Demographic Prediction from Purchase Data based on Knowledge-Aware Embedding

Yiwen Jiang[1,2,3], Wei Tang[1,2,3], Neng Gao[3], Ji Xiang[3], and Daren Zha[3]

[1] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[2] State Key Laboratory of Information Security, Chinese Academy of Sciences, Beijing, China
[3] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{jiangyiwen,tangwei,gaoneng,,xiangji,zhadaren}@iie.ac.cn

Abstract. Demographic attributes are crucial for characterizing different types of users in developing market strategy. However, in retail scenario, individual demographic information is not often available due to the difficult manual collection process. Several studies focus on inferring users' demographic attribute based on their transaction histories, but there is a common problem. Hardly work has introduced knowledge for purchase data embedding. Specifically, purchase data is informative, full of related knowledge entities and common sense. However, existing methods are unaware of such external knowledge and latent knowledge-level connections among items. To address the above problem, we propose a Knowledge-Aware Embedding (KAE) method that incorporates knowledge graph representation into demographic prediction. The KAE is a multi-channel and item-entity-aligned knowledge-aware convolutional neural network that fuses frequency-level and knowledge-level representations of purchase data. Through extensive experiments on a real world dataset, we demonstrate that KAE achieves substantial gains on state-of-the-art demographic prediction models.

Keywords: Demographic Prediction · Convolutional Neural Networks · Knowledge Graph Representation.

1 Introduction

Knowing customers' demographic attributes is significant for many companies and retailers to make market basket analysis [1], adjust marketing strategy [6], and provide personalized recommendations [16, 24]. For example, Nikes basketball shoes targets mainly a relatively young (age) and male (gender) customers with enough purchasing power (income). Additionally, in recommender systems, demographic information have been wildly used to improve the quality of the systems and solve the cold start problem.

Generally, the collection of customers' demographic information is difficult for most companies and retailers, as customers are reluctant to offer their personal information to companies in case of a series of data breaches. Besides, in a real

I Know You by Your Walking: Demographic Prediction with Separated Embedding and Correlation Learning

Yiwen Jiang [1,2,3], Wei Tang [1,2,3], Neng Gao [2], Ji Xiang [2]
[1] State Key Laboratory of Information Security, Chinese Academy of Sciences
[2] Institute of Information Engineering, Chinese Academy of Sciences
[3] School of Cyber Security, University of Chinese Academy of Sciences
{jiangyiwen,tangwei,gaoneng,xiangji}@iie.ac.cn

ABSTRACT

Knowing exact demographic attributes of users is crucial for human-computer interaction, intelligent marketing and automatic advertising. Ubiquitous sensor devices yield massive volumes of temporal data which hide a lot of valuable demographic information. In this paper, we bridge the gap between sensor data and demographic prediction to obtain real attributes of users from popular sensor devices: pedometer, which is widely used in mobile devices. We propose a novel model named Separated Embedding and Correlation Learning (SECL) for demographic prediction. Specifically, SECL first process the input data with a separated embedding layer to disentangle task-specific features for interference eliminating, and then capture the optimal correlations via a correlation learning layer, finally the refined task-specific features are fed into a multi-task prediction layer to predict demographic attributes. Experimental results show impressive performance of our model on a real-world pedometer dataset, which is made publicly available on https://github.com/deepdeed/SECL.

KEYWORDS

demographic prediction, sensor data, sequence learning

ACM Reference Format:
Yiwen Jiang [1,2,3], Wei Tang [1,2,3], Neng Gao [2], Ji Xiang [2]. 2019. I Know You by Your Walking: Demographic Prediction with Separated Embedding and Correlation Learning. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recently, sensing devices are ubiquitous in people's daily life. For example, many mobile devices like mobile phone embed pedometer, gyroscope, accelerometer, vibrometer and magnetometer. Some popular wearable devices such as Fitbit, Apple Watch, and Android Wear use pedometer, accelerometer and heart rate monitor [3]. All these sensing devices generate trillions of sensor data points per year, including rich signals such as step count variability, which closely correlate with users' daily activities as diverse as walking, exercise, or trip and indirectly hide the user's demographic attributes characteristics. As a result, extracting knowledge and emerging patterns from sensor data for user attribute prediction is a nontrivial task.

Obtaining individual attributes is crucial for the applications of human-computer interaction, intelligent marketing and automatic advertising. Beyond conventional applications of user attribute inference, knowing demographic attributes of users via sensor data has it's own unique applications in internet of things. For example, in smart home system, explicit attribute could be used to enable human-computer interaction more humanized and friendly. More specifically, when responding to a human with known gender, the computer could select a gender-aware response from many possible candidates to make the user more comfortable, which significantly enhance the competitiveness of the products [19]. However, it is usually not easy for smart device to obtain exact users' attributes.

In this paper, we make effort on the reasonable utilize of pedometer data (step count sequence) for demographic prediction. Most of earlier studies on attribute prediction are primarily involve analysis of the user-generated data derived from social media, including Facebook [25], Twitters [5, 26], microblogs [33], telephone conversations [12], YouTube [11], web search queries [15], social networking chats [24], and forum posts [9]. In this paper, we extend our sight to the ubiquitous mobile and sensing device to bridge the gap between sensor data and users' demographic attributes. We attempt to extract knowledge and emerge users' daily walking patterns from pedometer data, thereby inferring users' demographic. To the best of our knowledge, there is only one existing work that has used sensor data for prediction task in 2018. Ballinger et al. [2] combined step count with heart rate and proposed a semi-supervised learning method to predict cardiovascular risk in medical field. Nevertheless, the heart rate data is hard to obtain and full of privacy sensitivity on personal health. In this case, we use only step count data but make the finer granularity of analysis for a more general problem of demographic prediction.

