Pattern Recognition 46 (2013) 2228–2238

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number

Yiu-ming Cheung [a,b,*], Hong Jia [a]

[a] Department of Computer Science and Institute of Computational and Theoretical Studies, Hong Kong Baptist University, Hong Kong, China
[b] United International College, Beijing Normal University-Hong Kong Baptist University, Zhuhai, China

ARTICLE INFO

Article history:
Received 19 March 2012
Received in revised form
29 December 2012
Accepted 23 January 2013
Available online 31 January 2013

Keywords:
Clustering
Similarity metric
Categorical attribute
Numerical attribute
Number of clusters

ABSTRACT

Most of the existing clustering approaches are applicable to purely numerical or categorical data only, but not the both. In general, it is a nontrivial task to perform clustering on mixed data composed of numerical and categorical attributes because there exists an awkward gap between the similarity metrics for categorical and numerical data. This paper therefore presents a general clustering framework based on the concept of object-cluster similarity and gives a unified similarity metric which can be simply applied to the data with categorical, numerical, and mixed attributes. Accordingly, an iterative clustering algorithm is developed, whose outstanding performance is experimentally demonstrated on different benchmark data sets. Moreover, to circumvent the difficult selection problem of cluster number, we further develop a penalized competitive learning algorithm within the proposed clustering framework. The embedded competition and penalization mechanisms enable this improved algorithm to determine the number of clusters automatically by gradually eliminating the redundant clusters. The experimental results show the efficacy of the proposed approach.

1. Introduction

To discover the natural group structure of objects represented in numerical or categorical attributes [1], clustering analysis has been widely applied to a variety of scientific areas such as computer science [2] and bioinformatics [3]. Traditionally, clustering analysis concentrates on purely numerical data only. The typical clustering algorithms include the *k*-means [4], EM algorithm [5] and their variants. Since the objective functions of these two algorithms are both numerically defined, they are not essentially applicable to the data sets with categorical attributes. Under the circumstances, a straightforward way to overcome this problem is to transform the categorical values into numerical ones, e.g. the binary strings, and then apply the aforementioned numerical-value based clustering methods. Nevertheless, such a method has ignored the similarity information embedded in the categorical values and cannot faithfully reveal the similarity structure of the data sets [6]. Hence, it is desirable to solve this problem by finding a unified similarity metric for categorical and numerical attributes such that the metric gap between numerical and categorical data can be eliminated. Subsequently, a general clustering algorithm which is applicable to numerical and categorical data can be presented based on this unified metric. During the past decades, some works which try to find a unified similarity metric for categorical and numerical attributes have been presented, e.g. see [7]. However, a computational efficient similarity measure remains to be developed.

Another challenging problem encountered in clustering is how to determine the number of clusters. To the best of our knowledge, a lot of popular clustering methods, e.g. the *k*-means algorithm for numerical data clustering and the *k*-modes algorithm [8] for categorical data clustering, need to pre-assign the number of clusters exactly. Otherwise, they will almost always lead to a poor clustering result [9,10]. Unfortunately, in many cases, this vital information is not always available from the practical viewpoint. Hence, to explore an algorithm which can conduct clustering without knowing cluster number is also a significant work in clustering analysis. To address this issue, variant researches have been conducted in the literature and some feasible methods that can determine the number of clusters for purely numerical or categorical data have been presented [9–11]. Nevertheless, to the best of our knowledge, how to automatically select cluster number for mixed data during clustering process is still an unsolved problem.

In this paper, we will propose a unified clustering approach that is capable of selecting the cluster number automatically for both categorical and numeric data sets. Firstly, we present a general clustering framework based on the concept of object-

[*] Corresponding author. Tel.: +852 34115155.
E-mail addresses: ymc@comp.hkbu.edu.hk (Y.-m. Cheung), hjia@comp.hkbu.edu.hk (H. Jia).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.01.027

cluster similarity. Then, a new metric for both of numerical and categorical attributes is proposed. Under this metric, the object-cluster similarity for either categorical or numerical attributes has a uniform criterion. Hence, transformation and parameter adjustment between categorical and numerical values in data clustering are circumvented. Subsequently, an iterative clustering algorithm is introduced. This algorithm conducts a parameter-free clustering analysis and is applicable to the three types of data: numerical, categorical, or mixed data, i.e., the data with the both of numerical and categorical attributes. Moreover, empirical studies show that the proposed algorithm has higher accuracy as well as lower computational cost compared to the popular *k*-modes algorithm for categorical data clustering. For mixed data clustering, compared to *k*-prototype algorithm [12], the proposed method can get much better clustering results, but no parameter needs to be adjusted at all. Additionally, to overcome the cluster number selection problem, we further present a penalized competitive learning algorithm within the proposed clustering framework. The competition and penalization mechanisms in this improved algorithm can gradually fade out the redundant clusters. Hence, the number of clusters can be determined automatically during the clustering process. Experimental results on benchmark data sets have shown the effectiveness of this method.

