Knowledge-Based Systems 67 (2014) 401–411

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

MGR: An information theory based hierarchical divisive clustering algorithm for categorical data

Hongwu Qin [a,b], Xiuqin Ma [a,b,*], Tutut Herawan [c], Jasni Mohamad Zain [a]

[a] Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang, Lebuhraya Tun Razak, Gambang, 26300 Kuantan, Malaysia
[b] College of Computer Science & Engineering, Northwest Normal University, 730070 Lanzhou Gansu, PR China
[c] Faculty of Computer Science and Information Technology, University of Malaya, 50603 Pantai Valley, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 26 December 2011
Received in revised form 19 March 2014
Accepted 20 March 2014
Available online 27 March 2014

Keywords:
Data mining
Clustering
Categorical data
Gain ratio
Information theory

ABSTRACT

Categorical data clustering has attracted much attention recently due to the fact that much of the data contained in today's databases is categorical in nature. While many algorithms for clustering categorical data have been proposed, some have low clustering accuracy while others have high computational complexity. This research proposes mean gain ratio (MGR), a new information theory based hierarchical divisive clustering algorithm for categorical data. MGR implements clustering from the attributes viewpoint which includes selecting a clustering attribute using mean gain ratio and selecting an equivalence class on the clustering attribute using entropy of clusters. It can be run with or without specifying the number of clusters while few existing clustering algorithms for categorical data can be run without specifying the number of clusters. Experimental results on nine University of California at Irvine (UCI) benchmark and ten synthetic data sets show that MGR performs better as compared to baseline algorithms in terms of its performance and efficiency of clustering.

1. Introduction

Clustering is an important data mining technique which partitions a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters [37]. Most previous clustering algorithms focus on numerical data whose inherent geometric properties can be exploited naturally to define distance functions between objects. However, many fields, from statistics to psychology, deal with categorical data. Unlike numerical data, it cannot be naturally ordered. An example of categorical attribute is *color* whose values include *red*, *green*, *blue*, etc. Therefore, those clustering algorithms dealing with numerical data cannot be used to cluster categorical data. Recently, the problem of clustering categorical data has received much attention.

A number of algorithms have been proposed for clustering categorical data [1–23,25–34,38–41]. Similar to other clustering problems, categorical data clustering can also be considered as an optimization problem [17], thus a typical method for clustering categorical data is to define a dissimilarity measure between objects, an objective function, and then iteratively minimize or maximize the objective function until a solution is found. Unfortunately, this optimization problem is NP-complete. Therefore most researchers resort to heuristic methods to solve it. ROCK [2], k-modes [5], and k-ANMI [20] are representative examples of such type of methods. These methods require the user to specify the number of clusters first and then conduct the processes of initialization, iteration, and so on. They focus on the relationship between the objects and clusters during the process of clustering, as a result, their time complexity increases greatly with the increase in the number of objects. We can say these methods have implemented clustering from the viewpoint of objects. As we know, a data set consists of two elements: objects and attributes. Besides objects, attributes are also an important aspect to be considered for clustering. Generally, the number of attributes is much less than the number of objects in a data set, thus it is possible to improve the clustering efficiency if we employ attributes for clustering. The following example reveals the potential of attributes for categorical data clustering.

Table 1 shows a categorical data set with ten objects and five attributes. The column of real classes implies that the set of objects can be partitioned into three classes. We assume that the objects in each class are the same while completely distinct from the objects

* Corresponding author at: Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang, Lebuhraya Tun Razak, Gambang, 26300 Kuantan, Malaysia.

E-mail addresses: qhwump@gmail.com (H. Qin), xueener@gmail.com (X. Ma), tutut@um.edu.my (T. Herawan), jasni@ump.edu.my (J.M. Zain).

http://dx.doi.org/10.1016/j.knosys.2014.03.013
0950-7051/

Table 1
Example data set with ten objects and five attributes.

Objects	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Real classes
O_1	A_1	A_2	A_3	A_4	A_5	1
O_2	B_1	B_2	B_3	B_4	B_5	2
O_3	B_1	B_2	B_3	B_4	B_5	2
O_4	C_1	C_2	C_3	C_4	C_5	3
O_5	A_1	A_2	A_3	A_4	A_5	1
O_6	B_1	B_2	B_3	B_4	B_5	2
O_7	A_1	A_2	A_3	A_4	A_5	1
O_8	C_1	C_2	C_3	C_4	C_5	3
O_9	B_1	B_2	B_3	B_4	B_5	2
O_{10}	C_1	C_2	C_3	C_4	C_5	3

in other classes. A_i, B_i, and C_i for $i = 1, 2, 3, 4, 5$ denote different categories on ith attribute. The user is required to cluster the data set without knowing the real classes in advance.

