Learning from Multi-User Multi-Attribute Annotations

Ou Wu*, Shuxiao Li, Honghui Dong†, Ying Chen, Weiming Hu

Abstract

Mining the data source from a crowd of people has elicited increasing attention in recent years. In existing studies, multiple users are utilized, in which each user is generally required to annotate only one attribute for each sample. However, there are cases in numerous annotation tasks wherein despite of the presence of multiple users, each user should classify or rate multiple attributes for each sample. This situation is referred to as multi-user multi-attribute annotations in this paper. This work deals with the learning problem under multi-user multi-attribute annotations. A generative model is introduced to describe the human labeling process for multi-user multi-attribute annotations. Subsequently, a maximum likelihood approach is leveraged to infer the parameters in the generative model, namely, ground-truth labels, user expertise, and annotation difficulties. The classifiers for each attribute are also learned simultaneously. Furthermore, the correlations among attributes are taken into account during inference and learning using conditional random field. The experimental results reveal that compared with existing methods that ignore the characteristics of multi-user multi-attribute annotations, our approach can obtain better estimation of the ground truth labels, user experts, annotation difficulties as well as attribute classifiers.

1 Introduction

Crowdsourcing machine learning is a relatively new research area that has rapidly emerged in recent years [10]. This concept has a critical role in learning problems when labels are obtained from multiple non-expert users. In such context, labels are usually noisy and not equally reliable. Therefore, a simple majority strategy that treats different users equally is unsuitable. Sophisticated learning algorithms have been proposed to identify the expertise of users and ground-truth labels [5][14]. In existing crowdsourcing learning works, each training sample is associated with only one target attribute, that is, each user is required to annotate only one attribute for each sample.

Nevertheless, in recent computer vision studies, particularly those on attribute learning, each sample is associated with multiple attributes. If labels for each attribute are also obtained from multiple annotators, each user is required to annotate multiple attributes for each sample. In this context, the label procedure is relatively different from those of previous crowdsourcing learning studies where only one attribute was investigated. Furthermore, the attributes are usually correlated, and thus learning the classifiers for each attribute separately abandons the correlation information. Therefore, introducing new learning strategies to deal with the new context of multi-user multi-attribute annotations is necessary.

This study is also motivated by a number of human-computer-interaction (HCI) studies that employed multi-user multi-attribute annotations. A crowd of persons is usually recruited to rate/classify a set of attributes for a number of samples. For example, in [2], users rated six visual attributes ('complexity', 'clean', 'organize', 'interesting', 'clear', and 'beautiful') of given Web pages. In [7], the relationship between Web aesthetics and accessibility was investigate by recruiting thirty-two users to rate five attributes (i.e., 'clean', 'pleasing', 'fascinating', 'creative', and 'aesthetic') of 50 Web pages. The attributes in both studies are highly correlated. Nevertheless, the labels for each attribute in both studies are obtained by a majority strategy and the classification models for each attribute are separately constructed; thus, the correlation is ignored.

In this paper, a new generative model is proposed to describe the labeling process during user annotations. This model will allow better learning from multi-user multi-attribute annotations. The proposed model captures several key factors that affect the accuracy of user labels, namely, ground-truth labels, user expertise, and annotation difficulties. User expertise and annotation difficulties for each attribute are useful for user assessment and attribute analysis. To estimate these key factors and to learn the attribute classifiers, a maximum likelihood approach [1] is leveraged and combined with an EM optimization method. The attributes of a sample are generally correlated, so the conditional random field (CRF) theory used in [4] is utilized to improve the entire inference and learning approach. Our main contributions are as follows:

- A generative model is proposed to describe the multi-user multi-attribute annotations. The model contains several key factors affecting the accuracies of user labels. In addition, the classifier parameters for each attribute are considered and learned.
- A maximization likelihood approach is applied to si-

*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. {wuou, sxli, ychen, wmhu}@nlpr.ia.ac.cn

†Beijing Jiaotong University. hhdong@bjtu.edu.cn

multaneously infer the parameters of the generative model and learn the attribute classifiers. Furthermore, when EM is run, a CRF graph is used to describe the attribute correlations. The whole approach has been proved to be effective by experiments.

- We applied the proposed learning algorithm into visual attribute assessment for Web pages. The classifiers for the visual attributes, namely, visual quality, visual complexity, and interesting, are learned in an integrated manner. In previous literature [2], these visual attributes are separately learned.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 introduces the methodologies. Section 4 describes the improved model by considering attribute correlations. Section 5 reports experimental results. Section 6 concludes the paper.

2 Related works

2.1 Learning from multiple annotators Several studies [5][14][16] have been performed to deal with the setting involving multiple annotators. The aim of these studies is to pursue key factors, such as ground-truth labels and user expertise, as well as to learn classifiers. Therefore, a graphic model is usually introduced to describe the human labeling process. Subsequently, the model is used to infer the key factors and learn the classifiers. Thus far, no studies have investigated cases wherein multiple target attributes are annotated by each user. In this setting, the human labeling process is different from that in existing studies. A new factor, annotation difficulty for each attribute, should be considered.

2.2 Web appearance assessment Some visual attributes (e.g., visual quality, visual complexity) significantly affect user perception on Web pages. Therefore, some researchers explored the assessment for these attributes. Wu et al. [9] leveraged machine learning theories to assess the visual quality for Web pages. So far, all existing works construct the assessment models separately for each attributes. Nevertheless, studies from HCI [11] verified that there exist strong relationships among different attributes of Web pages. The proposed algorithm will be applied to this domain and attribute correlations will be considered in our work.

2.3 Attribute learning Learning attribute categories allow prediction of color or texture types, and can also provide a mid-level cue for object classification [12]. For example, in face recognition, attribute learning can classify face according to 'white/non-white', 'young/non-young', 'smiling/non-smiling', and so on. The classification results can help us obtain detailed information about a face image. When attribute learning meets crowdsourcing machine learning, a new problem arises: how to learn from the crowd when more than one attributes a user has to annotate for each sample? This problem is just the learning problem for multi-user multi-attribute annotations investigated in this paper.

