Applied Mathematics and Computation 335 (2018) 264–277

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

A sequential ensemble clusterings generation algorithm for mixed data

Xingwang Zhao, Fuyuan Cao, Jiye Liang*

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China

ARTICLE INFO

Keywords:
Ensemble clustering
Base clustering
Mixed data
Information entropy

ABSTRACT

Ensemble clustering has attracted much attention for its robustness, stability, and accuracy in academic and industry communities. In order to yield base clusterings with high quality and diversity simultaneously in ensemble clustering, many efforts have been done by exploiting different clustering models and data information. However, these methods neglect correlation between different base clusterings during the process of base clusterings generation, which is important to obtain a quality and diverse clustering decision. To overcome this deficiency, a sequential ensemble clusterings generation algorithm for mixed data is developed in this paper based on information entropy. The first high quality base clustering is yield by maximizing the entropy-based criterion. Afterward, a sequential paradigm is utilized to incrementally find more base clusterings, in which the diversity between a new base clustering and the former base partitions is measured by the normalized mutual information. Extensive experiments conducted on various data sets have demonstrated the superiority of our proposal as compared to several existing base clusterings generation algorithms.

1. Introduction

Clustering analysis is one of the primary techniques in data mining and machine learning. Its aim is to partition a set of unlabeled objects into several distinct clusters so that the data objects in the same cluster are similar and dissimilar to the data objects in other clusters. It has numerous applications in such areas as customer segmentation, target marketing, bioinformatics, social network analysis, and scientific data analysis [1–4].

In real applications, analyzed data sets are often comprised of mixed numerical and categorical attributes, particularly when they are merged from different sources. In other words, data are in a mixed mode. Such data may be encountered, for instance, in medical diagnosis analyses, in the analysis of survey data, as well as in image analysis. For example, the attributes of the data about medical diagnosis may include sex, age, weight, and blood pressure of the patients, where the attribute sex is categorical, the other attributes are numerical. In the past five decades, various clustering algorithms have been proposed in the literature [2–5]. However, the primary focus of these clustering algorithms has been on the data sets with either numerical attributes or categorical attributes. It is difficult to apply traditional clustering algorithm directly into mixed data. Therefore, mixed data clustering becomes not only a difficult task but also a challenging and promising one to attract many researchers in data mining and machine learning field.

* Corresponding author.
E-mail addresses: zhaoxw84@163.com (X. Zhao), cfy@sxu.edu.cn (F. Cao), ljy@sxu.edu.cn (J. Liang).

https://doi.org/10.1016/j.amc.2018.04.035
0096-3003/

Currently, in clustering analysis, there are usually two categories of methods to process mixed data. One category is to transform either categorical attributes into numerical attributes or numerical attributes into categorical attributes. Then, the clustering methods for numerical data or categorical data can be used. However, these methods are not effective since the similarity measure of the transformed data could not represent the similarity of original mixed data. The other category is to extend the clustering algorithms for numerical data or categorical data to match with mixed data to improve the clustering result. Using these two strategies, some clustering algorithms for mixed data have been developed in the literature [7–11].

Although there are many mixed data clustering algorithms, Kuncheva et al.[12] pointed out that there is no single clustering algorithm which performs best for all data sets and can discover all types of clusters and structures. Each algorithm has its own strength and weakness. For a given mixed data set, different clustering algorithms, or even the same algorithm with different parameters, usually obtain distinct clustering results. Therefore, it is difficult for users to decide which algorithm would be a proper choice for clustering the given data set. To overcome these limitations, ensemble clustering algorithms have recently emerged as a powerful alternative to standard clustering algorithms. Their main objective is to improve the robustness as well as the quality of clustering results, by combining different clustering decisions according to some criterion. Generating a set of base clusterings is a key process in ensemble clustering [13,14]. Examples of well-known ensemble clustering generation algorithms include running a single clustering algorithm with different initialization [16–18], carrying out one or more clustering algorithms on different subspaces or subsamples of a given data set [19–21], and performing different clustering algorithms [22,27,28].

Despite notable success, these algorithms generate the different base clustering results independently. The correlation between different base clusterings during the process of base clusterings generation is neglected, which is important to obtain a quality and diverse base clustering decision. To overcome this deficiency, a sequential ensemble clusterings generation algorithm is developed in this paper based on information entropy for mixed data. The first high quality base clustering is yield by maximizing the entropy-based criterion. Afterward, a sequential paradigm is utilized to incrementally find more base clusterings, in which the diversity between a new base clustering and the former base partitions is measured by the normalized mutual information. Extensive experiments conducted on various data sets have demonstrated the superiority of our proposal as compared to several existing base clusterings generation algorithms.

The rest of this paper is organized as follows: Section 2 reviews the related work on mixed data clustering and ensemble clustering problem. The proposed sequential ensemble clusterings generation algorithm is introduced in Section 3. Then, Section 4 exhibits the evaluation of this new algorithm against other ensemble clusterings generation algorithms over real data sets. The paper is concluded in Section 5.

2. Related work

In this section, mixed data clustering algorithms and some recent developments on ensemble clustering are reviewed.

2.1. Mixed data clustering

Data sets analyzed in practice are commonly characterized by mixed numerical and categorical attributes. One of the most common approaches to cluster mixed data involves converting the data set to a single data type, and applying standard clustering algorithms to the transformed data. For example, He et al. [6] considered a numerical attribute as a category by discretization. Then they extended their earlier clustering algorithm of categorical data to cluster mixed data.

An alternative approach is to design a generalized similarity or distance measure for mixed data, and apply it to the existing clustering algorithms. K-prototype [7] is one of the most famous algorithms. It integrates the k-means and the k-modes algorithms by defining a combined dissimilarity measure to enable clustering of mixed numerical and categorical attributes. Ahmad and Dey [8] proposed a distance metric for mixed data clustering based on the co-occurrence likelihood of two categorical attribute values. Li and Biswas. [9] presented an agglomerative hierarchical clustering algorithm based on Goodall similarity measure for mixed data. Hsu et al. [10] proposed a mixed data clustering algorithm applying the idea of distance hierarchy to calculate distance for every categorical attribute. This algorithm, however, requires domain-specific knowledge to build distance hierarchy which is not available for a large number of attribute domains. Liang et al. [11] proposed an algorithm to cluster mixed data by defining two kinds of information entropy measures for numerical and categorical data, respectively. Gower [29] introduced a similarity index that measures the similarity between two mixed data. And it is used to cluster mixed data in the framework of the k-means type algorithm.

Additionally, some mixed data clustering algorithms based on statistical models are developed recently, which typically assume the observations follow a normal-multinomial finite mixture model. Readers with interests can refer to the survey paper for more comprehensive understanding [30].

2.2. Ensemble clustering

Like ensemble methods in supervised learning, ensemble clustering methods work in two steps, clustering generation and clustering combination. The quality and diversity of the base clusterings are two major factors, which affect the performance of an ensemble clustering method. As a result, several heuristics have been proposed to generate different clusterings for a given data set, which can be classified into three categories: