From Common to Special: When Multi-Attribute Learning Meets Personalized Opinions

中国科学院信息工程研究所 INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学 University of Chinese Academy of Sciences

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

[1]SKLOIS, IIE, CAS [2]School of Cyber Security, UCAS

[3]UCAS [4]Key Lab of Intell. Info. Process., ICT, CAS

中国科学院计算技术研究所 INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Introduction

In this paper, we propose a novel model to learn user-specific predictors across multiple attributes. In our proposed model, the diversity of personalized opinions and the intrinsic relationship among multiple attributes are unified in a common-to-special manner. To this end, we adopt a three-component decomposition. Specifically, our model integrates a common cognition factor, an attribute-specific bias factor and a user-specific bias factor. Meanwhile Lasso and group Lasso penalties are adopted to leverage efficient feature selection. Furthermore, theoretical analysis is conducted to show that our proposed method could reach reasonable performance. Eventually, the empirical study carried out in this paper demonstrates the effectiveness of our proposed method.

Model Formulation

- Assume that we have n_a attributes to be evaluated, and that, for the ith attribute, we are given user-specific labels from n_{u_i} different workers. Then the training data could be represented as:

$$\mathcal{T} = \left\{(X^{(1,1)}, y^{(1,1)}), \cdots, (X^{(n_a, n_{u_{n_a}})}, y^{(n_a, n_{u_{n_a}})})\right\}$$

where n_{ij} is the number of images the jth user for the ith attribute labeled.

- Our goal is to learn a predictor $\boldsymbol{f}^{(i,j)}$ for each of the personalized label vectors $\boldsymbol{y}^{(i,j)}$. In this paper, we assume that $f^{(i,j)}(\cdot)$ has a linear form :

$$\boldsymbol{f}^{(i,j)} = X^{(i,j)}\boldsymbol{w}^{(i,j)}$$

where $\boldsymbol{w}^{(i,j)}$ is the corresponding model weight.

- We adopt a three-component additive decomposition of $\boldsymbol{w}^{(i,j)}$:

$$\boldsymbol{w}^{(i,j)} = \boldsymbol{\theta} + \boldsymbol{p}^{(i)} + \boldsymbol{u}^{(i,j)} \quad (1)$$

, where $\boldsymbol{\theta}$ is the general cognition factor, $\boldsymbol{p}^{(i)}$ is the attribute specific bias factor, $\boldsymbol{u}^{(i,j)}$ is the user specific bias factor)

- Objective function :

$$\min_{W} \sum_{i=1}^{n_a}\sum_{j=1}^{n_{u_i}} \left\| \boldsymbol{y}^{(i,j)} - X^{(i,j)}(\boldsymbol{\theta} + \boldsymbol{p}^{(i)} + \boldsymbol{u}^{(i,j)}) \right\|^2 + \lambda_1 \left\|\boldsymbol{\theta}\right\|_1 + \lambda_2 \left\|\boldsymbol{P}\right\|_{1,2} + \lambda_3 \left\|\boldsymbol{U}^{\top}\right\|_{1,2} \quad (2)$$