3 Methodologies

Some notations and symbols used in this paper are introduced first. Let $X = \{x_1, \ldots, x_J\}$[1] be the features of J objects. Assume that there are I attributes. Let Y be the output space whose elements are binary categories[2]. Assume that there are K users. Thus, after user annotation, all the obtained labels can be represented by a tensor $L^{I\times J\times K}$, where l_{ijk} is the label of the i-th attribute of the j-th object given by the k-th user. Let α_i be the annotation difficulty for the i-th attribute and β_{ik} be the user expertise[3] of the k-th user on the i-th annotation task.

With X and the label tensor L, our task is to (1) estimate the ground-truth label y_{ij} of x_j's i-th attribute, (2) estimate the user expertise β_{ik}, (3) estimate the annotation difficulties α_i, and (4) construct attribute classifiers.

3.1 Labeling process modeling In user labeling, a label by a user for one attribute of an object is determined by the following causal factors: (1) the user expertise for the attribute, i.e., β_{ik}, (2) the annotation difficulty for the attribute, i.e., α_i, (3) the inner/outer noise when labeling, and (4) the annotation difficulty for that object. Considering the last factor is generally ignored and can be easily added to our approach, this factor is not considered in this work.

With the above factors, a generative model (Fig. 1) is proposed to describe the labeling process. In the model, φ_{ijk} is the labeling accuracy when the i-th attribute of the j-th object is annotated by the k-th user. The parameter o_{jk} is used as the labeling noise because user labeling is also affected by some causal factors such as inner/outer

[1] The feature spaces for different attributes can be different. This study considers the feature spaces of all attribute are identical.

[2] In fact, the elements can be scores, orderings, and others. This study only considers two-category problems.

[3] User expertise reflects both the annotation ability and attitude of a user.

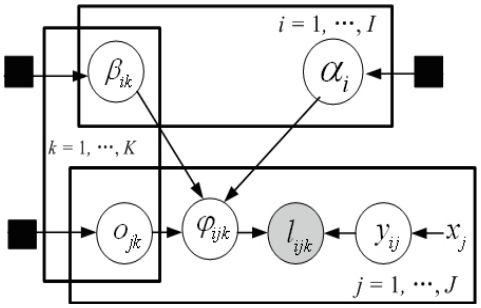


Figure 1: The graphical model of the labeling process.

environment. In Fig. 1, the user label, i.e., l_{ijk}, depends on the true label y_{ij} and the labeling accuracy φ_{ijk}. In this work, we only consider the binary classifications, namely, l_{ijk} which is either 0 or 1. Given y_{ij} and φ_{ijk}, l_{ijk} conforms to:

(3.1)
$$p(l_{ijk}|y_{ij},\varphi_{ijk}) = {\varphi_{ijk}}^{\Lambda(l_{ijk}=y_{ij})}(1-\varphi_{ijk})^{\Lambda(l_{ijk}\neq y_{ij})}$$

where Λ is the indication function. If $l_{ijk} = y_{ij}$, $\Lambda(l_{ijk} = y_{ij}) = 1$. The labeling accuracy φ_{ijk} depends on three factors: α_i, β_{ik}, and o_{jk}. Intuitively, the higher the degree of difficulty of the annotation task (the larger the value of α_i), the lower the labeling accuracy (the smaller of the value of φ_{ijk}); or the larger the user expertise (i.e., β_{ik}), the larger the value of φ_{ijk} becomes. Thus, we define:

$$\varphi_{ijk} = \frac{1}{1+e^{\alpha_i+o_{jk}-\beta_{ik}}} \tag{3.2}$$

The value of φ_{ijk} ranges from (0, 1). A larger value of φ_{ijk} indicates a higher labeling accuracy. Some observations can be obtained from Eq. (3.2).

- When the annotation task is extremely difficult (i.e., $\alpha_i \to +\infty$), φ_{ijk} approaches 0; when the annotation is extremely easy (i.e., $\alpha_i \to -\infty$)), φ_{ijk} approaches 1.
- When the user expertise is extremely high (i.e., $\beta_{ik} \to +\infty$)), φ_{ijk} approaches 1, and when the user expertise is extremely low (i.e., $\beta_{ik} \to -\infty$)), φ_{ijk} approaches 0.
- When $\alpha_i = \beta_{ik}$, φ_{ijk} depends only on o_{jk}, that is, if the user expertise and the annotation difficulty are exactly in the same level, φ_{ijk} depends only on noise.

The above observations are reasonable and in accordance with our practical experiences. In our model, α_i conforms to the Gaussian distribution, i.e., $\alpha_i \sim N(\mu_a, \sigma_a)$, where μ_a is the average annotation difficulty for all attributes and σ_a is the standard deviation. The prior distribution for β_{ik} also conforms to the Gaussian distribution: $\beta_{ik} \sim N(\mu_{uk}, \sigma_{uk})$ where μ_{uk} is the user expertise of the k-th user, and σ_{uk} is the standard deviation. The noise factor o_{jk} is also modeled using a Gaussian distribution: $o_{jk} \sim N(\mu_{nk}, \sigma_{nk})$, where μ_{nk} is the mean of the labeling noises for the k-th user, and σ_{nk} is the standard deviation.

The ground-truth label y_{ij} is modeled by a logistic sigmoid function similar to that in [14], that is,

$$p(y_{ij} = 1|x_j, w_i) = \frac{1}{1+e^{-w_i x_j}} \tag{3.3}$$

where w_i is the classifier parameter for the i-th attribute. The classification model is known as logistic regression. For brevity, let W represent the set $\{w_1, \ldots, w_I\}$.

