**

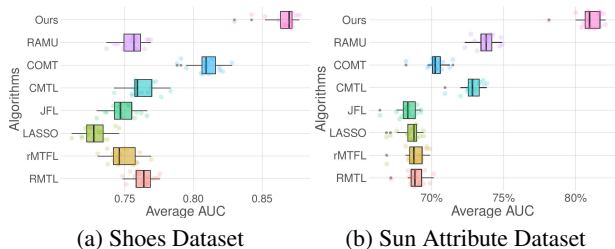(a) Shoes Dataset　　(b) Sun Attribute Dataset

Figure 5: Average performances on all attributes of shoes dataset

labels. To this end, we remove the users who give less than 8 annotations for at least one of the classes.

The left half of Table 3 shows the average performance of the 15 repetitions with the experimental setting (BR: Brown, CM:Comfortable, FA:Fashionable, FM:Formal, OP: Open, ON: Ornate, PT:Pointy). Furthermore, in Figure 5 ,we visualize the average result over the 7 attributes for 15 repetitions with a boxplot. Accordingly, we could reach the conclusion that our proposed algorithm consistently outperforms all the benchmark algorithms by a significant margin.

Sun Attribute Dataset

The SUN Attributes Dataset (Patterson and Hays 2012), is a well-known large-scale scene attribute dataset with roughly 1,4000 images and a taxonomy of 102 discriminative attributes. Recently, in (Kovashka and Grauman 2015), the personalized annotations over five attributes are collected with hundreds of annotators. For each person, 50 images are labeled based on their own comprehension and preference. Overall, this dataset contains 64,900 annotations collected from different users. As for dataset preprocessing, we adopt almost the same procedure as the shoes dataset. The difference here is that we use the second last fc layer of the Inception-V3 (Szegedy et al. 2016) network as the input feature. Furthermore, the PCA is done for each attribute preserving 90% of the total data variance.

The right half of Table 3 shows the average performance over 15 repetitions (CL:Cluttered, MO: Modern, OP: Open-

ing Area, RU: Rustic, SO: Soothe) , and Figure 5-(b) shows the average AUC scores over 5 attributes for the 15 repetitions. Similar to the shoes dataset, we see that our proposed algorithm consistently outperforms all the benchmark algorithms.

Conclusion

In this paper, we propose a novel multi-task model for learning user-specific attribute comprehension with a hierarchical decomposition to model the consensus-to-personalization evolution and an AUC-based loss function to learn the preference. Furthermore, we propose an efficient AUC-based evaluation method to significantly reduce the computational complexity of computing the loss and the gradients. Both theoretical and empirical analysis demonstrates the effectiveness of our proposed method.

References

Chen, J.; Zhou, J.; and Ye, J. 2011. Integrating low-rank and group-sparse structures for robust multi-task learning. In *KDD*, 42–50.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. A. 2009. Describing objects by their attributes. In *CVPR*, 1778–1785.

Gao, W.; Wang, L.; Jin, R.; Zhu, S.; and Zhou, Z. 2016. One-pass AUC optimization. *Artif. Intell.* 236:1–29.

Han, L., and Zhang, Y. 2016. Multi-stage multi-task learning with reduced rank. In *AAAI*, 1638–1644.

Kovashka, A., and Grauman, K. 2013. Attribute adaptation for personalized image search. In *CVPR*, 3432–3439.

Kovashka, A., and Grauman, K. 2015. Discovering attribute shades of meaning with the crowd. *International Journal of Computer Vision* 114(1):56–73.

Kumar, A., and Daumé, H. 2012. Learning task grouping and overlap in multi-task learning. In *ICML*.

Lin, Y.; Yang, L.; Lin, Z.; Lin, T.; and Zha, H. 2017. Factorization for projective and metric reconstruction via truncated nuclear norm. In *IJCNN*, 470–477.

Liu, R.; Cheng, S.; Liu, X.; Ma, L.; Fan, X.; and Luo, Z. 2018. A bridging framework for model optimization and deep propagation. In *NIPS*.

Lu, C.; Zhu, C.; Xu, C.; Yan, S.; and Lin, Z. 2015. Generalized singular value thresholding. In *AAAI*, 1805–1811.

Luo, C.; Li, Z.; Huang, K.; Feng, J.; and Wang, M. 2018. Zero-shot learning via attribute regression and class prototype rectification. *IEEE Trans. Image Processing* 27(2):637–648.

Massias, M.; Fercoq, O.; Gramfort, A.; and Salmon, J. 2018. Generalized concomitant multi-task lasso for sparse multi-modal regression. In *AISTATS*, 998–1007.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint l2,1-norms minimization. In *In NIPS*, 1813–1821.

Patterson, G., and Hays, J. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2751–2758. IEEE.

Rockafellar, R. T., and Wets, R. J.-B. 2009. *Variational analysis*, volume 317. Springer Science & Business Media.

Sadovnik, A.; Gallagher, A. C.; Parikh, D.; and Chen, T. 2013. Spoken attributes: Mixing binary and relative attributes to say the right thing. In *ICCV*, 2160–2167.

Song, F.; Tan, X.; and Chen, S. 2014. Exploiting relationship between attributes for improved face verification. *Computer Vision and Image Understanding* 122:143–154.

Sra, S.; Nowozin, S.; and Wright, S. J. 2012. *Optimization for machine learning*. Mit Press.

Su, C.; Zhang, S.; Yang, F.; Zhang, G.; Tian, Q.; Gao, W.; and Davis, L. S. 2017. Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping. *Pattern Recognition* 66:4–15.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Wang, Y.; Kwok, J. T.; Yao, Q.; and Ni, L. M. 2017. Zero-shot learning with a partial set of observed attributes. In *IJCNN*, 3777–3784.

Xu, L.; Huang, A.; Chen, J.; and Chen, E. 2015. Exploiting task-feature co-clusters in multi-task learning. In *AAAI*.

Yang, Z.; Zhang, T.; Lu, J.; Zhang, D.; and Kalui, D. 2017. Optimizing area under the ROC curve via extreme learning machines. *Knowl.-Based Syst.* 130:74–89.

Yang, Z.; Xu, Q.; Cao, X.; and Huang, Q. 2018. From common to special: When multi-attribute learning meets personalized opinions. In *AAAI*, 515–522.

Yu, S.; Tresp, V.; and Yu, K. 2007. Robust multi-task learning with *t*-processes. In *ICML*, 1103–1110.

Zhao, M.; An, B.; Yu, Y.; Liu, S.; and Pan, S. J. 2018. Data poisoning attacks on multi-task relationship learning. In *AAAI*.

Zhou, J.; Chen, J.; and Ye, J. 2011. Clustered multi-task learning via alternating structure optimization. In *NIPS*, 702–710.