Your Cart tells You: Inferring Demographic Attributes from Purchase Data

Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, Xueqi Cheng
CAS Key Lab of Network Data Science and Technology
Institute of Computing Technology, Chinese Academy of Sciences
wangpengfei@software.ict.ac.cn
{guojiafeng,lanyanyan,junxu,cxq}@ict.ac.cn

ABSTRACT

Demographic attributes play an important role in retail market to characterize different types of users. Such signals however are often only available for a small fraction of users in practice due to the difficulty in manual collection process by retailers. In this paper, we aim to harness the power of big data to automatically infer users' demographic attributes based on their purchase data. Typically, demographic prediction can be formalized as a multi-task multi-class prediction problem, i.e., multiple demographic attributes (e.g., gender, age and income) are to be inferred for each user where each attribute may belong to one of N possible classes (N≥2). Most previous work on this problem explores different types of features and usually predicts different attributes independently. However, modeling the tasks separately may lose the ability to leverage the correlations among different attributes. Meanwhile, manually defined features require professional knowledge and often suffer from under specification. To address these problems, we propose a novel Structured Neural Embedding (SNE) model to automatically learn the representations from users' purchase data for predicting multiple demographic attributes simultaneously. Experiments are conducted on a real-world retail dataset where five attributes (gender, marital status, income, age, and education level) are to be predicted. The empirical results show that our SNE model can improve the performance significantly compared with state-of-the-art baselines.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining

General Terms

Algorithm, Theory, Experimentation, Performance

Keywords

Demographic attribute, Structured Neural Embedding, multi-task multi-class prediction

WSDM'16, February 22–25, 2016, San Francisco, CA, USA.
Copyright 2015 ACM @2016 ACM. ISBN 978-1-4503-3716-8/16/02$15.00
DOI: http://dx.doi.org/10.1145/2835776.2835783 .

1. INTRODUCTION

Obtaining users' demographic attributes is crucial for retailers to conduct market basket analysis [21], adjust marketing strategy [11], and provide personalized recommendations [23, 27]. However, in practice, it is usually not easy to obtain this kind of personal data such as age, gender, and income, etc. This is particularly true for traditional offline retailers[1], who collect users' demographic information mostly in a manual way (e.g. requiring costumers to provide demographic information for registering some shopping cards). Most users are reluctant to provide detailed information or even refuse to register their demographics due to privacy and other reasons. Through our analysis on a large scale real-world retail dataset over shopping-cards where five demographic attributes (i.e., gender, marital status, income, age, and education level) are considered, as shown in Figure 1, more than 85% users only have partial attributes and 5% users have no attributes at all.

The difficulty in collecting demographic attributes in retail scenario thus raises an interesting research question: Can we inference users' demographic attributes automatically based on their purchase behaviors? Although some recent studies suggest that demographic attributes are predictable from different behavioral data, such as linguistics writing [6], web browsing [17], electronic communications [9, 13], social media [15, 29], and mobile data [4, 28, 29] to our best knowledge, seldom practice has been conducted on purchase behaviors in retail scenario.

In general, demographic prediction can be formalized as a multi-task multi-class problem, i.e., multiple demographic attributes (e.g., gender, age and income) are to be inferred for each user based on their behavioral data where each attribute may belong to one of N possible classes, $N \geq 2$ (e.g., age may refer to young, adult, or old). In the retail scenario, the behavioral data refer to users' purchase history typically recorded by the POS terminals. The prediction task may take two forms: 1) Given a set of users with partial demographic attributes, how to predict the unknown attributes? (referred to as Partial-Label prediction) 2) Given a set of users with partially/fully labeled attributes, how to predict the demographic attributes for new users? (referred to as New-User prediction)

[1] In this work, we mainly focus on traditional retailers in offline business rather than those in online e-commerce, where no additional behavioral data rather than transactions is available for analysis. Hereafter we will use retail/retailer for simplicity when there is no ambiguity.

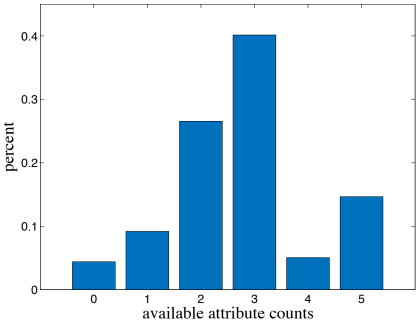

Figure 1: Distribution of demographic attribute count over users with shopping card. X axis stands for the number of available attributes per user, y axis indicates the proportion of users.

Previous work on demographic prediction usually predicts different attributes independently based on manually defined features [4, 19, 22, 28, 29]. For example, Zhong et al. [29] tried to predict six demographic attributes (i.e., gender, age, education background, sexual orientation, marital status, blood type and zodiac sign) separately using spatial, temporal and location knowledge features. However, manually defined features usually require professional knowledge and often suffer from under specification. Meanwhile, by predicting each attribute independently, one may not be able to leverage the potential correlations between different attributes (e.g., correlation between age and marital status). Some recent studies proposed to take the relations between different attributes into account [4, 28]. For example, Dong et al. [4] employed a Double Dependent-Variable Factor Graph model to predict gender and age simultaneously. Zhong et al. [28] attempted to capture pairwise relations between different tasks when predicting six demographic attributes from mobile data. However, these methods still rely on various human-defined features which are often costly to obtain.

To tackle the above problems, in this paper we propose a novel Structured Neural Embedding (SNE) model to automatically learn the representations (i.e., features) from users' purchase data for predicting multiple demographic attributes simultaneously in retail scenario. Specifically, we characterize each user by his/her purchase history using the bag-of-item representations. We then map each item to a vector in a continuous space, aggregate all the item vectors to form the user representation, and further feed this representation to a log-bilinear model for structured prediction. As compared with previous methods, the proposed SNE model enjoys the following two merits: 1) The features of users are automatically learned towards the goal of the prediction tasks. 2) By employing a structured prediction model, we can fully leverage the potential correlations between different attributes to improve the prediction accuracy. The proposed SNE model can be learned efficiently using the stochastic gradient descent (SGD) method.

We conduct extensive experiments on a real-world retail dataset to demonstrate the effectiveness of the proposed method. Some state-of-the-art baseline methods on demographic prediction and multi-task learning are taken into comparison. We tested different methods on both the Partial-Label and New-User prediction problems. The empirical results demonstrated that our approach is more effective than all the baseline methods.