With the above definitions, the graphical model in Fig. 1 represents the following labeling processes:

(3.4)
$$\begin{array}{l} y_{ij} \sim x_j, w_i; \alpha_i \sim N(\mu_a, \sigma_a); \beta_{ik} \sim N(\mu_{uk}, \sigma_{uk}) \\ o_{jk} \sim N(\mu_{nk}, \sigma_{nk}); \varphi_{ijk} = \frac{1}{1+e^{\alpha_i+o_{jk}-\beta_{ik}}} \\ l_{ijk} \sim y_{ij}, \varphi_{ijk}; i \in [1, I], j \in [1, J], k \in [1, K]. \end{array}$$

Let $\Theta = (\alpha_i, \beta_{ik}, o_{jk}, \mu_a, \sigma_a, \mu_{uk}, \sigma_{uk}, \mu_{nk}, \sigma_{nk})$. Given x_j, Θ, and w_i, the distribution of l_{ijk} can be written as

$$\begin{array}{l} p(l_{ijk}|x_j, w_i, \Theta) \\ = \sum\limits_{y_{ij}} \{p(l_{ijk}|y_{ij}, \varphi_{ijk}) p(y_{ij}|x_j, w_i) \\ \times p(\alpha_i|\mu_a, \sigma_a) p(\beta_{ik}|\mu_{uk}, \sigma_{uk}) p(o_{jk}|\mu_{nk}, \sigma_{nk})\} \end{array} \tag{3.5}$$

Based on Fig. 1 and Eq. (3.5), we obtain

$$\begin{array}{l} p(y_{ij} = z|l_{ijk}, x_j, w_i, \Theta) \sim p(l_{ijk}|y_{ij} = z, \Theta) \\ \times p(y_{ij} = z|x_j, w_i) \end{array} \tag{3.6}$$

where z equals 0 or 1.

3.2 Inference and learning Assuming the training instances are independently and identically distributed, the likelihood function of W and Θ given the features X and the user labels L can be factored as

$$\begin{array}{l} \mathrm{In}p(L|X, W, \Theta) = \mathrm{In}\prod\limits_i \prod\limits_j \prod\limits_k p(l_{ijk}|x_j, w_i, \Theta) \\ = \mathrm{In}\prod\limits_i \prod\limits_j \prod\limits_k \{\sum\limits_{y_{ij}} p(l_{ijk}|y_{ij}, \varphi_{ijk}) p(y_{ij}|x_j, w_i) \\ \times p(\alpha_i|\mu_a, \sigma_a) p(\beta_{ik}|\mu_{uk}, \sigma_{uk}) p(o_{jk}|\mu_{nk}, \sigma_{nk})\} \end{array} \tag{3.7}$$

To estimate the parameters W and Θ, the maximum-likelihood approach is used. The maximum-likelihood estimator is

$$(W_{ML}, \Theta_{ML}) = \arg\max_{W,\Theta}\{\mathrm{In}p(L|X, W, \Theta)\} \tag{3.8}$$

As Eq. (3.8) is difficult to optimize directly, the EM algorithm [1] is utilized. EM is an efficient iterative method that can achieve the maximum-likelihood estimator in presence of missing data. In our problem, the ground truth labels (y_{ij}) are taken as missing data. Thus y_{ij} and (W, Θ) can be estimated iteratively using EM. Once a certain number of criteria are met, the iteration is stopped. First, a new likelihood function is defined:

(3.9)
$$\begin{array}{l} \mathrm{In}p(L, Y|X, W, \Theta) = \mathrm{In}\prod\limits_i \prod\limits_j \prod\limits_k p(l_{ijk}, y_{ij}|x_j, w_i, \Theta) \\ = \mathrm{In}\prod\limits_i \prod\limits_j \prod\limits_k \{p(l_{ijk}|y_{ij}, x_j, \Theta) \times p(y_{ij}|x_j, w_i)\} \end{array}$$

The EM algorithm is run for Eq. (3.9) with two iterative steps: an Expectation (E) step and a Maximization (M) step. In the E-step, the expectation of Eq. (3.9) in terms of y_{ij} is calculated. Let l_{ij*} be $\{l_{ij1}, \cdots, l_{ijK}\}$. Let

$$\pi_{ij} = p(y_{ij}|l_{ij*}, x_j, W^{(t-1)}, \Theta^{(t-1)}) \tag{3.10}$$

and

$$\lambda_{ij} = p(y_{ij} = 0|l_{ij*}, x_j, W^{(t-1)}, \Theta^{(t-1)}) \tag{3.11}$$

where $W^{(t-1)}$ and $\Theta^{(t-1)}$ are from the previous iteration.

$$\begin{aligned} & E(\text{Inp}(L, Y|X, W, \Theta)) \\ &= \sum_i \sum_j \sum_k \sum_{y_{ij}} \{\text{Inp}(l_{ijk}, y_{ij}|x_j, W, \Theta)\pi_{ij}\} \\ &= \sum_i \sum_j \sum_k \sum_{y_{ij}} \{[\text{Inp}(l_{ijk}|x_j, y_{ij}, \Theta) + \text{Inp}(y_{ij}|x_j, w_i)]\pi_{ij}\} \\ &= E_1 + E_2 \end{aligned} \tag{3.12}$$

where

$$\begin{aligned} E_1 = \sum_i \sum_j \sum_k \{&\lambda_{ij} \times \text{Inp}(l_{ijk}|x_j, y_{ij} = 0, \Theta) \\ &+ (1 - \lambda_{ij}) \times \text{Inp}(l_{ijk}|x_j, y_{ij} = 1, \Theta)\} \end{aligned} \tag{3.13}$$

and

$$\begin{aligned} E_2 = \sum_i \sum_j \sum_k \{&\lambda_{ij} \times \text{Inp}(y_{ij} = 0|x_j, w_i) \\ &+ (1 - \lambda_{ij}) \times \text{Inp}(y_{ij} = 1|x_j, w_i)\} \end{aligned} \tag{3.14}$$

