Neurocomputing 92 (2012) 133–144

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Clustering in applications with multiple data sources—A mutual subspace clustering approach

Ming Hua [a], Jian Pei [b,*]

[a] Facebook Inc., Palo Alto, CA, USA
[b] Simon Fraser University, Burnaby, BC, Canada

ARTICLE INFO

Available online 24 February 2012

Keywords:
Clustering
Multiple sources

ABSTRACT

In many applications, such as bioinformatics and cross-market customer relationship management, there are data from multiple sources jointly describing the same set of objects. An important data mining task is to find interesting groups of objects that form clusters in subspaces of the data sources jointly supported by those data sources.

In this paper, we study a novel problem of mining mutual subspace clusters from multiple sources. We develop two interesting models and the corresponding methods for mutual subspace clustering. The density-based model identifies dense regions in subspaces as clusters. The bottom-up method searches for density-based mutual subspace clusters systematically from low-dimensional subspaces to high-dimensional ones. The partitioning model divides points in a data set into k exclusive clusters and a signature subspace is found for each cluster, where k is the number of clusters desired by a user. The top-down method interleaves the well-known k-means clustering procedures in multiple sources. We use experimental results on synthetic data sets and real data sets to report the effectiveness and the efficiency of the methods.

1. Introduction

In many applications, there are multiple data sources. It is important to analyze data using the multiple data sources in an integrative way.

1.1. Motivation application examples and challenges

To develop effective therapies for cancers, both clinical data and genomic data have been accumulated for cancer patients. Examining clinical data or genomic data independently may not reveal the inherent patterns and correlations present in both data sets. Therefore, it is important to integrate clinical and genomic data and mining knowledge from both data sources.[1]

Clustering is a powerful tool for uncovering underlying patterns without requiring much prior knowledge about data. To discover phenotypes of cancer, subspace clustering has been widely used to analyze such data. However, in order to understand the clusters on clinical attributes well, and find out the genomic explanations, it is highly desirable to find clusters that are manifested in subspaces in both the clinical attributes and the genomic attributes. For a cluster mutual in a clinical subspace and a genomic subspace, we can use the genomic attributes to verify and justify the clinical attributes. The mutual clusters are more understandable and more robust. In addition, mutual subspace clustering is also helpful in integrating multiple sources.

As another example, consider cross-market customer relationship management. Customer behaviors in multiple markets (e.g., financial planning and investment, vacation expenditure, reading, entertainment and leisure expense) can be collected. Mutual subspace clustering can achieve more reliable customer segmentation. A mutual cluster which is a set of customers that are exclusively similar to each other in a subspace (i.e., some features) in each market is interesting, since we may use the features in different markets to explain their behaviors in the other markets. Mutual subspace clustering not only generates more robust clusters, but also integrates data from multiple sources and produces more understandable knowledge.

Recently, in a few applications such as bioinformatics, health-informatics and cross-market customer relationship management, attribute data about the same set of objects is collected from multiple aspects and/or sources. The availability of such data enables the verification and justification of learning from multiple sources, as demonstrated in recent research [14,15,26,27]. Particularly, joint clustering from multiple sources (e.g., [19,23,7]) which discovers clusters agreed by multiple sources has been found interesting and important in those applications.

* Corresponding author.
E-mail addresses: arceehua@fb.com (M. Hua), jpei@cs.sfu.ca (J. Pei).

[1] https://science.pfizer.com/content/we-link-genomics-and-clinical-data-for-enhancing-the-discovery-and-development-of-new-therapies/

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.08.032

In this paper, we study mining mutual subspace clusters for those applications with multiple data sources. In the clinical and genomic data analysis example, a mutual cluster is a subset of patients that form a cluster in both a subspace of the clinical data source and a subspace of the genomic data source. Such a mutual cluster may suggest the inherent connection between the genomic features and the clinical features.

Is mutual subspace clustering computationally challenging? One may consider the straightforward generate-and-test methods. For example, a simple method works in two steps. In the first step, we can find the complete set of possible subspace clusters in the first data source, say clinical data. Then, in the second step, we can check for each subspace cluster whether it is a subspace cluster in genomic data. Similarly, when clustering in the union space is feasible, we can first find all clusters in the union space with the constraint that the subspace of each cluster must contain at least one attribute from each clustering space. Then, we can check each cluster against the mutual clustering criteria.

However, such a two-step, generate-and-test method is problematic. Finding the complete set of possible subspace clusters in the clinical space or the union space of clinical data and genomic data is often very costly or even infeasible. For example, in the partitioning model (e.g., [1,2,25]), it is impossible to find all possible subspace clusterings. In some other models where clusters are not exclusive, there may be many subspace clusters in a large, high dimensional data set. Enumerating all possible subspace clusters explicitly and checking them one by one is often computationally expensive. In some models such as density-based clustering [3] and pattern-based clustering [24,20], enumerating all possible clusters is NP-hard.

