Multiple Demographic Attributes Prediction in Mobile and Sensor Devices

Yiwen Jiang^{1,2,3}, Wei Tang^{1,2,3}, Neng Gao^{1,2}, Ji Xiang², Chenyang Tu^{1,2(✉)},
and Min Li^{1,2}

¹ State Key Laboratory of Information Security, Chinese Academy
of Sciences, Beijing, China

² Institute of Information Engineering, Chinese Academy
of Sciences, Beijing, China

{jiangyiwen,tangwei,gaoneng,xiangji,tuchenyang,minli}@iie.ac.cn

³ School of Cyber Security, University of Chinese Academy
of Sciences, Beijing, China

Abstract. Users' real demographic attributes is impressively useful for intelligent marketing, automatic advertising and human-computer interaction. Traditional method on attribute prediction make great effort on the study of social network data, but ignore massive volumes of disparate, dynamic, and temporal data derived from ubiquitous mobile and sensor devices. For example, daily walking step counts produced by pedometer. Multiple demographic prediction on temporal data have two problems. First one is that differential effectiveness of different time period data for prediction is unclear. And another one is how to effectively learn the complementary correlations between different attributes. To address the above problem, we propose a novel model named Correlation-Aware Neural Embedding with Attention (CANEAN), which first directly separates different attribute oriented feature using separated embedding layer, and use attention mechanism to assign a higher weight to dominant time point. Then it captures informative correlations using correlation learning layer. Finally we obtain the refined task-specific representations with optimal correlation information for predicting certain attributes. Experimental results show the effectiveness of our method.

Keywords: Demographic prediction · Pedometer data · Correlation learning

1 Introduction

Knowing users' real demographic attributes is extremely significant for the applications of intelligent marketing, automatic advertising and human-computer interaction. Conventional applications of demographic prediction tend to pay attention to users' behavior in social networks, and few people pay attention to mobile and sensor devices. However, sensor data for demographic prediction has it's own unique applications. For example, it's hard to use online speech to infer the age and gender of a person who rarely speaks on social networks, but it's

easier to infer such attribute via his or her pedometer records, as the walking is a daily stable behavior. Besides, the attributes predicted through sensor data could be directly used to enable human-computer interaction more humanized and friendly in smart home system. For example, when a young lady go home from outside, the air conditioning and audio could select a gender/age-aware response (air conditioning temperature and music) from many possible candidates to make the user more comfortable [10].

Most of earlier studies on attribute prediction are primarily involve analysis of the user-generated public data on the network, especially logs, comments as well as videos. They mostly come from Facebook [12], Twitters [2, 13], microblogs [17], YouTube [4], web search queries [6], social networking chats [11], and forum posts [3]. And little work is put into the management of sensor data. Besides, traditional demographic prediction model can not fully learn the complementary correlations between different attributes. Existing work usually infer users' different attributes separately without taking account of correlation learning or infer all the attributes together using shared embedding at the bottom of model. For example, wang et al. [16] proposed a Structured Neural Embedding (SNE) model that use shared embedding to leverages the potential correlations for multi-task learning by concatenating structured label. However, the concatenated formalization makes the prediction task more difficult than separated tasks since the output space for each task becomes much larger. Raehyun et al. [8] proposed an Embedding Transformation Network (ETN) model that shares user representations at the bottom and converts them to task-specific representations using a linear transformation. But this model does not take any steps to explicitly learn the complementary relationship between different task-specific features.

To tackle the above problem, we make effort on the reasonable utilize of pedometer data and go a further step to present a new model named Correlation-Aware Neural Embedding with Attention (CANEA) for multiple demographic attributes prediction. In CANEA, we first leverage a separated embedding layer with attention based on RNN branches to extract task-specific features, and obtain explicit day weight for better prediction. Then, to well learn complementary correlations between different tasks, we design a correlation learning layer with full connection neural pipelines to capture correlation features. Moreover, we enhance the user representation with correlation features using bilinear model. The final user representations are fed into multi-task prediction layer for multiple demographic attributes prediction.

We conduct extensive experiments on a real world pedometer dataset to test our method. Several state-of-the-art baselines are taken into comparison. And we also test our model in traditional application scenario of demographic prediction on transaction data. The experimental results prove the effectiveness of our model. Overall, our contributions are as follows:

- (1) We extend our sight to the ubiquitous mobile and sensor devices to bridge the gap between pedometer data and users' demographic attributes.
- (2) A new model named Embedding Separation and Correlation Learning with Attention (CANEA) is proposed for multiple demographic prediction, which has strong ability to learn the complementary correlations between tasks.