Previous work on demographic prediction, for example, Structured Neural Embedding (SNE) [31](Fig 1-(a)), usually employ shared embedding to capture the shared feature of user attribute. The advantages of this model are relatively simple structure and less parameters, but it ignores the interferences of multiple tasks. Another method of Embedding Transformation Network (ETN) [17] (Fig 1-(b)) address this problem using separated transform embedding upon the shared embedding to extract task-specific features. But ETN also have a significant limitation: insufficient ability to learn informative correlation features between multi-tasks. Commonly in multi-task learning, optimal correlation features are helpful for

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Your Pedometer Tells You: Attribute Inference via Daily Walking Step Count

Yiwen Jiang[*†], Wei Tang [*†], Neng Gao[*], Ji Xiang[*], Daren Zha[*], Xiang Li [*†]
[†]*School of Cyber Security, University of Chinese Academy of Sciences*
[*]*Institute of Information Engineering, Chinese Academy of Sciences*
Beijing, China
{jiangyiwen, tangwei, gaoneng, xiangji, zhadaren, lixiang9015} @iie.ac.cn

***Abstract*—Knowing individual real attributes is beneficial to intelligent marking, automatic advertising as well as a new application of smart home system. In this paper, we leverage seemingly innocent user information (step count) coming from ubiquitous mobile and sensing device for attribute inference. The utilization of sensor data breaks the traditional dependence on text, social relationship and online behavior in social media, and avoids the problem of users' gender variation in linguistic style on different social networks.**

Existing methods on attribute inference usually ignore the temporal information of data and purely explore different types of features manually. However, loss of temporal characteristic may reduce the representation ability to infer user's attributes. Meanwhile, excessively depending on manually defined features requires a great mass of human labor and often suffers from under specification. To address these problems, we propose a novel Hybrid Multiple Representations (HMR) model that combines simple human knowledge with automated deep learning by using stacked bidirectional LSTM, Bag-of-step and Holiday Activation methods to predict gender and age from users daily step count.

Experiments are conducted on a real-world pedometer dataset where gender and age are to be predicted. The empirical results show that our HMR model does a good job on the task of predicting attributes compared with state-of-the-art baselines.

***Index Terms*—Attribute inference, Sensor data, Hybrid Multiple Representations**

I. Introduction

Obtaining individual attributes is crucial for the applications of intelligent marketing and automatic advertising. For example, it could be used to conduct market basket analysis [1], adjust marketing strategy [16], and provide personalized recommendations [28], [33]. However, in practice, it is usually not easy to obtain exact users' attributes. As a result, the task of inferring users' attributes by analyzing the user-generated any trace, like content, behavior and social relationship, has become an active area of research [5], [9], [12], [21], [30].

In social media, it is worthwhile to highlight that the large number of users are increasingly aware and cautious about the publish text, open social relationship as well as online behavior to prevent being tracked [8]. Besides, there are research indicate that people have gender variation in linguistic style on different social platforms [3], which bring about the difference in inferring attributes for same user. Whereas most of earlier studies on attribute inference are primarily involve statistics of the social media data including text, image, and network behaviors, which are mainly derived from Facebook [25], Twitters [5], microblogs [31], telephone conversations [10], YouTube [9], social networking chats [24], scientific papers [4] and forum posts [7]. In this paper, we extend our sight to the ubiquitous mobile and sensing device. Most notably omitted in most methods is one kind of regularity of user's daily life data comes from sensor, like daily walking step count comes from pedometer. This kind of pedometer information potentially reflects people's stable lifestyle and habit, which is difficult to be modified and will not change with diverse social platform. To the best of our knowledge, only one existing method has attempted to use step count in 2018, which is combined with heart rate data for cardiovascular risk prediction [2].

Beyond conventional applications of attribute inference, this sensor data has its own unique applications on internet of things platform, like smart home system. For example, it could be used to enable human-computer interaction more humanized and friendly. More specifically, when responding to a human with known gender, the computer could select a gender-aware response from many possible candidates to make the user more comfortable [20].

Previous work on attribute inference is usually based solely on manually defined features, and ignore the temporality exploitation of data. For example, Liu et al. [21] used first name as features to infer twitter users' gender after long-term systematic investigation on a large corpus. Zhong et al. [18] tried to predict six attributes using spatial, temporal and location knowledge features. However, simply relying on manually defined features usually requires a great deal of professional knowledge and often suffers from under specification. Some studies proposed to automatically extract features from the raw data by employing feature learning method [18], [30]. For example, Wang et al. [30] characterized users' purchase history to form user representation using the bag-of-item, and feed this representation to a log-bilinear model for structured prediction. But the representation ability of this method is limited by the relatively simple structure, particularly for data with temporal characteristics. Recently, Ballinger et al. [2] proposed a semi-supervised Temporal Convolutions LSTM sequence learning method to predict cardiovascular risk without introduction of background knowledge. Nevertheless, the features learned in an semi-supervised manner and lacking of human knowledge