Overall, the major contributions of our work are as follows:

- We make the first attempt to investigate the prediction power of users' purchase data for demographic prediction in retail scenario.
- We propose a novel SNE model for the multi-task multi-class prediction problem which can not only learn the data representations automatically but also capture the relations between different attributes in a structured way.
- We conduct extensive experiments on a real-world retail dataset to demonstrate the effectiveness of the proposed SNE model as compared with different baseline methods.

The rest of the paper is organized as follows. After a summary of related work in Section 2, we describe the problem formalization of demographic prediction in retail scenario in Section 3. In section 4 we present our proposed model in detail. Section 5 concludes this paper and gives the future work.

2. RELATED WORK

In this section we briefly review three research areas related to our work: demographic prediction, multi-task & multi-class prediction, and representation learning.

2.1 Demographic Prediction

Demographic prediction has been studied in different scenarios in academia. Early work on demographic prediction attempted to predict demographic attributes based on the linguistics writing and speaking. For example, Schler et al. [22] found that there are significant differences in both writing style and content between male and female bloggers as well as among authors of different ages. Otterbacher [19] used logistic regression model to infer users' gender based on content of reviews.

Later, the digital communication and Internet offered new opportunities for inferring demographic attributes. Different approaches have been proposed to infer demographic attributes based on users' browsing history [9, 17]. Torres [5] found that the clicked pages were correlated with the demographic characteristics of users. Hu et al. [9] calculated demographic tendency of web pages, and modeled users' demographic attributes through a discriminative model. In [2], Bi et al. infers the demographic attributes of search users based on the models training on the independent social datasets. They demonstrated that by leveraging social and search data in a common representation, they can achieve better accuracy in demographic prediction.

Recently, the fast development of online social networks and mobile computing technologies accumulated large scale of user data, making it possible and also valuable to infer users' demographic attributes in these scenarios. Mislove [15] found that users with common profiles were more likely to be friends and often formed a dense community. Zhong et al. [28] proposed a supervised learning framework to predict users' demographic attributes based on mobile data. Dong et al. [4] focused on micro-level analysis of the mobile networks to infer users' demographic attributes. Culotta et al. [3] fitted a regression model to predict users' demographic attributes using information on followers of each website on Twitter.

As we can see, most existing work on demographic prediction focused on designing different features for the prediction tasks. Besides, to the best of our knowledge, seldom practice has been conducted on demographic prediction based on purchase behaviors in retail scenario.

2.2 Multi-task & Multi-class Prediction

The idea of learning multiple tasks together is to improve the generalization performance by leveraging the information contained in the related tasks. A typical way for this purpose is to learn tasks in parallel while using a shared representation [4, 10, 28]. Many algorithms have been proposed to solve multi-task learning with various kernels and regularizers to address the correlation between tasks. For example, Micchelli et al. [14] discussed how different kernels can be used to model relations between tasks and presented linear multi-task learning algorithms. Evgeniou et al. [7] presented an approach to multi-task learning based on the minimization of regularization functions.

Meanwhile, multi-class classification is the problem of classifying instances into one of the more than two classes. Usually two ways are used to solve this kind of problem: 1) one-against-one [1], which builds a classification for each pair of classes; and 2) one-against-all [25], which creates one binary problem for each of classes.

In this paper, we formalize the demographic prediction in retail scenario as a multi-task multi-class problem, where we propose to solve it by using structured prediction based on automatic representation learning.

2.3 Representation Learning

Learning representations of the data makes it easier to extract useful information when building classifiers or other predictors. That is why representation learning has attracted more and more attention and become a field in itself in the machine learning community.

Many remarkable empirical successes have been achieved based on representation learning in various applications in both academia and industry. For example, in speech recognition and signal processing, Alex Graves et al. [8] designed a deep recurrent neural network for speech recognition and obtain the best score on benchmark. In object recognition, Krizhevsky et al. [12] proposed to use convolutional neural network to classify image and achieved the record-breaking results. In natural language processing, Mnih [16] proposed three graphical models to define the distribution of next word in a sequence by using distributed representations.

In this work, we propose to use representation learning for demographic prediction in retail scenario, a new application area where representation learning might be helpful.

3. OUR APPROACH

In this section, we first introduce the formalization of demographic prediction problem in retail scenario. We then talk about the key idea of our approach. After that , we describe the proposed SNE in detail. Finally, we present the learning and prediction procedure of SNE and give some discussions on the model.

Table 1: List of demographic attributes used in this work

Attributes	Values
gender	male, female
age	young(14-24), adult(25-34), middle-age(35-49), old(>50)
marital status	single, married
income	ultra-low(<2k/month), low(2k-4k/month) medium(4k-6k/month), high(>6k/month)
education level	doctor, master, bachelor, college, high school, middle school

3.1 Problem Formalization

In our work, we aim to predict multiple demographic attributes based on users' behavioral data in retail scenario. Specifically, each user can be characterized by his/her purchase history, i.e., a set of items. The demographic attributes we are interested in include gender, age, marital status, income, and education level, which are useful signals for market basket analysis. The values each attribute may take are listed in Table 1, and for each attribute the possible values are exclusive. For each user, given part/none of his/her attributes, we want to predict all the unknown attributes.

Obviously, the above prediction task can be formalized as a multi-task multi-class problem. Specifically, let $T = \{T_1, T_2, \dots, T_K\}$ be a set of multi-class prediction tasks (i.e., predicting demographic attributes), where each task $T_k \in T$ is associated with C_k classes (i.e., multiple attribute values), $C_k \geq 2, k = 1, 2, \dots, K$. The total class number across all the tasks is $C = \sum_{k=1}^{K} C_k$. Let U be a set of $|U| = M$ users and I be a set of $|I| = N$ items. Each user is represented by $(x^{(i)}, y^{(i)}), i = 1, 2, \dots, M$, where $x^{(i)}$ denotes the purchase history of the i-th user, and $y^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \dots, y_K^{(i)}\}$ denotes the set of attribute labels of the i-th user. Note here $y_k^{(i)}$ denotes the attribute label under the k-th task $T_k \in T$ for the i-th user, which takes value from $\{1, 2, \dots, C_k\}$.

Given the notations defined above, we define the following two prediction problems:

- **Partial-Label prediction:** Given a set of users with partial demographic attributes, the objective is to learn a function to predict the remaining unknown attributes

$$f : X, Y^L \rightarrow Y^U$$

  where Y^L and Y^U denote the observed attributes and that to be predicted over the same set of users X respectively.

- **New-User prediction:** Given a set of users with partially/fully labeled attributes, the objective is to learn a function to predict the demographic attributes for new users

$$f : X^L, Y^L, X^N \rightarrow Y^N$$

  where X^L and Y^L denote the purchase histories and attributes of labeled users, X^N and Y^N denote the purchase history and the attributes of the new users. Note that here $X^L \cap X^N = \emptyset$.