In addition, we explicitly analyze different days' effect of pedometer records on demographic prediction.

- (3) Extensive experiments are conducted on a real world pedometer dataset as well as a public transaction dataset, which all demonstrate the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section 2 gives the problem formalization of multi-task demographic prediction. Section 3 discusses our model Architecture in detail, and Sect. 4 presents the experimental results and analysis. And finally, in Sect. 5, we conclude our work and present future work.

2 Problem Formalization

In our application scenario, we aim to predict demographic attributes based on pedometer data. Specifically, the profile of each user is represented as a sequence of daily step counts, which could be presented as follow:

$$S : [[s_1], [s_2], ..., [s_n]]$$

where S denotes the pedometer record of a certain user, s_i is his/her step count on the i -th day, and n is the total days.

Typically, our work is a multi-task prediction problem. Here, the attributes we are interested in are gender and age, and the attributes can be various if necessary in other applications. The values of gender are male and female regardless of non-traditional gender, and the values of age include young, adult, middle age and old [16]. Therefore, the prediction problem $\mathcal{T}$ could be summarized as:

$$\mathcal{T} : S \rightarrow \{y_1, y_2, ...y_t\}$$

where y_i is the attribute class of i -th task, and t is the total number of the tasks. The objective of our work is to train an effective end-to-end model to simultaneously predict demographic attributes for new users.

Fig. 1. The architecture of CANEA.

3 Model Architecture

In this section, we present the details of the CANEA. An overview of our framework is illustrated in Fig. 1. We leverage separated embedding layer to extract task-specific representations. And a correlation learning layer is used to learn the hidden correlations between different tasks. Finally, the multi-task prediction layer make the prediction of all tasks simultaneously.

3.1 Separated Embedding Layer

We use separated embedding branches showed in Fig. 2 to learn task-specific representations. Different from mapping user profiles to a shared representation at the bottom of model that ignores the interferences among multiple tasks, the separated embedding branches eliminate these interferences and produce a relatively pure task-specific representations directly. We have to admit that shared embedding retain informative correlations among multiple attributes, and separated embedding method seems to ignore these correlations. But this shared embedding lacks explanations for correlation extraction. And in the next section, we will introduce correlation learning layer to learn these correlations in a more interpretable way.

Fig. 2. The architecture of separated embedding branches.

Considering that the pedometer data is a kind of temporal data, we adopt the most popular sequence learning model of LSTM [5] as the backbone of each embedding branches. In this paper, we assume that the contexts from both past and future are useful and complementary to each other. Therefore we combine forward (left to right) and backward (right to left) recurrent to build a bidirectional LSTM (Bi-LSTM) [14]. Moreover, the stacked recurrent layers are used to build a deep RNN model to enhance the representation ability of the separated embedding layer. Here, we use a pairs of stacked bidirectional LSTM to respectively learn age oriented and gender oriented representations.

Note that the different days in each week play differential roles in tasks. For example, in our intuition, female are more likely to go shopping than male on weekends. That means users’ step counts on weekend is more important than weekday for gender inference. From this point of view, we adopt attention mechanism [1] to give the significant days higher scores. Thus, the separated embedding branches could be presented as:

$$\mathbf{w}_i = softmax(tanh(\mathbf{v}_i\mathbf{x} + \mathbf{b}_i))$$
$$\mathbf{E}_i = \mathcal{L}_{\theta_i}(\mathbf{w}_i\mathbf{x})$$

where $\mathcal{L}$ is the trainable model of Bi-LSTM, θ_i is the trainable parameters of embedding branches for i -th task, $\mathbf{x}$ is the input sequence, and $\mathbf{E}_i$ is the output task-specific feature for i -th task. Additionally, $\mathbf{v}_i$ and $\mathbf{b}_i$ are trainable parameters of the attention model in i -th branch, $\mathbf{w}_i$ is the attention weights describing the importance assigned to each element of the input.

3.2 Correlation Learning Layer

As we mentioned before, the separated embedding layer eliminate the interferences among multiple tasks, but the informative correlations between different tasks have been ignored. In this situation, we design a correlation learning layer to learn hidden correlation features. The key components of this correlation learning pipelines are a full connection network and a bilinear mixer (presented in Fig. 3).

Fig. 3. Architecture of the key component in correlation learning pipelines.

Full connection network are widely used in hidden representation learning on account of their excellent learning ability and desirable scalability. The correlation representation for i -th task is presented as:

$$\mathbf{C}_i = \mathcal{F}_{\phi_i}(\mathbf{w}_i \bar{\mathbf{E}}_i)$$

where $\mathcal{F}$ is the trainable model of full connection network, ϕ_i is the trainable parameters, C_i represents correlation feature for i -th task, $\bar{\mathbf{E}}_i$ is the concatenation of other task-specific features.