In the M-step, Eq. (3.12) is optimized to update $W^{(t-1)}$ and $\Theta^{(t-1)}$. The optimization of Eq. (3.12) can be performed by separately optimizing Eqs. (3.13) and (3.14). We first define

$$err_{ijk} = [\lambda_{ij}\Lambda(l_{ijk} \neq 0) + (1 - \lambda_{ij})\Lambda(l_{ijk} = 0)] \tag{3.15}$$

Parameter err_{ijk} is the expected labeling error for the i-th attribute of the j-th sample by the k-th user. The following equations are subsequently defined:

$$\begin{aligned} Err_i &= \sum_j \sum_k err_{ijk} \\ Err_{ik} &= \sum_j err_{ijk} \\ Err_{jk} &= \sum_i err_{ijk} \end{aligned} \tag{3.16}$$

We then have the following Lemma:

Lemma 1. The maximization of E_1 by Θ is attained by solving the following equation group:

$$\begin{aligned} & Err_i - \sum_{jk}(1 - \varphi_{ijk}) = \tfrac{\alpha_i - \mu_a}{(\sigma_a)^2}, \\ & -Err_{ik} + \sum_j (1 - \varphi_{ijk}) = \tfrac{\beta_{ik} - \mu_{uk}}{(\sigma_{uk})^2}, \\ & Err_{jk} - \sum_i (1 - \varphi_{ijk}) = \tfrac{o_{jk} - \mu_{nk}}{(\sigma_{nk})^2}, \\ & \mu_a = \sum_i \alpha_i / I, \sigma_a = \sqrt{\sum_i (\alpha_i - \mu_a)^2 / I}, \\ & \mu_{uk} = \sum_i \beta_{ik} / I, \sigma_{uk} = \sqrt{\sum_i (\beta_{ik} - \mu_{uk})^2 / I}, \\ & \mu_{nk} = \sum_j o_{jk} / J, \sigma_{nk} = \sqrt{\sum_j (o_{jk} - \mu_{nk})^2 / J}, \end{aligned} \tag{3.17}$$

Algorithm 1 Steps of LMM

Input: X, L, ε, T.
Output: Estimated ground-truth labels, W, and Θ.
Initialize: $W^{(0)}$, $\Theta^{(0)}$, $t = 1$.
Steps:

1. Calculate λ_{ij} for each y_{ij} using Eq. (11) with $W^{(t-1)}$ and $\Theta^{(t-1)}$.
2. Calculate the labeling errors ($Err_i, Err_{ik}, Err_{jk}$) using Eqs. (3.15) and (3.16).
3. Solve Eq. (3.17) to obtain new parameters $\Theta^{(t)}$.
4. Update the value of y_{ij} for each object with the following rule: $y_{ij}^{(t)} = \Lambda(1 - \lambda_{ij} > \lambda_{ij})$.
5. Learn the classifier parameters $W^{(t)}$ for each attribute based on the training set $(x_i, y_{ij}^{(t)})$ using Adaboost.
6. $t = t + 1$; If $t < T$ and $||(W^{(t)}, \Theta^{(t)}) - (W^{(t-1)}, \Theta^{(t-1)})|| > \epsilon$, goto Step 1; otherwise, exit and return $y_{ij}^{(t)}$, $W^{(t)}$, and $\Theta^{(t)}$.

where $i = 1, \cdots, I$; $j = 1, \cdots, J$ and $k = 1, \cdots, K$.

The proof of Lemma 1 is omitted due to lack of space. Equation (17) can be solved by conventional optimizing algorithms such as the Levenberg-Marquardt algorithm [8]. The following observations are obtained from Eq. (3.17): (1) The larger the value of Err_i, the larger the value of α_i is likely to be. (2) The lower the value of Err_{ik}, the larger the value of β_{ik} is likely to be. (3) The larger the value of Err_{jk}, the larger the value of o_{jk}.

Subsequently, E_2 is maximized. If w_i meets the following condition, the maximization of E_2 is attained.

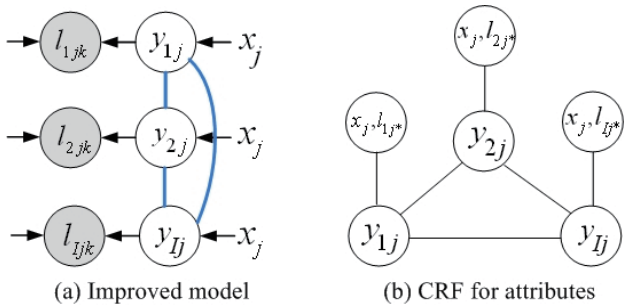
$$\begin{aligned} & p(y_{ij} = 1|x_j, w_i) = \Lambda(p(y_{ij} = 1|l_{ij*}, x_j, W^{(t-1)}, \Theta^{(t-1)}) \\ & > p(y_{ij} = 0|l_{ij*}, x_j, W^{(t-1)}, \Theta^{(t-1)})) \end{aligned} \tag{3.18}$$

where $i = 1, \cdots, I$; $j = 1, \cdots, J$ and $k = 1, \cdots, K$. The details of the proof are omitted due to lack of space. With Eq. (3.18), a heuristic procedure is introduced to optimize E_2, that is, we provide new labels for each attribute. The new label of the i-th attribute for each object is '1' if $\Lambda(p(y_{ij} = 1|l_{ij*}, x_j, W^{(t-1)}, \Theta^{(t-1)}) > p(y_{ij} = 0|l_{ij*}, x_j, W^{(t-1)}, \Theta^{(t-1)})) = 1$, and '0' otherwise. Based on Eq. (3.3) and the new training set, the classifier parameter w_i for each attribute can be learned using classical algorithms such as Adaboost [3].

The steps of the proposed learning algorithm for multi-user multi-attribute annotation (LMM) are shown in Algorithm 1.

4 Improved LMM with attribute correlations

As stated earlier, the visual attributes of an object are generally mutually correlated. For example, the visual quality of a Web page is proven to be correlated to the visual complexity [6], and visual quality also affects the level of interest of Web pages [2]. Attribute correlation has been employed in existing studies on attribute learning [4][15]. Chen et al.