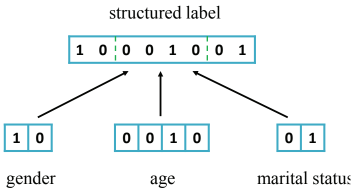

Figure 2: Structured representation of multi-task multi-class predictions.

We can see that the first problem focuses on predicting the missing attributes for the same users used in training, while the second one emphasizes the generalization ability of the prediction model to new users.

3.2 Key Idea

In our work, we introduce a novel structured prediction method based on representation learning to solve the above demographic prediction problems. The key motivation of this model comes from the following two folds.

Firstly, a fundamental problem in demographic prediction based on users' behavior data is how to represent users. Many existing work investigated different types of human defined features [4, 19, 28]. However, it is usually costly to define features manually since expertise knowledge is required and one has to do the same job task by task. Moreover, human defined features may often suffer from under specification since it is difficult to identify those hidden complicated factors for prediction tasks. Some recent work employs unsupervised feature learning methods [9, 13, 29], like Singular Vector Decomposition (SVD), to automatically extract low-dimension features from the raw data. However, the features learned in an unsupervised manner may not be optimal for the prediction tasks. Therefore, in this work we proposed to automatically learn representations of users for demographic prediction in a supervised way.

Secondly, as demographic prediction can be viewed as a multi-task problem, there might be correlations among different tasks that can be leveraged to improve the prediction accuracy. For example, users' marital statuses are more likely to be single if they are young, and a better educated person may have more chance to have higher income. However, most previous work treated different attributes as separate prediction tasks [3, 13, 29], thus ignored the correlations among these attributes (detailed discussion please refer to Section 3.4). In our work, we try to explicitly model the correlation information between different tasks by turning the multiple multi-class prediction tasks into a single structured prediction task.

Here we take the attributes gender, age, and marital status in our problem as an example. These three prediction tasks are 2-class, 4-class, and 2-class classification problems respectively. To turn them into a structured problem, we encode each task's label by a one-hot representation, and concatenate these labels to generate a single structured label, as shown in Figure 2. The benefit of this structured formalization is obvious, as the correlation among tasks can now be explicitly encoded in this label vector, e.g. label vector referring to young and single is much more popular than that referring to young and married. Therefore, by learning based on such structured labels, we can directly learn the features that are useful for revealing the correlation between multiple prediction tasks.

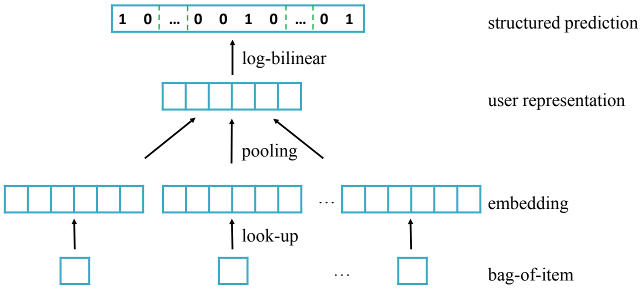

Figure 3: Architecture of SNE model.

3.3 Structured Neural Embedding Model

Based on the above two ideas, we now present the proposed Structured Neural Embedding (SNE) model in detail. In retail scenario, each user is characterized by his/her purchase history, i.e., a set of items. In SNE, we take the bag-of-item representation as the user input, and map each item to a vector in a continuous space. We then aggregate all the item vectors using some operator to form the user representation, and feed this representation to a log-bilinear model for the structured prediction. The architecture of our model is shown in Figure 3.

More formally, let $\mathbf{V}^I = \{\vec{v}_j^I \in \mathbb{R}^{D_v} | j = 1, \ldots, N\}$ denote all the item vectors in a D_v-dimension continuous space. For the i-th user with purchase history $x^{(i)}$ (i.e., a set of purchased items), we aggregate the item vectors to form the user representation by

$$\vec{v}^{(i)} = f(\vec{v}_j^I : j \in x^{(i)})$$

where $f(\cdot)$ denotes the aggregation function. In our work, we investigate three types of pooling functions as the aggregation operator for computational efficiency.

- *unique pooling*:

  $$\vec{v}^{(i)} = f_{uniq}(\vec{v}_j^I : j \in x^{(i)}) = \frac{1}{|uniq(x^{(i)})|} \sum_{j \in uniq(x^{(i)})} \vec{v}_j^I$$

  where $uniq(x^{(i)})$ represents the set of unique items purchased by i-th user.

- *average pooling*:

  $$\vec{v}^{(i)} = f_{avg}(\vec{v}_j^I : j \in x^{(i)}) = \frac{1}{|x^{(i)}|} \sum_{j=1}^{|x^{(i)}|} \vec{v}_j^I$$

- *max pooling*:

  $$\vec{v}^{(i)} = f_{max}(\vec{v}_j^I : j \in x^{(i)}) = \begin{bmatrix} max(\vec{v}_1^I[1], \ldots, \vec{v}_{|x^{(i)}|}^I[1]) \\ max(\vec{v}_1^I[2], \ldots, \vec{v}_{|x^{(i)}|}^I[2]) \\ \vdots \\ max(\vec{v}_1^I[D_v], \ldots, \vec{v}_{|x^{(i)}|}^I[D_v]) \end{bmatrix}$$

  Where $\vec{v}_j^I[l]$ denotes the l-th dimension in $\vec{v}_j^I$.

Based on the aggregated vector of the i-th user, SNE defines the probability of assigning demographic attributes $y^{(i)}$ to the user via a log-bilinear model:

$$p(y^{(i)} | x^{(i)}) = \frac{exp(\vec{v}^{(i)\top} \mathbf{W} \vec{y}^{(i)})}{\sum_{\vec{y} \in \mathcal{Y}} exp(\vec{v}^{(i)\top} \mathbf{W} \vec{y})} \quad (1)$$

where $\vec{y}^{(i)} \in \{0,1\}^C$ denotes the structured vector of $y^{(i)}$, $\mathcal{Y}$ denotes all the possible structured vectors of different combinations of attributes, and $W = \mathbb{R}^{D_v \times C}$ denotes the interaction matrix. Note that since each task is a multi-class problem (i.e., only one class can be assigned to each task), the total size of $\mathcal{Y}$ is $|\mathcal{Y}| = \prod_{k=1}^{K} C_k$.