1.2. Problem outline

While we will discuss the models of mutual subspace clustering in Sections 3.1 and 4.1, the problem can be generally described as follows.

We model a data source as a set of points in a clustering space. Let S_1 and S_2 be two *clustering spaces* where $S_1 \cap S_2 = \emptyset$, and O be a set of points in space $S_1 \cup S_2$ on which the subspace clustering analysis is applied. It is up to users to choose clustering spaces. The only requirement here is that each point appears in both clustering spaces.

A *mutual subspace cluster* is a triple (C,U,V) such that $C \subseteq O$, $U \subseteq S_1$, $V \subseteq S_2$, and C is a cluster in both U and V, respectively. U and V are called the *signature subspaces* of C in S_1 and S_2, respectively. To keep our discussion simple, we consider only two clustering spaces in this paper. However, our model can be easily extended to situations where more than two clustering spaces present.

What is the critical difference between mutual subspace clustering on multiple spaces and traditional subspace clustering in one space? Technically, one may think that we can find subspace clusters in the union space $S_1 \cup S_2$ with the constraint that the subspaces must contain attributes from both S_1 and S_2. Suppose C is a cluster in subspace $W \subseteq S_1 \cup S_2$. Then, we can assign $U = W \cap S_1$ and $V = W \cap S_2$ as the signature subspaces of C. Does this straightforward extension work?

Example 1 (*Mutual clustering*). Fig. 1 shows a synthetic data set. Let the clustering space S_1 be X and the clustering space S_2 be Y. The union space is the two-dimensional space as shown. There are three clusters (annotated as A, B and C in the figure) in the union space.

Mutual clustering from the clustering spaces S_1 and S_2 can help us to understand how the two attributes agree with each other in clusters. For example, cluster C is a good mutual cluster, since its projections on both S_1 and S_2 are also clusters. However, although

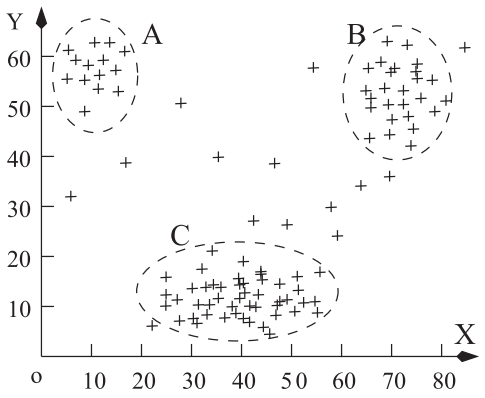

Fig. 1. An example of mutual clusters.

clusters A and B are clusters in the union space, each of them is not a distinguishing cluster in subspace S_2 (i.e., Y). They are mixed together in S_2. Thus, A and B are not good mutual clusters. □

Moreover, in real applications, different similarity measures and even clustering criteria may be adopted in different clustering spaces. In such a case, it is very difficult or even impossible to define an appropriate similarity measure and clustering criteria in the union space. Clustering in the union space becomes infeasible.

From the above example, we can see that mutual subspace clustering from multiple clustering spaces is critically different from subspace clustering in one (union) clustering space. A mutual cluster must be a cluster in a signature subspace of each clustering space. Mutual subspace clustering finds the common clusters agreed by subspace clustering in both clustering spaces, which cannot be handled by the traditional subspace clustering analysis.

In this paper, we study the mutual subspace clustering problem and make the following contributions. First, we identify the novel mutual subspace clustering problem, and elaborate its potential applications. Second, we develop two interesting models and the corresponding methods for mutual subspace clustering. The density-based model identifies dense regions in subspaces as clusters. The bottom-up method searches for density-based mutual subspace clusters systematically from low-dimensional subspaces to high-dimensional ones. Information from multiple sources is used to guide the search. The partitioning model divides points in a data set into k exclusive clusters and a signature subspace is found for each cluster, where k is the number of clusters desired by a user. The top-down method interleaves the well-known k-means clustering procedures in multiple sources. Third, we use experimental results on synthetic data sets and real data sets to report the effectiveness and the efficiency of the methods.

The rest of the paper is organized as follows. The related work is reviewed in Section 2. The density-based, bottom-up method and the partitioning, top-down method are developed in Sections 3 and 4, respectively. The experimental results are reported in Section 5. Section 6 discusses the related issues and concludes the paper.

2. Related work

Our work is generally related to subspace clustering, clustering from multiple sources, and multiview learning. In this section, we review those areas briefly.

2.1. Subspace clustering

Subspace clustering has attracted substantial interest due to its successful applications in some new domains such as