Pattern Recognition 86 (2019) 248–264

Contents lists available at ScienceDirect

Pattern Recognition

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

Unsupervised Multi-task Learning with Hierarchical Data Structure

Wenming Cao[a], Sheng Qian[a], Si Wu[b], Hau-San Wong[a,*]

[a] Department of Computer Science, City University of Hong Kong, Hong Kong
[b] School of Computer Science and Engineering, South China University of Technology, China

ARTICLE INFO

Article history:
Received 24 November 2017
Revised 31 May 2018
Accepted 31 August 2018
Available online 21 September 2018

Keywords:
Multi-task learning
hierarchical structure
unsupervised learning
structural similarity,

ABSTRACT

Unsupervised multi-task learning exploits the shared knowledge to improve performances by learning related tasks simultaneously. In this paper, we propose an unsupervised multi-task learning method with hierarchical data structure. It strengthens similarities between instances in the same cluster, and increases diversities of instances by utilizing instances from related clusters. Firstly, we introduce **Rep**resentative **D**ual **F**eature**s** (RepDFs) that possess representative capabilities in the feature space and the sample space for each cluster concurrently. Secondly, we explore hierarchical structural similarities between clusters in related tasks from the topological perspective: 1) feature basis matrix, which learns compact representations for features in the feature space; and 2) sample refined matrix, which preserves local structures in the sample space. Thirdly, we adopt RepDFs to measure correlations between clusters and incorporate hierarchical structural similarities to conduct knowledge transfer among tasks. Experimental results on real-world data sets demonstrate the effectiveness and superiority of the proposed method over existing multi-task clustering methods.

1. Introduction

Unsupervised learning focuses on modeling the underlying structures or distributions of unlabeled input data. As one of unsupervised learning techniques, clustering has been applied in fields such as sequence analysis [1–4], medical imaging [5], and image segmentation [6–9], *etc*. The goal of traditional clustering methods is to split data instances into several clusters, thus data in the same cluster are similar. The assumption behind is that data in a given task follow certain independent and identical distribution, *i.e.*, traditional clustering deals with data of a single task. However it is common to encounter the situation where multiple unsupervised learning tasks are related, but the number of unlabeled instances in each task is not enough to achieve desiring performances, especially when features are high-dimensional. Although one can resort to the labelled information, the time cost and expenditure are prohibited. Considering the limitation of conventional clustering, the thought emerges that effectively explores the knowledge shared by related tasks and learns these tasks jointly. This thought coincides with supervised multi-task learning methods [10–14]. These methods identify the shared knowledge from task-specific knowledge and utilize the common knowledge to boost performances when tasks are related by constructing optimization objectives that integrate knowledge transfer into single-task cost function.

Inspired by supervised multi-task learning, unsupervised multi-task learning methods are proposed to improve clustering performance by utilizing common knowledge shared by related tasks. This shared knowledge is in various forms, such as shared features, shared reusable instances, shared model parameters and relatedness information. Methods [15–18] hold that related tasks follow an identical marginal label distribution and learn a shared subspace, where the knowledge in features can be transferred to each other. Specifically, the same orthonormal projection [15] is applied on tasks to obtain the shared subspace. The embedded feature space [16] is learnt to reduce the distributional divergence among tasks measured by the maximum mean discrepancy (MMD). In [17], two types of kernels are studied to explore the optimal reproducing kernel Hilbert space (RKHS) as the shared subspace: nonparametric kernel and spectral kernel. The low-dimensional shared subspace is investigated with the help of a covariance matrix containing relationships between features [18]. Furthermore, shared reusable instances are used to enrich the diversity of instances [19–21] by reweighing distances between pairs of instances or importance sampling. To be more specific, distances between instances of related tasks [19] are adjusted to control the bias between distributions of corresponding tasks. This bias balances the knowledge transfer among tasks to improve the clustering quality. In [20], reusable instances in the nearest clusters of two tasks are identified by evaluating the Jensen-Shannon divergence. For each

* Corresponding author. Tel.: +(852) 34428624.
E-mail addresses: wenmincao2-c@my.cityu.edu.hk (W. Cao), sqian9-c@my.cityu.edu.hk (S. Qian), cswusi@scut.edu.cn (S. Wu), cshswong@cityu.edu.hk (H.-S. Wong).

https://doi.org/10.1016/j.patcog.2018.08.021
0031-3203/

task, one can construct a similarity matrix by accumulating the useful information about reusable instances from other tasks to increase instance diversities. In [21], distances between instances of two tasks are reweighed in the shared subspace learnt by MMD. A similarity matrix is built using the nearest neighbors from inter-task and intra-task for each instance, thus both the within-task and cross-task information are used to increase instance diversities. Besides, shared model parameters provide an alternative under the assumption that label distributions of related tasks are similar. The method [18] holds that parameters determined by the covariance matrix containing the relatedness among task-specific models can help the knowledge transfer. Additionally, the relatedness information about clusters from tasks [22–24] is utilized to aid clustering. Particularly, imposing $\ell_{2,p}$ regularization on feature mapping matrices [22] helps uncover the relationships among tasks and adjust the amount of common knowledge shared by tasks. In [23], the task relationships are contained in the task regularization to boost individual clustering performances. This task regularization is measured by the divergence among density models of tasks learnt via mixture density models. It is noticed that the relationships among tasks have potential negative effects on the information transfer. This issue gets mitigated with kernel Bregman clustering [24]. Although unsupervised multi-task learning above has explored the transferable knowledge, the hierarchical structural information in the feature and sample spaces is seldom investigated. We consider that these hierarchical structural similarities in two spaces can provide supplementary information for correlations among tasks to help the shared information transfer among tasks.