It is worth noting that when $y^{(i)}$ only contains partial attributes, we will construct a set of structured vectors $\vec{y}^{(i)}$ corresponding to the same $y^{(i)}$ by fixing the values for the known attributes and enumerating all the possible values for the missing attributes. Let $\mathcal{Y}^{(i)}_{partial}$ denote the set of corresponding vectors, the computation of $p(y^{(i)}|x^{(i)})$ becomes

$$p(y^{(i)}|x^{(i)}) = \frac{\sum_{\vec{y}^{(i)} \in \mathcal{Y}^{(i)}_{partial}} exp(\vec{v}^{(i)\top} \mathbf{W} \vec{y}^{(i)})}{\sum_{\vec{y} \in \mathcal{Y}} exp(\vec{v}^{(i)\top} \mathbf{W} \vec{y})} \tag{2}$$

Since the enumeration of all the possible values of missing attributes will appear in both numerator and denominator in Equation (2), they can be eliminated for computation simplicity. Therefore, we can obtain a general version of Equation (1) handling both partial and full attributes

$$p(y^{(i)}|x^{(i)}) = \frac{exp(\vec{v}^{(i)\top} \mathbf{W}_c \vec{y}^{(i)}_c)}{\sum_{\vec{y}_c \in \mathcal{Y}_c} exp(\vec{v}^{(i)\top} \mathbf{W}_c \vec{y}_c)} \tag{3}$$

Where the subscript c denotes a *compact* version of the variable. A compact structured vector $\vec{y}^{(i)}_c$ is a concatenation of one-hot representations of known attributes in $y^{(i)}$, while a compact interaction matrix $\mathbf{W}_c$ is formed by removing the columns corresponding to the missing attributes from the original $\mathbf{W}$.

The objective function of SNE is then defined as the log likelihood over all the users as follows:

$$\ell_{SNE} = \sum_{i=1}^{M} \log p(y^{(i)}|x^{(i)}) - \lambda \|\Theta\|_F^2 \tag{4}$$

where λ is the regularization constant and Θ are the model parameters (i.e. $\Theta = \{\mathbf{W}, \mathbf{V}^I\}$).

3.4 Learning and Prediction

Learning SNE model involves maximize the objective function defined in Equation (4). However, the direct optimization is intractable due to the high computational cost of the normalization term which is proportional to $|\mathcal{Y}|$. Therefore, we adopt the negative sampling technique [20, 24] for efficient optimization, which approximates the original objective ℓ_{SNE} with the following objective function:

$$\ell_{NEG} = \sum_{i=1}^{M} \Big(\log \sigma(\vec{v}^{(i)\top} \mathbf{W} \vec{y}^{(i)}) + k_{neg} \cdot \mathbb{E}_{\vec{y}^{neg} \sim P_{\mathcal{Y}}} [\log \sigma(-\vec{v}^{(i)\top} \mathbf{W} \vec{y}^{neg}))] \Big) - \lambda \|\Theta\|_F^2$$

where $\sigma(x)$ is the logistic function $\sigma(x) = 1/(1+e^{-x})$, k_{neg} is the number of "negative" samples, and $\vec{y}^{neg}$ is the sampled structured vector, drawn according to the noise distribution $P_{\mathcal{Y}}$ which is modeled by empirical distribution over all possible attribute combinations. As we can see, the objective of SNE with negative sampling aims to differentiate the ground truth from noise by increasing the probability of the correct label combination given the user input and deceasing that of any wrong combinations.

We then apply stochastic gradient descent algorithm to maximize the new objective function for learning the model. The updating algorithm is shown in Algorithm 1.

Algorithm 1 Learning algorithm of SNE model

```
Initialize model $\Theta$: $\{\mathbf{W}, \mathbf{V}^I\}$ randomly
t=0
repeat
    $t \leftarrow t+1$;
    for i=1,...,|U| do
        $\vec{v}^{(i)} = f(\vec{v}^I_j : j \in x^{(i)})$
        for each $\vec{w}$ in $\mathbf{W}_c \vec{y}^{(i)}_c$ do
            $\vec{w} \leftarrow \vec{v}^{(i)} \sigma(-\vec{v}^{(i)} \sum_{\vec{w}} \vec{w})$
        end for
        for k=1,...,n do
            for each $\vec{w}$ in $\mathbf{W}_c \vec{y}^{neg}_c$ do
                $\vec{w} \leftarrow -\vec{v}^{(i)} \sigma(\vec{v}^{(i)} \sum_{\vec{w}} \vec{w})$
            end for
        end for
        for each $j \in x^{(i)}$ do
            $\vec{v}^{(i)}_j \leftarrow \sum_{\vec{w} \in \mathbf{W}_c \vec{y}^{(i)}_c} \vec{w} \sigma(-\vec{v}^{(i)} \sum_{\vec{w} \in \mathbf{W}_c \vec{y}^{(i)}_c} \vec{w}) - \sum_k^n \sum_{\vec{w} \in \mathbf{W}_c \vec{y}^{neg}_c} \vec{w} \sigma(\vec{v}^{(i)} \sum_{\vec{w} \in \mathbf{W}_c \vec{y}^{neg}_c} \vec{w})$
        end for
    end for
until converge or t> num
return $\mathbf{W}, \mathbf{V}^I$
```

With the learned item representations $\mathbf{V}^I$ and interaction matrix $\mathbf{W}$, the prediction process is to find the best attributes for a given user according to

$$y* = arg \max_{y \in \mathcal{Y}} p(y|x)$$

For Partial-Label prediction problem, part of the attributes are fixed and we want to decide the rest; For New-User prediction problem, we need to predict the whole set of attributes. These two problems can be solved similarly in an efficient way. We first re-write the log-bilinear model as follows

$$\begin{aligned} p(y|x) &= \frac{exp(\vec{v}^\top \mathbf{W} \vec{y})}{\sum_{\vec{y} \in \mathcal{Y}} exp(\vec{v}^\top \mathbf{W} \vec{y})} \\ &\propto exp(\vec{v}^\top \mathbf{W} \vec{y}) \\ &= \sum_{j: I(\vec{y}[j]=1)} v^\top \mathbf{W}_{*j} \end{aligned} \tag{5}$$

Where $\vec{y}[j]$ denotes the j-th entry in $\vec{y}$, $\mathbf{W}_{*j}$ denotes the j-th column of the interaction matrix, and $I(\cdot)$ denotes the indicator function.

The Equation (5) shows that the probability of an attribute set is proportional to the sum of scores (i.e.,$v^\top \mathbf{W}_{*j}$) corresponding to the attribute assignments (i.e., $\vec{y}[j]$ set as 1). Since in our work each task is a multi-class problem where only one class can be assigned, the best attribute set is then a combination of assignments with the highest score from each task given the input. In this way, for each user input, we only need to conduct a forward computation to generate the scores for each attribute entry, and select the highest one for each task as the final prediction. For Partial-Label problem, we simply select for those missing attributes while leaving the known attributes fixed.

