Pattern Recognition Letters 51 (2015) 70–78

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Incorporating side information into multivariate Information Bottleneck for generating alternative clusterings ☆

Yangdong Ye , Ruina Liu, Zhengzheng Lou*

School of Information Engineering, Zhengzhou University, Zhengzhou 450052, China

ARTICLE INFO

Article history:
Received 25 January 2014
Available online 9 October 2014

Keywords:
Alternative clustering
Multivariate Information Bottleneck
Multi-view clustering
Side information

ABSTRACT

Traditional clustering algorithms aim to find a single clustering of data. However, it is difficult to put an accurate interpretation on the complex data and there will be multiple different meaningful explanations. For such situation, this paper presents a novel alternative clustering algorithm, which takes existing reference clusterings as side information and incorporates such information into the multivariate Information Bottleneck (IB) method. The side information is used to lead the learning algorithm to generate an alternative clustering that is different from the existing reference clusterings, while the multivariate IB method guarantees the quality of new clustering results. Our method has the ability to incorporate multiple existing reference clusterings into the alternative cluster learning process, and can be used to analyze both co-occurrence data and non co-occurrence data. Moreover, our method is able to discover non-linear alternative clusterings. The experimental results on synthetic and real-world datasets demonstrate that the performance of the proposed algorithm is superior to the existing state-of-the-art alternative clustering algorithms.

1. Introduction

Clustering aims to discover reasonable partitions of data. However, traditional clustering algorithms focus on learning a single partition, while complex high-dimensional data can be usually interpreted from multiple views. For example, the movie data can be categorized either by genres or by directors, and the gene data can be interpreted according to gene functions or structures. For such situation, only one clustering solution is not enough to thoroughly understand the data, and we need to analysis the data from multiple views. This issue has recently led to the emerging research area of multi-view clustering [1–4], which aims to mine multiple high-quality and non-redundant clusterings of a given dataset.

Current algorithms for multi-view clustering can be roughly classified into two paradigms. The first adopts a simultaneous way to explore multiple non-redundant clustering solutions from the given dataset [5,6,2]. The other one, which is called alternative clustering, adopts an alternative way to generate a new clustering partition that is different from existing reference clusterings [1,7–11]. The main difference between these two paradigms is that the first one explores multiple clusterings simultaneously, while the latter one utilizes prior knowledge of existing clustering partitions to discover new clustering solutions one by one. In this paper, we focus on the research of alternative clustering.

There are two basic goals of alternative clustering. The first one is that the new clustering partition should reveal some meaningful explanations of the data. The second goal requires that the new clustering solution should be non-redundant referring to existing reference clusterings. To achieve these goals, some works [1,7,8,12] assume that different clustering partitions can be revealed in different feature spaces, and transform the data feature space via some feature transformation methods (e.g., feature selection, feature weighting) so that the traditional clustering algorithms can discover different clustering partitions in the transformed feature spaces. Other works [13–15,9,16,10] define some objective functions to lead the alternative clustering algorithms to find a high-quality clustering partition that is different from the existing reference clusterings.

This paper presents a novel alternative clustering algorithm, named SmIB, which takes existing reference clusterings as side information and incorporates such information into the multivariate Information Bottleneck (IB) method [15]. The side information is used to lead the learning algorithm to generate an alternative clustering that is different from the existing reference clusterings, while the multivariate IB method guarantees the quality of new clustering results.

Our approach is closely related to the Conditional Information Bottleneck (CIB) algorithm [13], Parallel Information Bottleneck (PIB) algorithm [15], NACI algorithm [9] and minCEntropy algorithm

☆ This paper has been recommended for acceptance by A. Petrosino.

* Corresponding author. Tel.: +86 13623859902.
E-mail addresses: yeyd@zzu.edu.cn (Y. Ye), rnliu.xt@gmail.com (R. Liu), zzlou@zzu.edu.cn (Z. Lou).

http://dx.doi.org/10.1016/j.patrec.2014.08.007
0167-8655/

Table 1
The comparison of different algorithms.

Algorithms	Reference clusterings	Data types	Information measurements
PIB	n	Co-occurrence data	Multi-information
CIB	1	Co-occurrence data	Conditional mutual information
NACI	1	Non-limited, but more suitable for low dimension data	Quadratic mutual information based on Parzen window density estimation technique
minCEntropy	n	Non-limited, but more suitable for low dimension data	Conditional quadratic Havrda–Charvat's entropy based on Parzen window density estimation technique
SmIB	n	Non-limited	Mutual information and MeanNN differential entropy estimator

[16]. All these approaches employ the information measurements to measure clustering quality and redundancy between the alternative clustering result and the existing reference clusterings. The main differences between the SmIB algorithm and the aforementioned methods include the number of existing clusterings that can be incorporated, the data types that can be analyzed and the information measurements employed. These differences are summarized in Table 1. As observed from this table, there are some limitations of these state-of-the-art methods: (1) CIB and NACI incorporate only one reference clustering; (2) CIB and PIB are just suitable for analyzing co-occurrence data; (3) NACI and minCEntropy employ the Parzen window probability density estimator [17], which needs to specify Gaussian kernel parameters in advance and is unsuitable to estimate the probability density of high-dimensional data. Compared with these methods, SmIB has the following properties:

- Based on the multivariate IB method, SmIB can incorporate multiple reference clusterings into the alternative cluster learning process;
- SmIB employs mutual information [18] and MeanNN differential entropy estimator [19] to measure the information resided in data, which makes it be suitable for analyzing both co-occurrence data and non co-occurrence data, and have the ability to find non-linear alternative clusterings of the data;
- SmIB does not require any supernumerary parameters when estimating the information resided in data.

The experimental results on synthetic datasets, CMUFace dataset [20], WebKB dataset[1] and other real-world datasets [20] demonstrate the above properties of the SmIB algorithm and show that its performance is superior to the existing state-of-the-art alternative clustering algorithms.

2. Background

The following notations are used throughout this paper. $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ denotes the collection of data and $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$ denotes the corresponding features, where $x_i = [y_1^{x_i}, y_2^{x_i}, \ldots, y_m^{x_i}]$. $\mathcal{T} = \{t_1, t_2, \ldots, t_M\}$ denotes the clusters. Let X, Y, T be three discrete random variables, taking values from $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{T}$, respectively.

2.1. Side information

Traditionally, only the feature vectors are provided for the clustering algorithms. However, in real world applications, there will be some side information other than what is contained in feature vectors, which can be used to help the clustering algorithms reveal data patterns. Examples of side information include instance-level constraints [21], existing partitions [13], auxiliary attributes (e.g. links, user-access behavior, etc.) in text documents [22], etc.

In alternative clustering algorithms, existing reference clusterings are employed to lead the learning algorithms to find a new clustering solution of the data. In this paper, the existing clusterings are taken as one type of side information. Fig. 1 demonstrates the alternative clustering with side information. In the original data space, there are four data points. When we want to divide these four data points into two groups, the partitions $\{\{x_1, x_2\}, \{x_3, x_4\}\}$ and $\{\{x_1, x_4\}, \{x_2, x_3\}\}$ are relatively obvious and can be discovered easily. Given these two partitions, they can be taken as side information, which are denoted by S_1 and S_2 in Fig. 1. After incorporating the side information into the cluster learning process, the alternative clustering partition $\{\{x_1, x_3\}, \{x_2, x_4\}\}$ can be discovered.

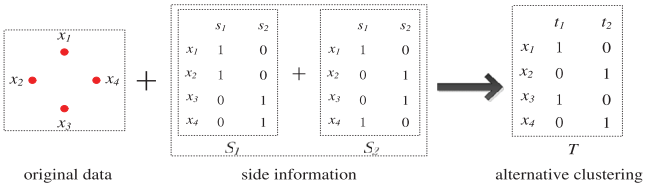

Fig. 1. Alternative clustering with side information.

2.2. Multivariate Information Bottleneck

The Information Bottleneck (IB) method [23] is an information-theoretic based data analysis method, which treats the pattern extraction from data as a process of data compression. As an extension of IB method, the multivariate IB method [24,15] provides a general principled framework for multiple variable compressing problem, which can be used to deal with more complex data analysis tasks [25,13,26,27].

Given a set of observed variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ and a set of compressed variables $\mathbf{T} = \{T_1, \ldots, T_k\}$, the multivariate IB method uses two Bayesian networks G_{in} and G_{out} to specify the relationship among variables. G_{in} specifies the compression relationship where $T_j \in \mathbf{T}$ is the compressed version of a subset of the observed variables denoted by $U_j \subset \mathbf{X}$. G_{out} specifies the dependent relationship from $\mathbf{T}$ to $\mathbf{X}$ which talks what information the compressed variables should maintain. Let $I^{G_{\text{in}}}$ denote the information that we want to minimize in G_{in} and $I^{G_{\text{out}}}$ denote the information that we would like to preserve in G_{out}, the objective function of multivariate IB is defined as:

$$L_{\min}[P(T_1|U_1), \ldots, P(T_k|U_k)] = I^{G_{\text{in}}} - \beta I^{G_{\text{out}}}, \tag{1}$$

where $\beta \in [0, \infty)$ is a positive Lagrange multiplier controlling the trade-off between minimizing the compression-information $I^{G_{\text{in}}}$ and maximizing the relevant information $I^{G_{\text{out}}}$.

2.3. Information measurements

Mutual information [18] and MeanNN differential entropy estimator [19] are adopted to measure the information resided in co-occurrence and non co-occurrence data, respectively.

2.3.1. Information measurement for co-occurrence data

Since the co-occurrence data can be easily converted into a joint distribution according to the corresponding co-occurrence matrix,

[1] www.cs.cmu.edu/~webkb.