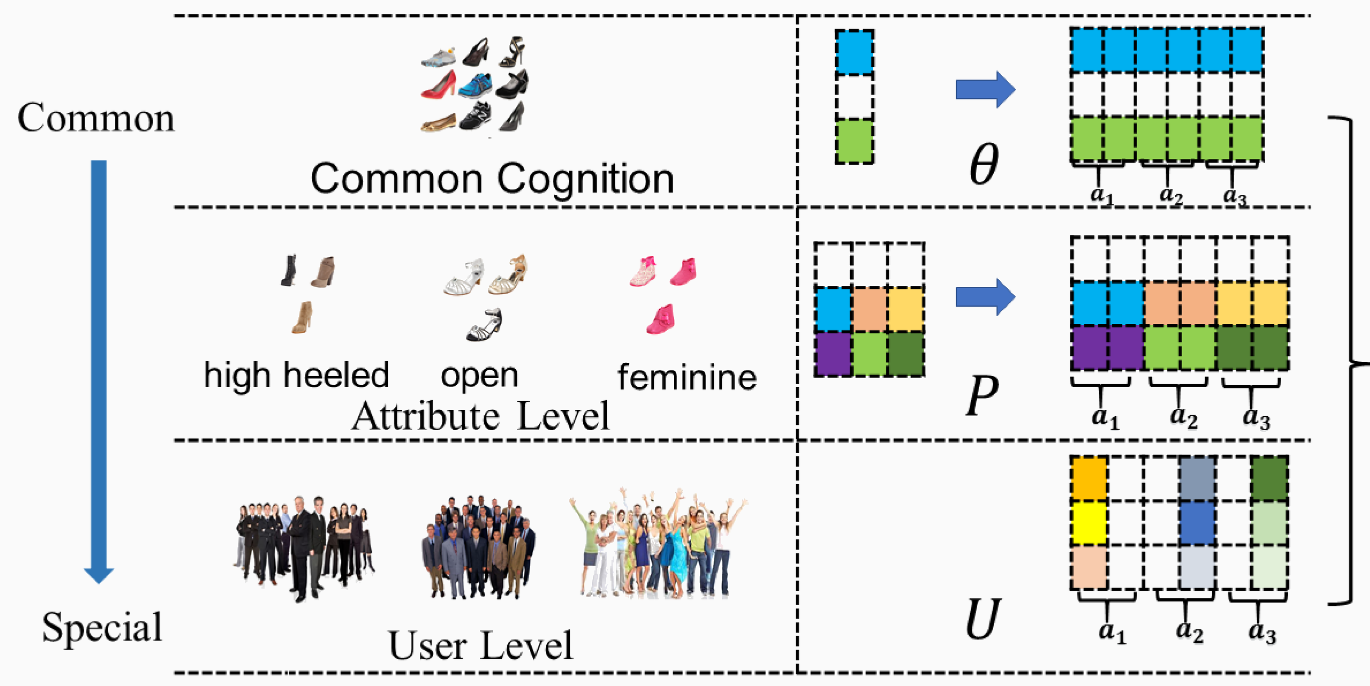

Fig. 1: illustration of our proposed model

Fig. 1 illustrates the three-component decomposition. Here a_1, a_2 and a_3 are the three mentioned attributes: high heeled, open and feminine. We assume that, for each attribute, there are two annotators who labeled the corresponding images. Note that we extend θ and P to match the size of U.

Theoretical Analysis

Proposition 1 (Global Optimality). *P1 is jointly convex with respect to* $\boldsymbol{\theta}$, $\boldsymbol{P}$, $\boldsymbol{U}$

Proposition 2 (Lipschitz Continuous Gradient). *Given two arbitrary feasible solutions* W *and* W'*, we have :*

$$\left\|\nabla L(\tilde{W}) - \nabla L(\tilde{W}')\right\| \le \rho \left\|\tilde{W} - \tilde{W}'\right\|$$

Theorem 1 (Performance Bounds).

$$\mathbb{P}\left(\frac{1}{n_{min}n_u}\left\|X\overline{W} - F\right\|^2 \le \left(\frac{2\zeta}{n_{min}n_u}\right)^2\right) \ge \delta(t)$$

$$\mathbb{P}\left(\left\|\hat{\theta} - \theta^*\right\|_1 \le \frac{2(\beta_\theta + 1)\sqrt{n_\theta}}{\kappa_\theta n_{min}n_u}\zeta\right) \ge \delta(t)$$

$$\mathbb{P}\left(\left\|\hat{P} - P^*\right\|_{1,2} \le \frac{2(\beta_p + 1)\sqrt{n_p}}{\kappa_p n_{min}n_u}\zeta\right) \ge \delta(t)$$

$$\mathbb{P}\left(\left\|\hat{U}^\top - U^{*\top}\right\|_{1,2} \le \frac{2(\beta_{u,a} + 1)\sqrt{n_{u,a}}}{\kappa_{u,a} n_{min}n_u}\zeta\right) \ge \delta(t)$$