3.5 Discussion

In this section we try to compare the difference between our structure learning model and conventional multi-task learning methods.

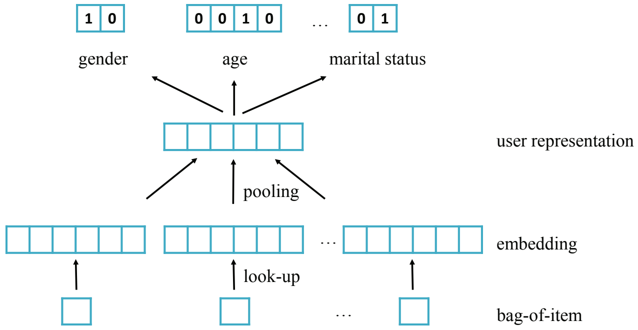

Figure 4: Architecture of joint model for multi-task multi-class prediction.

In conventional multi-task learning, a joint model is typically employed to learn several related tasks at the same time by using a shared representation, as shown in Figure 4. In this model, each task is viewed as a separate prediction problem and the objective function is a sum over these tasks:

$$\ell_{JOINT} = \sum_{i=1}^{M}\sum_{k=1}^{K}\log p(y_k^{(i)}|x^{(i)}) - \lambda\|\Theta\|_F^2$$

Where $y_k^{(i)}$ denotes the k-th attribute of the user.

The joint model can improve the prediction performance by learning the commonality among multiple tasks through the shared representation. This is the same as our SNE model. The major difference lies in how we model the prediction tasks. In joint model, each attribute is modeled as a separate prediction task, thus the correlation between attributes (like seeing one attribute makes it more likely to see another simultaneously) is ignored. Some multi-task learning methods try to consider the correlation among tasks by kernels or regularizers [7, 30], but they usually rely on explicit knowledge of relationships among tasks. While in our structured model, we turn the multiple prediction tasks into a single structured prediction task. Therefore, we can see that in the objective function of the SNE model (Equation (4)), there is no summation over the tasks. The learning over the structured label vector makes us be able to learn the important patterns revealing the correlation among multiple tasks.

One may argue that the structured formalization makes the prediction task more difficult than original separate tasks since the output space becomes much larger. However, with the large scale user behavioral data, this sparse problem can somehow be alleviated and our experimental results show that we can indeed improve the prediction performance by our structured formalization.

4. EXPERIMENTS

In this section, we conduct empirical experiments to demonstrate the effectiveness of our proposed SNE model on demographic attribute prediction in retail scenario. We first introduce the experimental settings. Then we analyze the effect of different aggregation operators and negative sampling strategies to our SNE model. Finally, we compare our SNE model to the baseline methods to demonstrate the effectiveness in both Partial-Label prediction and New-User prediction scenarios.

Table 2: Distribution of demographic attributes on BeiRen dataset.

attributes	value	distribution
gender	male	29%
	female	71%
age	young	7%
	adult	39%
	middle age	42%
	old	12%
martial status	single	40%
	married	60%
income	ultra-low	17%
	low	50%
	medium	19%
	high	14%
education level	doctor	<1%
	master	8%
	bachelor	49%
	college	3%
	high school	12%
	middle school	28%

4.1 Experimental Settings

Here we introduce the experimental settings including the dataset, baseline methods, and evaluation metrics.

4.1.1 Dataset

We conduct our empirical experiments over a real world large scale retail dataset, namely BeiRen dataset[2]. This dataset comes from a large retailer[3] in China, which records its supermarket purchase histories during the period from 2012 to 2013. It contains $49,290,149$ transactions over $220,828$ items belonging to $1,206,379$ users. For research purpose, the dataset has been anonymized with all the users and items denoted by randomly assigned IDs for the privacy issue. We first conduct some pre-process on the BeiRen dataset. We only keep the users who have all the five demographic attributes (i.e., gender, age, marital status, income, and education level) provided. We then extract all the transactions related to these users to form their purchase histories, and remove all the items bought by less than 5 times. After preprocessing, the dataset contains $61,097$ distinct items and $57,693$ distinct users with full attributes. In average, each user has bought about 110.6 distinct items. The detailed distribution of different attributes are listed in Table 2.

For experiments on Partial-Label prediction, we randomly set the observed ratio of users' demographic attributes from 10% to 90% with the step length as 10%. All the users' with their partially observed attributes and purchase histories are taken as training data, and the task is to predict the hidden attributes of these users.

For experiments on New-User prediction, we split the dataset into two non overlapping set, i.e. a training set and a testing set, with the ratio $9:1$. The resulting training set contains $51,923$ users, and the test set contains $5,770$ users.

4.1.2 Baseline Methods

We evaluate our model by comparing with several state-of-the-art methods on the demographic attribute prediction task:

- POP: The most popular combination of demographic attributes in the training set is taken as prediction.

[2] http://www.bigdatalab.ac.cn/benchmark/bm/bd?code=SNE
[3] http://www.brjt.cn/

Table 3: Comparison of different aggregation operators in SNE with varied observed attribute ratio from 10% to 90%

ratio	wPrecision			wRecall			wF1			Hamming Loss		
	SNE$_{uniq}$	SNE$_{max}$	SNE$_{avg}$	SNE$_{uniq}$	SNE$_{max}$	SNE$_{avg}$	SNE$_{uniq}$	SNE$_{max}$	SNE$_{avg}$	SNE$_{uniq}$	SNE$_{max}$	SNE$_{avg}$
10	0.071	0.163	**0.166**	0.034	0.072	**0.079**	0.047	0.106	**0.115**	0.543	0.455	**0.445**
20	0.169	0.228	**0.239**	0.096	0.108	**0.121**	0.123	0.147	**0.161**	0.492	0.452	**0.433**
30	0.214	0.294	**0.310**	0.129	0.138	**0.168**	0.161	0.188	**0.221**	0.487	0.455	**0.431**
40	0.247	0.318	**0.320**	0.181	0.214	**0.224**	0.208	0.256	**0.264**	0.488	0.441	**0.428**
50	0.338	0.368	**0.376**	0.247	0.275	**0.283**	0.302	**0.321**	0.319	0.455	**0.427**	0.429
60	0.419	0.407	**0.410**	0.307	0.339	**0.343**	0.354	0.370	**0.372**	0.457	**0.427**	0.428
70	0.451	0.448	**0.481**	0.370	0.402	**0.403**	0.407	0.424	**0.431**	0.457	0.428	**0.425**
80	0.485	0.503	**0.514**	0.428	0.462	**0.471**	0.455	0.482	**0.491**	0.451	0.431	**0.423**
90	0.528	0.534	**0.556**	0.491	0.511	**0.530**	0.511	0.522	**0.543**	0.451	0.432	**0.421**

Obviously this heuristic baseline ignores users' purchase history, and only relies on the correlations among demographic attributes for prediction.