(a) Improved model (b) CRF for attributes

Figure 2: The improved model (left), and the CRF model for the attibute correlations.

[4] leveraged a conditional random field (CRF) to model attribute correlations with the aim of improving the classification on each attribute. The study describes attribute correlation using CRF as its good performances reported in [4].

Figure 2 (a) shows the generative model that has been improved by considering the correlations among ground-truth labels y_{ij}. Figures 1 and 2(a) differ. In Fig. 1, the attribute labels for each object are independent, whereas in Fig. 2(a), they are mutually dependent. Let y_{*j} be $\{y_{1j}, \cdots, y_{Ij}\}$, l_{*jk} be $\{l_{1jk}, \cdots, l_{Ijk}\}$, and l_{*j*} be $\{l_{*j1}, \cdots, l_{*jK}\}$. Let

$$\pi_{*j} = \sum_{y_{*j}} p(y_{*j}|l_{*j*}, x_j, W^{(t-1)}, \Theta^{(t-1)}) \tag{4.19}$$

and

$$\phi_{ij} = \sum_{y_{*j}/y_{ij}} p(y_{*j}/y_{ij}, y_{ij} = 0|l_{*j*}, x_j, W^{(t-1)}, \Theta^{(t-1)}) \tag{4.20}$$

Then, by using Fig. 2(a), E_1 and E_2 in Eq. (3.12) are redefined as

$$E_1 = \sum_j \sum_k \sum_{y_{*j}} \{\mathrm{In}p(l_{*jk}|x_j, y_{*j}, \Theta) \times \pi_{*j}\} \tag{4.21}$$

and

$$E_2 = \sum_j \sum_k \sum_{y_{*j}} \{\mathrm{In}p(y_{*j}|x_j, W) \times \pi_{*j}\} \tag{4.22}$$

Note that

$$p(l_{*jk}|x_j, y_{*j}, \Theta) = \prod_i p(l_{ijk}|x_j, y_{ij}, \Theta) \tag{4.23}$$

Then

$$\begin{aligned} E_1 &= \sum_j \sum_k \sum_{y_{*j}} \{\mathrm{In}[\prod_i p(l_{ijk}|x_j, y_{ij}, \Theta)] \times \pi_{*j}\} \\ &= \sum_i \sum_j \sum_k \sum_{y_{*j}/y_{ij}} \phi_{ij} \times \{\mathrm{In}p(l_{ijk}|x_j, y_{ij} = 0, \Theta) \\ &+ (1 - \phi_{ij})\mathrm{In}p(l_{ijk}|x_j, y_{ij} = 1, \Theta)\} \end{aligned} \tag{4.24}$$

Algorithm 2 Steps of LMMc

Input: X, L, ε, T.
Output: Estimated ground-truth labels, W, and Θ.
Initialize: $W^{(0)}$, $\Theta^{(0)}$, $t = 1$.
Steps:

1. If $t < 3$, perform Steps 1-6 in Algorithm 1.
2. Calculate the potential functions η and ρ using Eq. (4.27) based on $W^{(t-1)}$ and $\Theta^{(t-1)}$.
3. Calculate the probabilities ϕ_{ij} in Eq. (4.20) with Eqs. (4.25) and (4.26).
4. Calculate the labeling errors $(Err_i, Err_{ik}, Err_{jk})$ using Eqs. (3.15) and (3.16).
5. Solve Eq. (3.17) to obtain $\Theta^{(t)}$.
6. Maximize Eq. (4.25) to update y_{ij} for each object and construct new training sets for each attribute.
7. Learn the classifier parameters $W^{(t)}$ for each attribute based on the training set $(x_i, y_{ij}^{(t)})$ using Adaboost.
6. $t = t + 1$; If $t < T$ and $||(W^{(t)}, \Theta^{(t)}) - (W^{(t-1)}, \Theta^{(t-1)})|| > \epsilon$, goto Step 2; else return $y_{ij}^{(t)}$, $W^{(t)}$, and $\Theta^{(t)}$.

By comparing Eq. (4.24) with Eq. (3.13), we can derive the conclusion that by merely replacing λ_{ij} (Eq. (3.11)) with ϕ_{ij} (Eq. (4.20)), the optimization of Eq. (4.24) is the same as that of Eq. (3.13). To calculate (ϕ_{ij}) Eq. (4.20), we should first calculate $p(y_{*j}|l_{*j*}, x_j, W^{(t-1)}, \Theta^{(t-1)})$. CRF is used and the model is shown in Fig. 2 (b). Similar to [4], the union distribution of all attribute labels in Fig. 2(b) is defined as follows.

$$p(y_{*j}|l_{*j*}, x_j, W^{(t-1)}, \Theta^{(t-1)}) = e^{-\Psi(y_{*j})} \tag{4.25}$$

where Ψ is a potential function defined as follows:

$$\Psi(y_{*j}) = \sum_{i \in [1,I]} \eta(y_{ij}) + \upsilon \sum_{i,i' \in [1,I]} \rho(y_{ij}, y_{i'j}) \tag{4.26}$$

where υ assigns a relative weight between the two functions η and ρ. The value of υ is set to 0.5 in the experiments. The function η is the unary potential and ρ is the edge potential. They are defined as follows:

$$\begin{aligned} \eta(y_{ij}) &= -\log[\tfrac{p(y_{ij}|l_{ij*}, x_j, W^{(t-1)}, \Theta^{t-1})}{p(y_{ij})}] \\ \rho(y_{ij}, y_{i'j}) &= -\log(p(y_{ij}, y_{i'j})) \end{aligned} \tag{4.27}$$

For E_2, a heuristic optimizing method is used. First, we maximize Eq. (4.25) to determine an optimal y_{*j}. The optimal y_{*j} is taken as the new labels for x_j. The new labels and features of all samples are used to achieve the classifier parameters W in Eq. (3.3). In our implementation, the above inference is performed within each EM step. Therefore, the probabilities such as $p(y_{ij})$ and $p(y_{ij}, y_{i'j})$ are obtained from the estimated ground-truth labels in the previous iteration. Considering that the estimated ground truth in the initial steps has low accuracy, the CRF inference is performed after the initial three iterations. The improved LMM, which

considers attribute correlations (called LMMc), is shown in Algorithm 2.