Experiments

Simulated Dataset

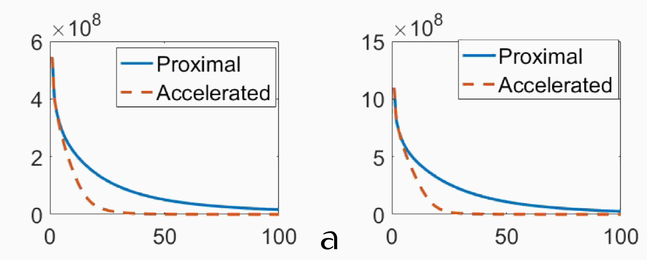

Fig. 2: Effect of nestrov's Acceleration with (a) 40% random selected samples as training data and (b) 80% samples as training data. The y axis represents the average loss function of ($P1$) on 10 repetitions, and the x axis represents the iteration number

Table 1 Simulated dataset with 40% samples selected as training data

SVR	Ridge	Lasso	Elastic Net	ours
1.000	0.830	2.99E-05	9.96E-05	**2.70E-05**
1.000	0.830	3.20E-05	9.92E-05	**2.82E-05**
1.000	0.830	3.15E-05	1.07E-04	**2.78E-05**
0.998	0.829	4.20E-05	1.14E-04	**3.75E-05**
1.005	0.834	2.88E-05	9.34E-05	**2.72E-05**
1.001	0.830	3.28E-05	1.03E-04	**2.95E-05**

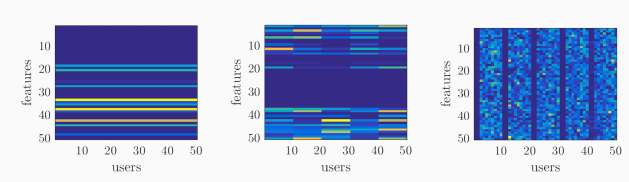

Fig. 3: Figures for the magnitude of θ,P and U. For comparison convenience, both θ and P are extended to $\mathbb{R}^{d \times n_u}$ matrix

Table 2 Performance comparison for the simulated dataset with 80% samples selected as training data

SVR	Ridge	Lasso	Elastic Net	ours
1.005	0.884	2.27E-05	5.19E-05	**2.11E-05**
1.009	0.888	2.37E-05	5.27E-05	**2.17E-05**
1.009	0.887	2.30E-05	5.14E-05	**2.12E-05**
1.010	0.889	2.79E-05	5.64E-05	**2.60E-05**
1.019	0.896	2.03E-05	4.79E-05	**1.89E-05**
1.010	0.889	2.35E-05	5.20E-05	**2.18E-05**

Real World Datasets

Table 3 Performance comparison for the relative attribute dataset

algorithm	accuracy	
	40%	80%
rel_attr	0.4797	0.5195
RankNet	0.4791	0.4721
RankBoost	0.4669	0.5251
user exclusive	0.4753	0.5303
user adaptive	0.4777	0.5336
rMTFL-G	0.4807	0.5074
rMTFL-U	0.4838	0.5433
ours	**0.5119**	**0.5546**

Table 4 Performance comparison for the binary attribute dataset

algorithm	accuracy	
	40%	80%
SVM	0.6278	0.6756
MLP	0.5057	0.4997
user exclusive	0.6549	0.6913
user adaptive	0.6771	0.6956
rMTFL-G	0.6421	0.6941
rMTFL-U	0.6659	0.7056
ours	**0.6894**	**0.7121**

Conclusion In this paper, we propose a hierarchical multi-task model for user specific attribute learning across multiple attributes with a common-to-special decomposition of the model weights. Specifically, our model weights include a common cognition factor, an attribute-specific factor and a user specific factor. The well-known accelerated proximal gradient method is employed to solve this model. Based on assumption 1, we prove theoretically that the proposed algorithm could both leverage good performance and estimate the true parameters well with high probability. The experiment results further verify the effectiveness of our proposed model.