- SVD-Single: A singular value decomposition (SVD)[4] is first conducted over the user-item matrix to obtain low dimensional representations of users. Then a logistic model is learned over the low dimensional representation to predict each demographic attribute separately. This method has been widely used in demographic attribute prediction [9, 17, 29].

- SVD-Structured: Different from SVD-Single, a structured learning model (i.e., log-bilinear model) is used to predict multiple demographic attributes based on the low dimensional representations obtained by SVD decomposition.

- JNE: The joint neural embedding (JNE) model is a typical multi-task learning method as discussed in Section 3.5. All the tasks are assumed to share the same latent representation of the user, and a joint model is employed to predict multiple attributes in parallel.

For both SVD and neural embedding based methods, we run several times with random initialization by setting the dimensionality as 100. We compare the average results of different methods and demonstrate the results in the following sections.

4.1.3 *Evaluation Metrics*

We employ the following evaluation metrics to evaluate the performance of demographic prediction methods against the groundtruth.

- Hamming Loss: the hamming loss is a wildly used metric [18, 26], which calculates how many times an instance-label pair is misclassified. The metric is calculated as follows:

$$\text{Hamming Loss} = \frac{1}{|U|}\sum_i \frac{|y^{*(i)} \triangle y^{(i)}_{test}|}{|y^{(i)}_{test}|}$$

  Where $\triangle$ stands for the symmetric difference between two sets, $y^{(i)}_{test}$ denotes the set of attributes to be predicted for the i-th user, and $y^{*(i)}$ denotes the set of predicted attributes. Note that Hamming Loss is an attribute level metric, which takes each demographic attribute independently for evaluation. As we can see, The smaller the value of Hamming Loss is, the better performance the model obtains.

- Weighted F1: we follow the idea in [4] to use weighted F1 as an evaluation metric since we consider each class is as important as each other. The weighted F1 is computed as follows:

$$\text{wPrecision} = \frac{1}{|\mathcal{Y}|} \frac{\sum_i I(y^{*(i)} = y^{(i)}_{test})}{\sum_i I(y = y^{(i)}_{test})}$$

$$\text{wRecall} = \frac{1}{|U|}\sum_i I(y^{*(i)} = y^{(i)}_{test})$$

$$\text{wF1} = \frac{2 \times \text{wPrecision} \times \text{wRecall}}{\text{wPrecision} + \text{wRecall}}$$

  where $I(\cdot)$ is an indicator function. Note that these weighted metrics are more strict than Hamming Loss in that the prediction for a user is correct only when the set of attributes are all correctly predicted. As we can see, the weighted precision is the prediction accuracy in the label combination view while the weighted recall is the prediction accuracy in the user view.

4.2 Study of the SNE Variations

We first analyze the multiple variations of the proposed SNE model, including the aggregation operators and negative sampling strategies.

4.2.1 *Effect of Aggregation Operators*

In SNE model, we can employ different aggregation operators to obtain users' representations from item vectors. In this work, we introduced three types of pooling functions, namely unique pooling, average pooling, and max pooling. Here we study which kind of operator works better with respect to the demographic prediction. We denote the corresponding SNE model as SNE$_{uniq}$, SNE$_{avg}$, and SNE$_{max}$, and show the performance results over the Partial-Label prediction problem in Table 3.

As we can see, among all the three variations of SNE model, SNE$_{avg}$ performs better than both SNE$_{uniq}$ and SNE$_{max}$ in terms of different measures in most cases. The results indicate that the frequency information of items is important for demographic prediction, which is ignored in both SNE$_{uniq}$ and SNE$_{max}$. This is reasonable since someone who frequently buys wine and cigarette is more likely to be an adult man than someone happens to buy these items.

4.2.2 *The Impact of Negative Sampling*

To learn the proposed SNE model, we employ negative sampling procedure for optimization. One parameter in this procedure is the number of negative samples we draw each time, denoted by k_{neg}. Here we investigate the impact of the sampling number k_{neg} to the performance of Partial-Label

[4] http://tedlab.mit.edu/∼dr/SVDLIBC/

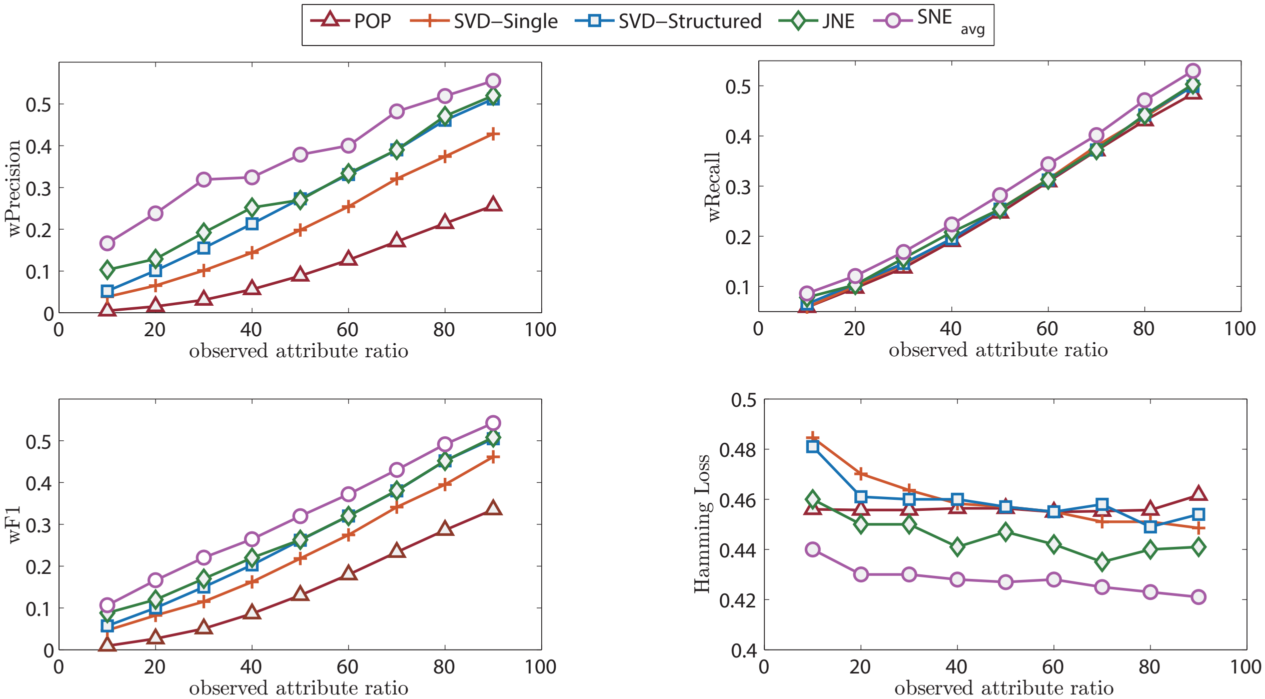

Figure 5: Performance comparison over different methods on Partial-Label prediction in terms of wPrecision, wRecall, wF1 and Hamming Loss.