We combine the separation feature and correlation feature in a bilinear model. The bilinear model is a two-factor model with mathematical property of separability: their outputs are linear in either factor when the others held constant, which has been demonstrated that the influences of two factors can be efficiently separated and combined in a flexible representation [15]. The combination function can be formulated as:

$$\mathbf{M}_i = \mathbf{E}_i \mathbf{W}_i \mathbf{C}_i$$

where $\mathbf{W}_i$ is the tensor of trainable parameters, $\mathbf{E}_i$ denotes separation representation of i -th task, and $\mathbf{C}_i$ is the correlation representation for i -th task.

3.3 Multi-task Prediction Layer

With the separation representation and correlation representation obtained by the previous two layers, we obtain the prediction probability for the demographic attribute of a given user by:

$$p(y_i|S) = \text{softmax}(\mathbf{O}_i \mathbf{M}_i)$$

where $\mathbf{M}_i$ is the mixed representation of user and $\mathbf{O}_i$ is the trainable parameter. The parameter is responsible for converting the mixed representation for each task into prediction through linear transformation.

The goal of demographic prediction is to infer all demographic attribute of users from their pedometer profiles. For i -th task, we minimize the sum of the negative log-likelihoods defined as:

$$Loss_i = - \sum_{j=1}^m \log p(y_{i,j}|S_j)$$

where m is the total number of users. S_j and $y_{i,j}$ are the input of j -th user's daily step count sequence and his/her inferred attribute class of i -th task.

Combining all these task-specific losses, the total multi-task loss function is:

$$Loss = \sum_{i=1}^t \lambda_i Loss_i$$

where the hyper-parameter λ controls the trade-off between all of t task-specific losses. Considering that all tasks are equal important in our experiments, we set all λ to be 1.

4 Experiments

In this section, we present the details of our experiments and analyze the effectiveness of our approach.

4.1 Dataset

We build a large-scale real world pedometer dataset came from WeChat, a famous mobile application in China. WeChat develops a subfunction called WeChat Sport that collects and ranks users’ as well as their net friends’ daily step counts online based on the pedometers embedded in mobile phones, and we crawl the data at the chance. Our dataset contains 39,246 users’ 300-days step counts during the period from 2018.6.11 to 2019.4.6. All of the users are annotated with gender and age. To guarantee the reliability of the data, we removed those unsuitable users who have more than 150 days of zero step count records. The distribution of users’ attributes are listed in Table 1. In experiment, we use 60% and 20% of the data for training and test, and leave the rest 20% as a validation set for tuning.

Table 1. Distribution of users’ attributes.

Attributes	Value	Users	Distribution
Gender	Male	22134	56%
	Female	17112	44%
Age	Young	8635	22%
	Adult	19230	49%
	Middle age	7064	18%
	Old	4317	11%

4.2 Evaluation Metrics

We employ F-measure to evaluate our model. F1-measure is a widely used measure method as a complement for accuracy, and it is the most popular evaluation metrics for demographic prediction. Specifically, F1 score is calculated as the harmonic mean of precision and recall. Note that, in our experiment, we use the weighted precision (wP), recall (wR) and F1 ($wF1$) score as the evaluation metrics since we consider all classes to be equal important. The weighted F1 is calculated as follows:

$$wP = \sum_{y \in Y} (\frac{\sum_{i=1}^u I(y_i^* = \hat{y}_i \& y = \hat{y}_i)}{\sum_{i=1}^u I(y = \hat{y}_i)} * weight)$$
$$wR = \sum_{y \in Y} (\frac{\sum_{i=1}^u I(y_i^* = \hat{y}_i \& y = \hat{y}_i)}{\sum_{i=1}^u I(y = y_i^*)} * weight)$$

$$wF1 = 2 \times \frac{wP \times wR}{wP + wR}$$

where $I(\cdot)$ is an indicator function, u denotes the total number of new users, Y is the set of all label combinations to be predicted, y_i^* denotes the ground truth of attributes for the i -th new user, $\hat{y}_i$ denotes the predicted attributes, and $weight = \frac{1}{u} \sum_{i=1}^u I(y = y_i)$. The weighted F1 assigns a high weight to the large classes to account for label imbalance.

4.3 Baseline Models

We compare our models with several baselines on demographic prediction. The description of these baselines are listed below:

POP. POP is a naive method that always predicts the given sample as the majority classes. In previous work [8, 16], POP is used as a baseline model which ignores characteristics of users.