5 Experiments

Learning from multi-user multi-attribute annotation has not been investigated in previous literature. For this reason, the proposed algorithms (LMM and LMMc) are compared using several classical methods: the GLAD algorithm proposed by Whitehill et al. [5], the algorithm (termed as LFC) proposed by Raykar et al. [14], and the simple majority strategy. These algorithms are used on each attribute separately. For GLAD and LFC, their parameters conform to the suggested settings used in [5] and [14]. For LMM and LMMc, the initial values of α_i are sampled from $N(0.5, 0.5)$; the initial values of β_{ik} are sampled from $N(1, 1)$; and the initial values of o_{jk} are sampled from $N(0.05, 0.01)$. The maximum number of iterations is set to 50.

There have been no benchmark data sets for the present study. We compiled three data sets (a synthetic data set and two real-world data sets). One real-world set is for Web attribute assessment; the other is for facial attribute classification.

For the estimated ground-truth labels on training data and predicted labels on test data, the classification accuracy, which is defined as the proportion of correctly classified samples, is used . For the estimated user expertise, annotation difficulty, and noise, the orders may be more useful in practice than the absolute values. Thus the orders of the estimated values of these factors are measured according to the ground-truth orders. Taking user expertise as an example, the ground-truth orders are obtained as follows. For a specific user, we calculate the classification accuracies of the user's labels on each attribute. Subsequently, the average accuracy on all attributes is obtained. The order of the average accuracies of all users is taken as the ground-truth order of user expertise. The orders of $\mu_{u1}, \cdots, \mu_{uK}$ are taken as the estimated order. The Kendall's tau coefficient (τ) [13] is calculated for two orders:

$$\tau = (C - D)/[K(K-1)/2] \tag{5.28}$$

where C is the number of concordant pairs and D is the number of discordant pairs. If the two orders are the same, then $\tau = 1$; if the two orders are inverse, then $\tau = -1$.

Table 1: Classification accuracies of each attribute.

	Attribute 1	Attribute 2	Attribute 3	Average
LMM	0.9414	0.7318	0.7886	0.8206
LMMc	0.9398	0.7692	0.8078	0.8389
Majority	0.8880	0.5680	0.6520	0.7060
GLAD	0.9280	0.6360	0.7004	0.7548
LFC	0.9280	0.6460	0.7107	0.7616

Table 2: Kendall's tau coefficients of key factors.

	User expertise	Annotation difficulty	Noise
LMM	0.9400	0.8333	0.7215
LMMc	0.9800	0.9333	0.7873
Majority	/	/	/
LFC	0.8410	/	/
GLAD	0.8476	/	/

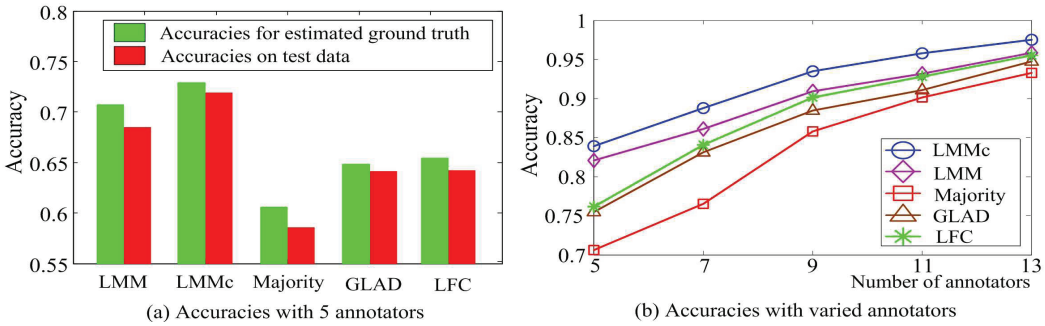
5.1 Results on the synthetic data The rules for creating the synthetic data are as follows. First, 1000 points are randomly sampled according to the uniform distribution on a square area $[0, 10] \times [0, 10]$. Next, three classification rules are used to assign three attribute labels to each point. Let ω be a random variable uniformly distributed in [-0.5, 0.5]. The first rule is $y_1 = \Lambda(-x(1) + x(2) + \omega > 0)$; the second rule is $y_2 = \Lambda(x(1) + x(2) + \omega > 10)$; and the third rule is $y_3 = \Lambda(x(1) + \omega > 5)$. Finally, the 1000 points are reselected to ensure that approximately 50% of the points have the same y_1 and y_2 labels, and approximately 60% of the points have the same y_1 and y_3 labels.

Five users are simulated. Their user expertise values are sampled from the Gaussian distributions with the means [0.4 0.8 0.9 1 1.8] and the standard deviations [0.15 0.15 0.15 0.15 0.15 0.15], respectively. The annotation difficulties are set as 0.43, 0.93, and 0.75, respectively. The noise factors are sampled from the Gaussian distributions with the means [0.02 0.04 0.06 0.03 0.07] and the standard deviations [0.01 0.01 0.01 0.01 0.01], respectively.

With the above pre-defined parameters, simulated labels for all the attributes of each sample are obtained. The data set is randomly divided into 2 fold: one is for training and the other is for testing. The randomly division is repeated 10 times. The competing algorithms are run on the data set and simulated user labels.

Figure 3(a) shows the average accuracies of the five competing methods. Both LMM and LMMc significantly outperform the other methods. The majority method obtains the most unreliable results. The accuracies of the estimated ground truth labels on the three attributes are presented in detail in Table 1. Overall, the accuracies on attribute 2 are the lowest for almost all of the algorithms. This result is caused by the lower than 0.5 accuracies obtained by attribute 2, in several runs, which produced low quality user labels.