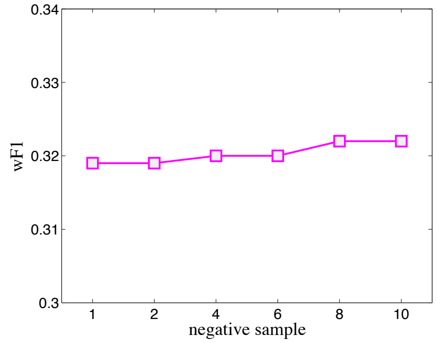

Figure 6: Performance variation with the increase of negative samples in terms of wF1 with the observed attribute ratio set as 50%.

prediction. Specifically, we tried $k_{neg} \in \{1, 2, 4, 6, 8, 10\}$, and depict the test performance of SNE$_{avg}$ in terms of wF1 against k_{neg} in Figure 6, where the observed attribute ratio is set as 50%.

As we can see, the test performance is quite stable with the increase of the negative sampling number. We have also tried other observed attribute ratios, and find similar stable results. Therefore, the results demonstrate that the optimization of the SNE model is not sensitive to the negative sampling number. We set $k_{neg} = 1$ in our learning procedure for efficiency.

4.3 Performance Comparison on Demographic Prediction

Now we compare our SNE model with the state-of-the-art baseline methods on demographic predictions, including both Partial-Label prediction and New-User prediction. Here we choose SNE$_{avg}$ as the representative of our model for clear comparison.

4.3.1 Partial-Label Prediction

The results of different methods on Partial-Label prediction are shown in Figure 5. We have the following observations:

(1) It is unsurprising to see that with the increase of the observed label ratio (i.e., more observed attributes in learning), all the methods can obtain better performances in prediction. (2) By simply using the most popular combination of attributes as prediction, the POP method can achieve reasonably good performance especially in terms of wRecall and Hamming Loss. This is due to the fact that the attribute distribution is extremely skewed with the most popular combination (female, adult, married, medium income, high school) takes up to 5.6% of users. Not surprisingly, POP is the worst in terms of wPrecision and wF1, since it cannot predict other combinations of attributes. (3) Using SVD to obtain low dimensional representations of users can obtain better performance than POP. For example, the relative improvement of SVD-Single over POP is 10.2%, and by SVD-Structured over POP is 18.5%, in terms of wPrecision when observed label ratio is 50%. (4) The structured learning methods can improve the performances over the single models, by considering the correlation among multiple prediction tasks. For example, the relative improvement of SVD-Structured over SVD-Single is about 10.1% in terms of wF1 when the observed label ratio is 50%, and that of SNE over JNE is about 14.0%. (5) By learning representations towards the end task, we can achieve better performances than method-

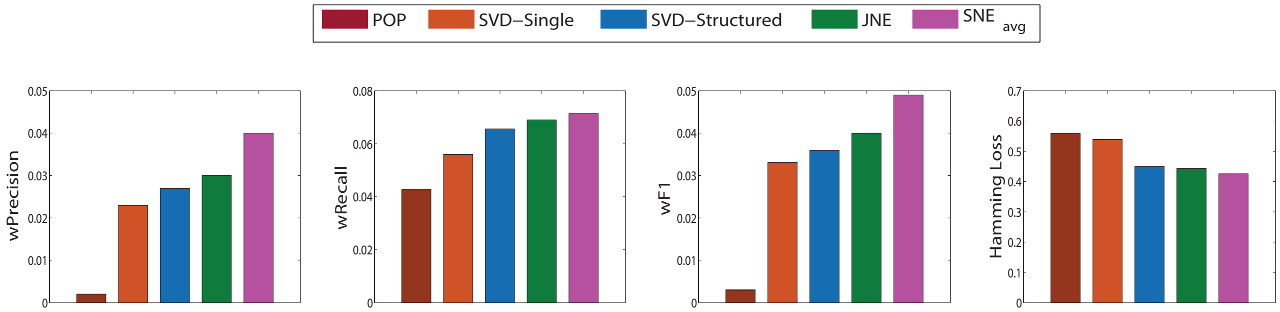

Figure 7: Performance comparison over different methods on New-User prediction in terms of wPrecision, wRecall, wF1 and Hamming Loss.

Table 4: Performance comparison on Partial-Label prediction over different user groups.

user activeness	method	wPrecision	wRecall	wF1	Hamming Loss
Inactive	POP	0.083	0.238	0.122	0.467
	SVD-Single	0.182	0.234	0.211	0.476
	SVD-Structured	0.259	0.242	0.251	0.466
	JNE	0.304	0.269	0.286	0.452
	SNE$_{avg}$	**0.350**	**0.281**	**0.312**	**0.431**
medium	POP	0.093	0.259	0.138	0.445
	SVD-Single	0.200	0.278	0.232	0.426
	SVD-Structured	0.324	0.285	0.304	0.422
	JNE	0.334	0.289	0.310	0.414
	SNE$_{avg}$	**0.371**	**0.289**	**0.324**	**0.411**
active	POP	0.102	0.271	0.148	0.438
	SVD-Single	0.189	0.289	0.229	0.412
	SVD-Structured	0.327	0.286	0.305	0.417
	JNE	0.339	0.297	0.318	0.411
	SNE$_{avg}$	**0.361**	**0.299**	**0.327**	**0.410**

s based on representations learned in an unsupervised way (i.e., SVD). For example, the relative improvement of JNE over SVD-Single is 11.1% in terms of wF1 when the observed label ratio is 50%. (6) Finally, by learning the representations to predict multiple tasks in a structured way, our SNE can achieve the best performance in terms of all the evaluation measures under different observed label ratios. The improvement of SNE over the second best method (JNE) is significant (p-value<0.01) in terms of all the evaluation metrics.