JNE. Joint Neural Embedding [16] maps users' all walking histories into latent vectors. These vectors are processed by average pooling and then fed into a linear prediction layer for each task.

SNE. Structured Neural Embedding [16] has similar structure with JNE. The only difference between SNE and JNE is that the loss of SNE is computed via a log-bilinear model with structured predictions.

ETN. Embedding Transformation Network [8] uses a shared embedding just as SNE. The shared embedding is fed into an embedding transformation layer to obtain the transformed representation. Then the transformed representation is directly fed into the prediction layer.

ETNA. Embedding Transformation Network with Attention [8] is an improved version of ETN. The transformed representation produced by embedding transformed layer is fed to a task-specific attention layer to take into account the importance of each element in users' profile.

4.4 Experimental Settings

Considering that the recurrent units maintain activation for each time-step which have already make the network to be extremely deep, we use only one layer of bidirectional RNN with 128 LSTM units in each separated embedding branch. And the full connection networks in correlation learning layer is set to be the shallow architectures using only 3 hidden layer with 256 sigmoid units. We use random values drawn from the Gaussian distribution with 0 mean and 0.01 standard deviation to initialize the weight matrices in LSTM, full connected layer, and prediction layer. Learning from [7], the forget gate bias are initialized to be 5 to let the forget gate close to 1, namely no forgetting. Thus, long-range

Table 2. Performance comparison of different models.

Model name	Results		
	wP	wR	wF1
POP [8]	0.137	0.354	0.198
JNE [16]	0.609	0.643	0.626
SNE [16]	0.615	0.648	0.631
ETN [8]	0.634	0.679	0.656
ETNA [8]	0.641	0.688	0.660
CANEA ^a	0.496	0.513	0.504
CANEA ^b	0.681	0.726	0.703
CANEA	0.695	0.741	0.717

^aabandon correlation learning layer.^babandon attention mechanism.

dependencies can be better learned at the beginning of training. All other bias, the cell as well as hidden states of LSTMs in our work are initialized at 0. Adam [9] is used as the optimization algorithm and the mini-batch size is 128. The learning rate is set to be $1e^{-5}$. After each epoch, we shuffle the training data to make different mini-batches.

4.5 Comparison with Baselines

Table 2 shows the experimental results of our model and the baseline models. Based on these results, we have the following findings:

- (1) If we abandon correlation learning layer, our model will degenerate into a parallel integration of multiple RNNs, CANEA^a. Note that CANEA is much better than CANEA^a, this result proves that the correlation learning layer plays an important role in our models. Furthermore, CANEA^a is far behind JNE, SNE, ETN and ETNA, because it ignores the informative correlations among multiple tasks. According to this result, it could be demonstrated that capturing such correlations has great significance in multi-task prediction.
- (2) As we emphasized in this paper, the ability to learn correlations between different tasks is important in multi-task learning. JNE, SNE, ETN and ETNA use the shared embedding that implicitly leverage these correlations, but ignore the interferences among multiple tasks. We first use the separated embedding to avoid such interference, and then employ the correlation learning to obtain the correlations. Although we use the most simple architectures of our models for baseline comparison (see Sect. 4.4), which still outperform all the baselines with a significant gap.
- (3) The attention mechanism is helpful for separation embedding. As shown in Table 2, the CANEA is obviously better than the CANEA^b. As aforementioned, they have the similar architecture, except that the former model introduces the attention mechanism into separated Embedding layer to

weaken unimportant factors and highlight the important factors of the input. The results are a strong proof that the step count records of different days in week play differential roles in gender and age prediction.

4.6 Visualization of Attention

To further analyze the effectiveness of attention mechanism in separated embedding layer, we provide visualization of attention scores. We pick example that provide insights for users' pedometer records from Sunday to Saturday during 10 weeks with the average attention scores in each task. Based on the attention scores from our model, we draw heatmap in Fig. 4.

Fig. 4. Comparison of attention weights calculated by separated embedding layer.

Noting that Saturday and Sunday obtain highest attention in gender prediction task. Exactly, in our intuition, women prefer to go shopping on the weekends, which fits the gap between male and female in Fig. 5-(Gender).

Fig. 5. Mean value (*y*-axis) of attention scores from Sunday to Saturday (*x*-axis).

For age prediction, our model give relatively higher attention scores to weekdays but lower scores to weekends. Intuitively, young people are more energetic and active, and most adults tend to stay at their desks on weekdays. But on the weekends, parents and children may go out to play together. Figure 5-(Age) empirically demonstrates the reasonability of this intuition from a statistical point of view, as the gap between adult and young is relatively large on weekdays but small on the weekends.