All of the average Kendall's tau coefficients over ten runs obtained from different methods are listed in Table 2. The orderings obtained by the proposed LMMc algorithm have the highest Kendall's tau coefficients on all the three factors that affect label accuracy. In addition, LFC and GLAD do not yield annotation difficulty and noise because they deal with each task (attribute) separately and do not consider the noise factor. The Kendall's tau coefficients of the estimated orderings for the noise factor are relatively



(a) Accuracies with 5 annotators (b) Accuracies with varied annotators

Figure 3: The average accuracies (a), and the accuracy variations under different numbers of annotators.

Non-complex, high-VisQ, interesting | Non-complex, Non-high-VisQ, interesting | Complex, high-VisQ, interesting | Complex, non-high-VisQ, non-interesting | Complex, non-high-VisQ, non-interesting

Figure 4: Web pages and their attributes.

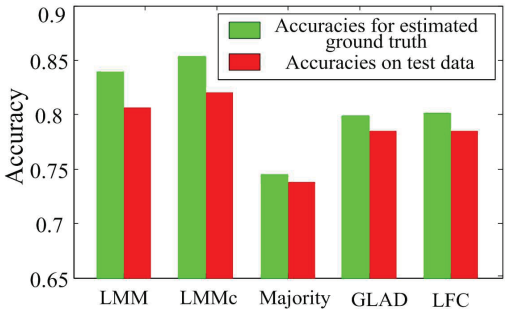


Figure 5: Average accuracies on Web data.

Table 3: Classification accuracies of each attribute.

	Visual complexity	Visual quality	Interesting	Average
LMM	0.9330	0.7560	0.8190	0.8360
LMMc	0.9280	0.7810	0.8430	0.8507
Majority	0.8370	0.6720	0.7200	0.7430
GLAD	0.9146	0.7071	0.7686	0.7968
LFC	0.9080	0.7155	0.7805	0.8013

Table 4: Kendall's tau coefficients for key factors.

	User expertise	Annotation difficulty
LMM	0.8400	0.8333
LMMc	0.8400	0.9420
Majority	/	/
GLAD	0.7133	/
LFC	0.7261	/

low. One underlying reason is that the noise factor is simply modeled with a Gaussian distribution, whereas in real situations it is a time-related factor and a Markov chain may be more suitable for modeling.

In addition, we investigate the classification accuracies when the number of reliable learners is increased. Figure 3(b) shows the accuracies when the numbers of users, whose mean user expertise values are 1 and 1.8, are increased. With the increase, the accuracies of all the methods are increased. However, to achieve the same level of accuracy, fewer users are needed for both LMM and LMMc.

5.2 Results on Web appearance evaluation In this article, 1,000 homepages are collected, mainly from company, university, and personal sites among others. In this experiment, three visual attributes of the appearances of Web pages, namely, visual complexity, visual quality, and interesting, are considered. The features in [9] are used includ-

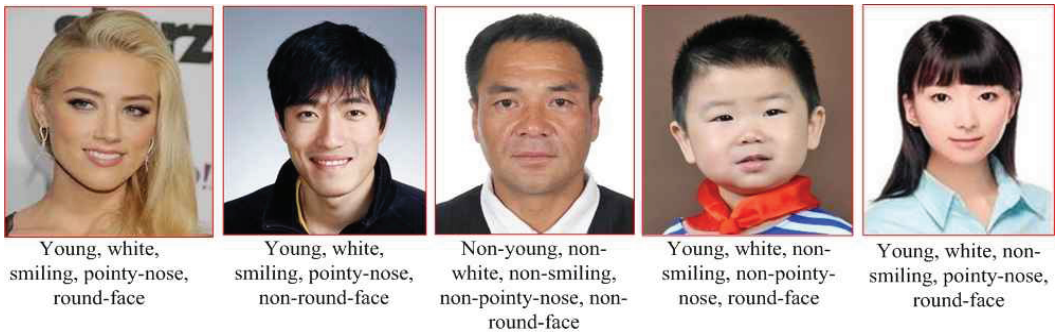


Figure 6: Facial images and their attributes.

ing the quantities of elements (e.g., texts, links, and images), color features (brightness, hue, and colorfulness), and structural features (e.g., number of major blocks).

Five students in our laboratory are invited to label the three attributes of the collected pages. The candidate categories are complex/non-complex, beautiful/non-beautiful, and interesting/non-interesting. During the user-rating session, a Web page is randomly loaded. The participant then views the page and subsequently rates it. After rating, the participant clicks "next" to load the succeeding random page. If the participant fails to rate a page within a fixed amount of time, a page is randomly selected among the unrated pages and is loaded automatically. After all the 1000 pages were rated, the participant's rating task is concluded.

To obtain high-quality labels, three professional Web designers were invited to rate the three attributes. If two designers choose the same category for each attribute for a page, then the category is taken as the label. Finally, all the labels are used as the ground truth labels. Figure 4 shows several examples.

The data set is randomly divided into 2 fold: one is for training and the other is for testing. The randomly division is repeated 10 times. Figure 5 shows the average accuracies of estimated ground-truth labels on training data and the prediction results on test data. The proposed approaches (LMM and LMMc) significantly outperform the other three methods.

Table 3 lists the accuracies of the five methods in terms of the three attributes. The comparison between LMM and LMMc reveals that the utilization of attribute correlation increases the estimation performances. The visual quality is highly correlated to both visual complexity and interest. Table 4 lists the Kendall's tau coefficients of the estimated orders of user expertise and annotation difficulty. LMM and LMMc are observed to be superior to other methods. LMMc successfully estimates the orders of task difficulty, i.e., visual quality $>$ interesting $>$ visual complexity in the collected pages.