4.3.2 New-User Prediction

We further compare the performance of SNE against baseline methods on predicting new users' demographic attributes. Results of different methods on new user prediction are shown in Figure 7.

From the results we can obtain similar conclusions as in Partial-Label prediction. Both SVD decomposition and structured learning can improve the performance, while supervised representation learning can work better than unsupervised one. The proposed SNE model can achieve the best performance in New-User prediction, and the improvement of SNE over the second best method (JNE) is significant (p-value<0.01) in terms of all the evaluation metrics.

4.3.3 Performance on Different User Group

To further investigate the performance of different methods, we split the users into three groups (i.e., inactive, medium and active) based on their activeness. A user is taken as inactive if there are less than 100 items in his/her purchase history, and active if there are more than 500 items in the purchase history. The remaining users are taken as medium. In this way, the proportions of inactive, medium and active are 62.6%, 31.3%, and 6.1% respectively. The results of Partial-Label prediction when observed ratio is 50% are shown in Table 4. As we can see, the relative performance improvement of SNE over JNE is about 2.6%, 1.4%, 0.9% in terms of wF1 on inactive, medium and active users respectively. In other words, the performance gain of SNE is larger on inactive users than medium and active users. The results indicate that structured prediction can work better by leveraging the correlation between tasks to compensate the limited input information, as compared with joint prediction.

5. CONCLUSION

In this paper,we address the problem of demographic prediction based on users' purchase behaviors. We propose a novel SNE model which can automatically learn the representations to predict a set of demographic attributes simultaneously. Experiments on the real-world purchase dataset demonstrate that our model can outperform the state-of-the-art baselines consistently under different evaluation metrics.

Although the SNE model is proposed in this retail scenario, it is in fact a general model which can be applied on other multi-task multi-class problems. In the future, we would like to extend the usage of our SNE model to other applications to verify its effectiveness. Moreover, the proposed SNE model is still a *shallow* model. Therefore, it would also be interesting to try some deeper architectures to extract more expressive representations for demographic prediction.

6. ACKNOWLEDGE

This research work was funded by 973 Program of China award number under Grant 2014CB340401, 863 Program of China award number under Grant 2014AA015204, 973 Program of China award number under Grant 2012CB316303, National Natural Science Foundation of China award numbers under Grant 61472401, 61433014, 61425016, 61203298, 61425016, and Key Research Program of the Chinese Academy of Sciences under Grant NO.KGZD-EW-T03-2, and the Youth Innovation Promotion Association CAS, Grant no.20144310.

7. REFERENCES

[1] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *J. Mach. Learn. Res.*, 1:113–141, Sept. 2001.

[2] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 131–140, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[3] A. Culotta, N. R. Kumar, and J. Cutler. Predicting the demographics of twitter users from website traffic data. In *Twenty-ninth National Conference on Artificial Intelligence (AAAI)*, 2015.

[4] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 15–24, New York, NY, USA, 2014. ACM.

[5] S. Duarte Torres and I. Weber. What and how children search on the web. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 393–402, New York, NY, USA, 2011. ACM.

[6] P. Eckert. Gender and sociolinguistic variation. *Readings in Language and Gender*, 1997.

[7] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 109–117, New York, NY, USA, 2004. ACM.

[8] A. Graves, A. Mohamed, and G. E. Hinton. Speech recognition with deep recurrent neural networks. *CoRR*, abs/1303.5778, 2013.

[9] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 151–160, New York, NY, USA, 2007. ACM.

[10] Y. Ji and S. Sun. Multitask multiclass support vector machines: Model and experiments. *Pattern Recogn.*, 46(3):914–924, Mar. 2013.

[11] K. Kalyanam and D. S. Putler. Incorporating demographic variables in brand choice models: An indivisible alternatives framework. *Marketing Science*, 16(2):166–181, May 1997.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[13] D. S. M. Kosinski and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 2013.

[14] C. Micchelli and M. Pontil. Kernels for multi-task learning. *NIPS*, 2005.

[15] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: Inferring user profiles in online social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 251–260, New York, NY, USA, 2010. ACM.

[16] A. Mnih and G. Hinton. Three new graphical models for statistical language modelling. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 641–648, New York, NY, USA, 2007. ACM.

[17] D. Murray and K. Durrell. Inferring demographic attributes of anonymus internet users. In *Revised Papers from the International Workshop on Web Usage Analysis and User Profiling*, WEBKDD '99, pages 7–20, London, UK, UK, 2000. Springer-Verlag.

[18] S. Nowozin and C. H. Lampert. Structured learning and prediction in computer vision. *Found. Trends. Comput. Graph. Vis.*, 6(3&#8211;4):185–365, Mar. 2011.

[19] J. Otterbacher. Inferring gender of movie reviewers: Exploiting writing style, content and metadata. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 369–378, New York, NY, USA, 2010. ACM.

[20] T. M. Quoc V. Le. distributed representations of sentences and documents. *The 31st International Conference on Machine Learning*, 2014.

[21] I. S. C. Rick L. Andrews. Identifying segments with identical choice behaviors across product categories: An intercategory logit mixture model. *International Journal of Research in Marketing*, 2002.

[22] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *AAAI*, 2006.

[23] S. Sedhain, S. Sanner, D. Braziunas, L. Xie, and J. Christensen. Social collaborative filtering for cold-start recommendations. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 345–348, New York, NY, USA, 2014. ACM.

[24] K. C. G. C. J. D. Tomas Mikolov, Ilya Sutskever. Distributed representations of words and phrases and their compositionality. *Conference on Neural Information Processing Systems 2013. Proceedings*, pages 3111–3119, 2013.

[25] Y. Yi Liu, Zheng. One-against-all multi-class svm classification using reliability measures. *IEEE International Joint Conference on Neural Networks*, 2005.

[26] M.-L. Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 999–1008, New York, NY, USA, 2010. ACM.

[27] X. W. Zhao, Y. Guo, Y. He, H. Jiang, Y. Wu, and X. Li. We know what you want to buy: A demographic-based system for product recommendation on microblogs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1935–1944, New York, NY, USA, 2014. ACM.

[28] E. Zhong, B. Tan, K. Mo, and Q. Yang. User demographics prediction based on mobile data. *Pervasive Mob. Comput.*, 9(6):823–837, Dec. 2013.

[29] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie. You are where you go: Inferring demographic attributes from location check-ins. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 295–304, New York, NY, USA, 2015. ACM.

[30] Y. J. Zhou J, Chen J. Clustered multi-task learning via alternating structure optimization. *Advances in neural information processing systems.*, 2011.