The rest of this paper is organized as follows. Related works are reviewed in Section 2. Section 3 proposes a general clustering framework based on object-cluster similarity, whose metric is also defined. Section 4 describes an iterative clustering algorithm and Section 5 presents an improved one with capability of automatically selecting cluster number. Experiments are conducted in Section 6. Finally, we draw a conclusion in Section 7.

2. Related works

This section reviews the related works on: (1) data clustering with categorical-and-numerical attributes and (2) cluster number selection.

In the former, several methods have been presented which can be grouped into two lines. In the first line, the algorithms are essentially designed for purely categorical data, although they have been applied to the mixed data as well by transforming the numerical attributes to categorical ones via a discretization method. Along this line, several methods have been proposed based on the perspective of similarity metric, graph partitioning or information entropy. For example, ROCK algorithm proposed by Guha et al. [13] is an agglomerative hierarchical clustering procedure based on the concepts of neighbors and links. In this method, a pair of objects are regarded as neighbors if their similarity exceeds a certain threshold, and the desired cluster structure is obtained by merging the clusters sharing a pre-assigned number of neighbors gradually. ROCK has shown its superiority over traditional hierarchical algorithms in the experiments, but its performance is generally sensitive to the setting of similarity threshold. Also, the computation of links between objects is quite time-consuming [14]. By contrast, CLICKS algorithm proposed in [15] mines subspace clusters for categorical data sets. This method encodes a data set into a weighted graph structure, where each weighted vertex stands for an attribute value and two nodes are connected if there is a sample in which the corresponding attribute values co-occur. Experiments have shown that CLICKS outperforms ROCK algorithm and scales better for high-dimensional data sets. However, its performance also depends upon a set of parameters whose tuning is quite difficult from the practical viewpoint. Additionally, the COOLCAT algorithm, an entropy-based method proposed by Barbara et al. [16], utilizes the information entropy to measure the closeness between objects and presents a scheme to find a clustering structure via minimizing the expected entropy of clusters. The performance of this algorithm is stable for different data sizes and parameter settings. Furthermore, a scalable algorithm for categorical data clustering called LIMBO [17], which is proposed based on the Information Bottleneck (IB) framework [18], employs the concept of mutual information to find a clustering with minimum information loss. In general, all of the above-stated algorithms can be applied to mixed data via a discretization process, which may, however, cause loss of important information, e.g. the difference between numerical values.

By contrast, the second line attempts to design a generalized clustering criterion for numerical-and-categorical attributes. For example, Li and Biswas [7] presented the Similarity Based Agglomerative Clustering (SBAC) algorithm which is based on Goodall similarity metric [19] that assigns a greater weight to uncommon feature value matching in similarity computations without the prior knowledge of the underlying distributions of the feature values. This method has a good capability of dealing with the mixed attributes, but its computation is quite laborious. He et al. [20] extended the Squeezer algorithm to cluster mixed data and proposed the usm-squeezer method, in which the similarity measure for categorical attributes is the same as the Squeezer while the similarity of numerical attributes is defined by relative difference. However, the clustering effectiveness of this method has not been sufficiently demonstrated. In [21], an Evidence-Based Spectral Clustering (EBSC) algorithm has been proposed for mixed data clustering by integrating the evidence based similarity metric into the spectral clustering structure. Moreover, the AUTOCLASS proposed by Cheeseman and Stutz [22] assumes a classical finite mixture distribution model on mixed data and utilizes a Bayesian method to derive the most probable class distribution for the data given prior information. Among this category of approaches, the most cost-effective one may be the *k*-prototype algorithm proposed by Huang [12]. In this method, the distance between two categorical values is defined as 0 if they are the same, and 1 otherwise while the distance between numerical values are quantified with Euclidean distance. Subsequently, the *k*-means paradigm is utilized for clustering. However, since different metrics are adopted for numerical and categorical attributes, a user-defined parameter is utilized to control the proportions of numerical distance and categorical distance. Nevertheless, the clustering result is very sensitive to the setting of this parameter. A simplified version of *k*-prototype algorithm namely *k*-modes [8,23,24], which is applicable for purely categorical data clustering, has also been widely utilized. Thus far, different improvement strategies on this method have been explored, e.g. see [25–27].

In general, all of the aforementioned methods need to pre-assign the number of clusters exactly, which is, however, a nontrivial task from the practical viewpoint. In the literature, a variety of methods have been proposed for cluster number estimation. For example, some computational demanding methods choose the optimal number of clusters via different statistic criteria, such as Akaike's Information Criterion (AIC) [28] and Schwarz's Bayesian inference criterion (BIC) [29]. By contrast, another kind of methods within the framework of competitive learning often introduce some competitive mechanisms, such as penalization [9,11] and cooperation [30], into the clustering process so that the number of clusters can be automatically selected. Nevertheless, these existing methods focus on numerical data only and cannot be directly applied to data sets with categorical attributes. Recently, Liao and Ng [10] have introduced an entropy penalty term into the objective function of *k*-modes algorithm. Then, by choosing different values for the regularization parameter, variant clustering results with different cluster