Pattern Recognition 73 (2018) 33–46

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

User attribute discovery with missing labels☆

Yang Cong[a,*], Gan Sun[a], Ji Liu[b], Haibin Yu[a], Jiebo Luo[b]

[a] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
[b] Department of Computer Science, University of Rochester, Rochester, NY 14611, USA

ARTICLE INFO

Article history:
Received 31 December 2016
Revised 20 April 2017
Accepted 7 July 2017
Available online 20 July 2017

Keywords:
User attribute
Smart sensor
Multi-task learning
Semi-supervised learning
Missing labels
Low rank

ABSTRACT

In this paper, we focus on user attribute analysis by recasting such a problem as a multi-task learning issue, where each attribute is considered as an independent task. In comparison with traditional data analysis, the missing labels problem broadly presents for smart sensor data due to some objective / subjective factors, where the label incompleteness increases the difficulty significantly. Therefore, we design a semi-supervised multi-task learning model (S2MTL) to handle the missing labels issue. For modeling, we integrate the matrix factorization to learn the mapping feature dictionary and attribute space information simultaneously, and adopt the pairwise affinity similarity to incorporate the unlabeled data information, where the low rank property and model efficiency can be well controlled. For model optimization, we convert our model as two individual convex subproblems with one non-smooth, and implement an alternating direction method to generate an efficient optimal solution. State-of-the-art models have validated the effectiveness and efficiency of our proposed model via extensive experiments and comparisons, on two public datasets and our new smart building dataset.

1. Introduction

User attributes such as gender, age, financial, employment, marriage status, are crucial for many intelligent business applications, e.g., targeted marketing [1] or social science. Traditionally, the user attribute information is manually collected by survey questionnaire. Automatically discovering the user attribute is amazing but very hard, some researches use the social network such as blog, twitter, micro-blogs [2].

In this paper, we intend to discover multiple user attributes simultaneously depending on the smart sensor data, which can be considered as a multi-task learning issue. For example, we adopt the smart meter data (recording the consumer electricity consumption every 30 min) to estimate the user attributes, e.g., age, children number, single or not, home income, etc; these user attributes information can then help to set the multi-step electricity price, recommend targeted advertising for business, or analyze social behavior.

In our opinion, one discriminative characteristic of the smart sensor data in comparison with traditional data analysis issues, is the missing data phenomenon. For example, there are some missing data caused by various objective factors, such as the data acquisition error (the hardware error when collecting sensor data); the transmission error happened due to electromagnetic interference; or even the harsh environment or cost factor, where we cannot install sufficient sensors or the sensor cannot be mounted in some specific locations. There are also some subjective factors, e.g., human subjectively missed some questions due to personal privacy reasons when collecting the survey questionnaire manually. Generally in our opinion, the missing data problem can be categorized into four types as shown in Fig. 1, where $X \in \mathbb{R}^{d\times n}$ and $Y \in \mathbb{R}^{m\times n}$ are the training data and the corresponding labels, respectively (d, m and n are the feature dimension, task number and data size).

a. Fully labeled data, where both X and Y are fully completed, i.e., there are no missing entry.
b. Missing X, i.e., the input X is partially missing.
c. Partially missing Y, including the whole column of Y is missing and some entries of Y are missing. (Missing all of Y can be considered as an unsupervised learning problem and out of the scope of this study).
d. Missing both X and Y.

In this paper, we focus on the missing labels problem ([c] partially missing Y) for user attribute discovery. In order to address this, each attribute can be separately predicted, which is unrobust without considering the embedded correlation among different attributes. Therefore, we consider each attribute as an independent task and adopt the multi-task learning [3–9] to discover

☆ This work is supported by NSFC (61375014, 61533015, U1613214).
* Corresponding author.
E-mail addresses: congyang81@gmail.com (Y. Cong), sungan@sia.cn (G. Sun), jliu@cs.rochester.edu (J. Liu), yhb@sia.cn (H. Yu), jiebo.luo@cs.rochester.edu (J. Luo).

http://dx.doi.org/10.1016/j.patcog.2017.07.012
0031-3203/

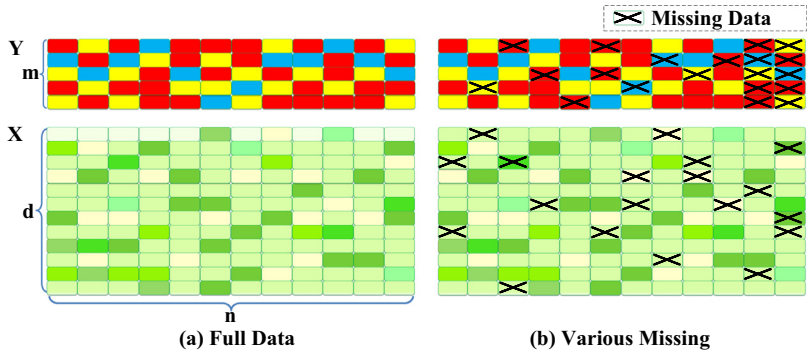

Fig. 1. Multi-task learning with (a) fully labeled data of both X and Y; (b) partially labeled data with various missing entries, such as missing data X, partially missing labels y_i and completely missing labels y_i (two columns on the right of Y).

the user attribute concurrently, which can improve the generalization performance than standard multi-class classification problem. Most state-of-the-art supervised multi-task learning methods only adopt the limited labeled Y for model training, e.g., [10–16]. As we know, the unlabeled data, i.e., the missing labels data, also include useful information. Therefore, some semi-supervised methods are attempted to use the unlabeled data. For example, [17] uses Gaussian process for multi-task learning [4,18]; assume that all tasks should be related with each other, which limit the generality of their methods [19]; designs a multi-label classification algorithm using only one Laplacian matrix to represent all tasks and low-rank structure in the origin label space, which cannot handle multi-task learning well and limit its generalization performance. In order to handle this, we propose a semi-supervised multi-task learning model (S2MTL) to overcome missing labels by integrating mapping feature dictionary and attribute space information to address multi-task matrix recovery. The proposed formulation is non-convex; we convert it as two independent convex optimization problems, and we then develop an efficient alternating direction framework to solve it with a global optimal solution. We adopt our S2MTL model for user attribute analysis depending on smart sensor data with missing labels. We also build a new smart building dataset for multi-task learning, and adopt our model for smart meter dataset as well. Generally, the main contributions of our paper are as follows:

i. Missing labels phenomenon is a frequent problem especially for real smart sensor data analysis. In this paper, we propose a semi-supervised multi-task learning model (S2MTL) with low rank constraint to overcome missing labels issue for user attribute discovery.
ii. For modeling, we adopt the matrix factorization by learning the mapping feature dictionary and attribute space information simultaneously. Therefore, we can handle the high-dimensional big data efficiently in practice, and meanwhile, control the model complexity with low rank as well.
iii. To our best knowledge, ours is the first work about multiple user attributes discovery with missing label data. We also build a new smart building dataset and compare ours with the state-of-the-arts to validate the effectiveness of our model via three real-world datasets.

2. Related works

Multi-task learning (MTL) [20–24] intends to explore the task relationships. According to whether the data is well labeled, the multi-task learning algorithms can be categorized as **fully labeled multi-task learning** and **partially labeled multi-task learning**.

For the **fully labeled multi-task learning**, some supervised MTL algorithms assume that all tasks are related [20,21]. For example, multi-layered feed forward neural networks [3,25] use the hidden layer to represent common features from different tasks and predict the result using the output layer. However, unrelated tasks can be violated in many real applications and will degrade the performance accordingly. Multiple tasks clustering algorithms [26–28] aim to group all the tasks into several clusters where tasks within the cluster are either close to each other with some distance metric or share a common probabilistic prior. These algorithms are robust to outlier tasks due to separated clusters cannot affect each other; however, they will put negatively correlated tasks in different clusters. The Bayesian models [12,27,29,30] are also proposed for multi-task learning, such as Gaussian process [31,32], t process [33], Dirichlet process [7], etc. Another common assumption is that different tasks lie in a low dimensional subspace, which captures the predictive structure for all tasks. For example, [13–15,34–36] assume that there are a set of features (either in original space or in a transformed space) sharing for all tasks. There are also some multi-task learning algorithms using sparse constraints, such as ℓ_1 norm constraint [11], $\ell_{2,1}$ norm constraint [37], trace norm constraint [15,36], and the combination of them such as $\ell_1 + \ell_{1,q}$ norm multi-task learning [16], sparse and low-rank multi-task learning [13], robust multi-task learning using group sparse and low rank constraints [38], robust multi-task feature learning [39].

For the **partially labeled multi-task learning**, most multi-task learning algorithms design a semi-supervised learning framework by using both labeled and unlabeled data to improve the performance. Two questions need to be addressed: what semi-supervised classifier is designed with partially labeled data in the single task, and how to embed multiple classifiers within a unified sharing structure. In order to solve these, [40] formulates a reconstruction error for semi-supervised multi-view learning, which constructs a linear classifier for each task depending on the underlying task-specific data manifolds and integrates all the classification vectors together by a K-means like inter-task regularization term [17]. integrates semi-supervised regression and multi-task by assuming the kernel parameters of all tasks distributed in the same Gaussian process, which cannot take full advantage of the relationship between features and tasks underlying in the labeled data. With the help of graph Laplacian regularizer, [41] presents an online multi-task learning framework called ORION to estimate the optimal weights for combining the ensemble member forecasts [42]. proposes a semi-supervised autoencoder for multi-task learning [43]. designs an online semi-Supervised multi-task metric learning model [44]. analyzes the use of deep features applied in a semi-supervised multi-task framework [[45]. designs a semi-supervised multi-task learning using task regularizations [4]. proposes a semi-supervised multi-task learning by trying to find a common low-dimensional feature space structure shared by the multi-problems