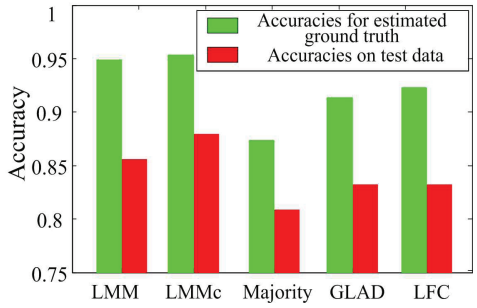


Figure 7: Average accuracies on facial data.

Table 5: Classification accuracies of each attribute.

	A 1	A 2	A 3	A 4	A5	Average
LMM	0.9675	0.9725	0.9550	0.9125	0.9400	0.9495
LMMc	0.9675	0.9700	0.9550	0.9375	0.9400	0.9540
Majority	0.8950	0.9075	0.8750	0.8575	0.8375	0.8745
GLAD	0.9350	0.9400	0.9200	0.9125	0.8625	0.9140
LFC	0.9450	0.9350	0.9300	0.9050	0.9075	0.9245

5.3 Results on facial attribute classification A facial database containing 500 images are collected from Google and Bing search engines. The features used in [12] are extracted, i.e., the gist descriptor and a 45-dimensional Lab color histogram. Five facial attributes are considered: young/non-young, white/non-white, smiling/non-smiling, pointy-nose/ non-pointy-nose, and round-face/non-round-face. The labeling procedure is similar to those in Section 5.2, that is, five students in our laboratory are invited to label all the attributes of each facial image. To obtain high-quality labels, three facial sketch professional designers are invited to classify the images. If two designers choose the same category, then the category is the label. Finally, all the labels are used as the ground truth labels. Figure 6 shows several rated examples.

The data set is randomly divided into 2 fold. One is for

Table 6: Kendall's tau coefficients for key factors.

	User expertise	Annotation difficulty
LMM	0.8150	0.7600
LMMc	0.8220	0.7600
Majority	/	/
GLAD	0.7265	/
LFC	0.8012	/

training and the other for testing. The randomly division is repeated 10 times. Figure 7 shows the average accuracies of estimated ground truth on training data and the prediction results on test data. The average labeling accuracies are listed in Table 5. For brevity, 'A1' - 'A5' denote the five attributes. LMMc achieves the highest average accuracy and slightly better than LMM.

The average Kendall's tau coefficients of different methods are listed in Table 6. Likewise, LMM and LMMc are superior to the other methods. LFD is better than GLAD. The estimation of annotation difficulty in this set is challenging because the differences of user label accuracies on the five attributes are not distinct. Therefore, the Kendall's tau coefficients of task difficulty are relatively low, as shown in Table 6.

Table 7: Attribute correlation and accuracy improvement (LMMc $>$ LMM) of the three data sets.

Data set	Correlation	Improvement
Synthetic data	0.5928	1.83%
Web data	0.5171	1.47%
Facial data	0.0965	0.45%

5.4 Discussion on attribute correlations Table 7 lists the average correlation coefficients of attributes and the accuracy improvements between LMM and LMMc of the three data sets. The results are in accordance with the motivation of LMMc. The higher the correlation, the higher the improvement.

6 Conclusions

This paper has proposed a new generative model to describe the labeling process of multi-user multi-attribute annotations. The model is successful in capturing several important factors that affect the accuracy of user labels, namely ground-truth labels, user expertise, and annotation difficulties. The logistic regression model is used to connect ground-truth labels, features, and classifiers. To infer the key factors and learn the attribute classifiers, a maximization likelihood approach is applied and an EM-based optimization algorithm is leveraged. Furthermore, the correlations among attributes are taken into account using the conditional random field theory. Experimental results on three data sets indicate that the proposed approach outperforms all the competing methods in terms of classification accuracies and estimated key factors.

7 Acknowledgement

This work is partly supported by NSFC (Grant No. 61379098, 61003115, 61103056).

References

[1] Dempster et al. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B, 39:1-38, 1977.

[2] E. Michailidou et al. Visual complexity and aesthetic per- ception of web pages. ACM Special Interest Group on the Design of Communication Conference (SIGDOC), pages 215-223, 2008.

[3] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. European Conference on Computational Learning Theory (EuroCOLT), pages 23-37, 1995.

[4] H. Chen et al. Describing clothing by semantic attributes. European Conference on Computer Vision (ECCV), pages 609-623, 2012.

[5] J. Whitehill et. al. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. Neural Information Processing Systems (NIPS), pages 2035-2043, 2009.

[6] K. Reinecke et al. Predicting users first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. Proceedings of the SIGCHI Conference on Human Factors in Computing System (CHI), pages 2049-2058, 2013.

[7] G. Mbipom and S. Harper. The interplay between web aes- thetics and accessibility. ACM SIGACCESS International Conference on Computers and Accessibility (ASSETS), pages 147-154, 2011.

[8] J. Nocedal and S. J. Wright. Numerical optimization, 2nd edition, springer. 2006.

[9] O. Wu et al. Evaluating the visual quality of web pages using a computational aesthetic approach. ACM WSDM, pages 337-346, 2011.

[10] P. Welinder et. al. The multidimensional wisdom of crowds. NIPS, pages 2424-2432, 2010.

[11] M. Pandir and J. Knight. Homepage aesthetics: The search for preference factors and the challenges of subjectivity. Interacting with Computers, 18:1351-1370, 2006.

[12] D. Parikh and K. Grauman. Relative attributes. ICCV, pages 503-510, 2011.

[13] R. Fagin et al. Comparing top k lists. SIAM Journal on Discrete Mathematics, 17:134C160, 2003.

[14] V. C. Raykar et al. Learning from crowds. JMLR, 11:1297C 1322, 2010.

[15] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. ECCV, pages 155C168, 2010.

[16] P. Donmez and J. Garbonell, A Probabilistic Framework to Learn from Multiple Annotators with Time-Varying Accuracy, SIAM SDM, pages 862-871, 2010.