4.7 Experiments on Transaction Data

We also conduct experiments on transaction dataset used in [8]. This dataset is the first public dataset containing both transaction records and demographic information. It consists of purchasing histories of 56,028 users and contains the gender, age, and marital status of all the users. And this dataset support two types of different prediction applications: partial label prediction and new user prediction. Table 3 reports the experimental results. Results show that CANEA also outperforms all baseline models with impressive improvement on transaction data, which also proves the strong generalization ability of our model.

Table 3. Results on transaction data.

Model name	Partial label			New user		
	wP	wR	wF1	wP	wR	wF1
POP	0.289	0.514	0.370	0.086	0.294	0.134
JNE	0.518	0.563	0.539	0.321	0.348	0.334
SNE	0.521	0.563	0.542	0.295	0.351	0.321
ETN	0.540	0.576	0.557	0.310	0.368	0.336
ETNA	0.554	0.584	0.569	0.339	0.382	0.360
CANEA	0.568	0.596	0.582	0.356	0.401	0.377

5 Conclusion and Future Work

In this paper, we introduce the Correlation-Aware Neural Embedding with Attention (CANEA) model for demographic prediction. Different from previous method using shared embedding to retain hidden correlations between multiple tasks, our approach leverages separated embedding to eliminate the interferences between multiple tasks and then propose correlation learning to obtain the correlations between multiple tasks. Furthermore, we further analyze different effects on gender and age prediction of different days in week, and take advantage of this phenomenon to optimize our model.

In the future, we would like to verify the effectiveness of our approach in a larger number of tasks, which need more annotations. More advanced models are expected as some shortcomings are still existed in current methods, such as lack of external knowledge and excessive reliance on large amounts of training data.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015 (2015). <http://arxiv.org/abs/1409.0473>

2. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Conference on Empirical Methods in Natural Language Processing, pp. 1301–1309 (2011)
3. Dong, N., Smith, N.A.: Author age prediction from text using linear regression. In: ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 115–123 (2011)
4. Filippova, K.: User demographics and language in an implicit social network. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1478–1488 (2012)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>
6. Jones, R., Kumar, R., Pang, B., Tomkins, A.: “I know what you did last summer”: query logs and user privacy. In: Sixteenth ACM Conference on Conference on Information and Knowledge Management, pp. 909–914 (2007)
7. Józefowicz, R., Zaremba, W., Sutskever, I.: An empirical exploration of recurrent network architectures. In: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015. pp. 2342–2350 (2015). <http://jmlr.org/proceedings/papers/v37/jozefowicz15.html>
8. Kim, R., Kim, H., Lee, J., Kang, J.: Predicting multiple demographic attributes with task specific embedding transformation and attention network. CoRR abs/1903.10144 (2019). <http://arxiv.org/abs/1903.10144>
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR abs/1412.6980 (2014). <http://arxiv.org/abs/1412.6980>
10. Li, S., Wang, J., Zhou, G., Shi, H.: Interactive gender inference with integer linear programming. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, 25–31 July 2015. pp. 2341–2347 (2015). <http://ijcai.org/Abstract/15/331>
11. Peersman, C., Daelemans, W., Vaerenbergh, L.V.: Predicting age and gender in online social networks. In: International Workshop on Search and Mining User-Generated Contents, pp. 37–44 (2011)
12. Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., Crowcroft, J.: The personality of popular facebook users. In: ACM Conference on Computer Supported Cooperative Work, pp. 955–964 (2012)
13. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: International Workshop on Search and Mining User-Generated Contents, pp. 37–44 (2010)
14. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**(11), 2673–2681 (1997). <https://doi.org/10.1109/78.650093>
15. Tenenbaum, J.B., Freeman, W.T.: Separating style and content. In: Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, 2–5 December 1996. pp. 662–668 (1996). <http://papers.nips.cc/paper/1290-separating-style-and-content>
16. Wang, P., Guo, J., Lan, Y., Xu, J., Cheng, X.: Your cart tells you: inferring demographic attributes from purchase data. In: WSDM, pp. 173–182 (2016)
17. Zhang, L., Huang, X., Liu, T., Li, A., Chen, Z., Zhu, T.: Using linguistic features to estimate suicide probability of Chinese microblog users. In: Zu, Q., Hu, B., Gu, N., Seng, S. (eds.) HCC 2014. LNCS, vol. 8944, pp. 549–559. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-15554-8_45