In this paper, we propose an unsupervised multi-task learning method with hierarchical data structure (MTCHDS). Specifically, we first perform clustering in the sample space, followed by hierarchical clustering in the feature space. The latter clustering allows that there exist some overlaps between feature groups for different clusters. Second, we introduce RepDFs to measure correlations between clusters, and utilize instances in the nearest cluster pairs to enrich the diversities when performing multiple task clustering simultaneously. Third, we explore hierarchical structural similarities between clusters with the help of feature basis matrix and sample refined matrix. The former matrix learns the compact representations for features in the feature space, whereas the latter matrix aims to strengthen similarities between instances in the sample space. The hierarchical structural information obtained from the two spaces is combined with correlations of clusters to instruct the shared knowledge transfer among tasks. The main contributions of this paper are summarized as follows: 1) Overlaps between feature groups of different clusters are allowed to encourage the shared information. Meanwhile, we introduce RepDFs to evaluate correlations between clusters; 2) Hierarchical structural similarities between clusters are explored in the two spaces, and combined with correlations from RepDFs to guide knowledge transfer when utilizing instances from related clusters to enrich samples diversities; 3) We present an alternate optimization algorithm, and verify the effectiveness of MTCHDS via conducting experiments on several data sets.

2. Related Work

We review multi-task clustering methods which have been extensively investigated in the last decade. Jothi *et al.* [25] give the summary of multi-task clustering methods. Based on whether tasks are completely related and follow an identical distribution, these methods can be partitioned into multi-task completely related clustering and multi-task partially related clustering methods. The first category [15–18,22,26] assumes that the distributional divergences among related tasks can be ignored when learning the shared subspace. Specifically, Gu *et al.* [15] adopt the same orthonormal projection on related tasks to learn the shared subspace. This method fails to consider the distribution differences between tasks. Besides, it is required to set a proper dimensionality for the shared subspace. Two types of kernels [17] including nonparametric kernel and spectral kernel are studied to learn the optimal RKHS for multi-task clustering, which avoids setting the dimensionality of the subspace. But it needs to balance between the intra-task structural preservation and inter-task joint distribution learning. Zhang *et al.* [16] present a multi-task clustering via domain adaptation to search the optimal RKHS, in which MMD is used to minimize the distribution gap to strengthen relationships among tasks. But the assumption limits the feasibility that pairs of cluster centers from related tasks should be close to each other in the common embedded space. Yang *et al.* [22] propose to learn the feature mapping matrix and partition matrix concurrently, in which the shared low-dimensional features are explored by imposing the $\ell_{2,p}$ regularization on the feature mapping matrix. These features allow measuring the inter-task and intra-task correlations to achieve knowledge transfer among tasks. However, the divergences of feature weights on clusters of related tasks are easily suppressed due to this regularization. Al-Stouhi *et al.* [19] propose the symmetric multi-task clustering non-negative matrix factorization, where instances between related tasks are reweighed to adjust correlations among them. However, the biases among tasks are introduced, and this method focuses on binary classification problems. Xie *et al.* [26] combine within-task clustering and cross-task regularization into a unified multi-task clustering framework. This regularization contains the mutual information between feature spaces from tasks, which is evaluated by their joint probability distribution. However, the clustering quality depends on the difficult matrix approximation. Convex discriminative multi-task feature clustering [18] explores the task relationships via a covariance matrix that models correlations between features from tasks. This method fuses multi-task feature learning with maximum marginal clustering. All the methods work under the assumption that label distributions of related tasks are similar in the learnt shared subspace. This assumption naturally restricts the real-world applications of these methods.

To alleviate the above assumption, methods that belong to the second category including [18,20,21,23,24] are proposed and obtain better clustering performances when dealing with partially related tasks. These methods consider the distributional divergences between tasks, which plays an important role in learning task relationships and guiding knowledge transfer [17]. Specifically, Zhang *et al.* [23] develop the multi-task Bregman clustering method that exploits distributional divergences between clusters to construct the task relationships. This method converts multi-task clustering into approximating the joint data distributions of tasks, under the requirement that the learnt mixture densities for related tasks are similar to each other. Zhang et al. [20] propose the self-adapted multi-task clustering method. It treats instances from the nearest clusters from related tasks as reusable instances. These instances can evaluate correlations among tasks via kernel mean matching, thus the useful information from correlations are transferred. Zhang *et al.* [21] propose a multi-task clustering algorithm which transfers knowledge of instances. This method adjusts the distances between instances of related tasks and constructs a similarity matrix by considering the nearest neighbors from a given task and the other tasks. Zhang [18] introduces the convex discriminative multi-task relationship clustering by combining the multi-task relationship learning with the maximum marginal clustering. The task relationships are captured by a covariance matrix that models connections between task-specific regression coefficient matrices. Smart multi-task Bregman clustering [24] deals with negative effects of the cluster relationship transfer on within-task clustering [23] by comparing the values of the loss function before and af-