Using the methods mentioned above, the user has to specify the number of clusters first. Imagine the number of clusters is set to two, the accuracy of clustering will be affected. In fact, from the viewpoint of attributes, it can be seen that each attribute partitions the data set in the same way. If we can find such relation between the attributes, a perfect clustering of the data set including three clusters will be obtained by using the partition defined by any attribute without specifying the number of clusters in advance. Obviously, using attributes to cluster the data set in this example is a more natural way.

In a real life categorical data set, the partitions defined by attributes are not as perfect as that in the above example (i.e. the partitions defined by attributes are not always the same); however, if the real classes are sufficiently distinguishable from each other, the objects in the same real classes will create distinct values on some attributes from the objects in the other real classes, consequently, there exist some partitions defined by attributes which are similar to the real clustering of objects; at least, there exist some equivalence classes (the set of objects which has the same value of the attribute) in these partitions which are similar to the real classes. Our goal is to find such partitions and equivalence classes to construct the clustering of the objects.

In this paper, a novel information theory based hierarchical divisive clustering algorithm for categorical data, namely MGR, is proposed. MGR iteratively performs two steps on the current data set: selecting a clustering attribute and an equivalence class on the clustering attribute. Information theory based concepts of mean gain ratio and entropy of clusters are used to implement these two steps respectively. Experimental results on nine UCI real life and ten synthetic data sets show that our algorithm has lower computational complexity and comparable clustering results.

The rest of the paper is organized as follows. Section 2 briefly reviews the related work. Section 3 describes our algorithm MGR, with an illustrative example. Section 4 analyzes the limitations of MMR [16] algorithm, the most similar work to our method, and compares it with MGR. Section 5 presents experimental results, with a comparison with other algorithms. Finally, Section 6 presents conclusions and recommendations for future work.

2. Related work

Ralambondrainy [1] proposes a method to convert multiple categories attributes into binary attributes using 0 and 1 to represent either a category absence or presence, and to treat the binary attributes as numeric in the k-means algorithm. ROCK algorithm [2] is an adaptation of agglomerative hierarchical clustering algorithm in which the notion of "links" is defined to measure the closeness between clusters. STIRR [3] is an iterative algorithm, which maps categorical data to non-linear dynamic systems. If the dynamic system converges, the categorical data can be clustered. Based on a novel formalization of a cluster for categorical data, a fast summarization based algorithm, CACTUS, is presented in [4]. CACTUS finds clusters in subsets of all attributes and thus performs a subspace clustering of the data.

The k-modes algorithm [5,6] extends the k-means paradigm to categorical domain by using a simple matching dissimilarity measure for categorical objects, i.e., modes instead of means for clusters, and a frequency-based method to update modes. Subsequently, based on k-modes, many algorithms are proposed including adapted mixture model [7], fuzzy k-modes [8], tabu search technique [9], iterative initial points refinement algorithm for k-modes clustering [10], an extension of k-modes algorithm to transactional data [11], fuzzy centroids [12], initialization methods for k-modes and fuzzy k-modes [13,40,14], a dissimilarity measure for k-modes [38], attribute value weighting in k-modes clustering [40], and genetic fuzzy k-modes [15]. k-ANMI [20] is also a k-means like clustering algorithm for categorical data that optimizes the mutual information sharing based objective function.

Besides k-means, classical information theory is another widely used technique in categorical data clustering. COOLCAT [17] explores the connection between clustering and entropy: clusters of similar points have lower entropy than those of dissimilar ones. LIMBO [25] is a hierarchical algorithm that builds on the Information Bottleneck (IB) framework to detect the clustering structure in a data set. "Best K" [26] proposes a BkPlot method to determine the best K number of clusters for a categorical data set.

He et al. [19] formally define the categorical data clustering problem as an optimization problem from the viewpoint of cluster ensemble, and apply cluster ensemble approach for clustering categorical data. Simultaneously, Gionis et al. [27] use disagreement measure based cluster ensemble method to solve the problem of categorical data clustering.

Recently, several works try to solve the problem of categorical data clustering by direct optimization. In algorithms ALG-RAND [18], G-ANMI [21] and the iterative Monte-Carlo procedure in [22], some concepts of information theory, such as generalized conditional entropy, mutual information are used to define the objective function and some optimization methods like Genetics are used to solve the problem. While these algorithms improve clustering accuracy on some data sets, as pointed out in [21], considerable obstacles still remain before they can be widely used in practice. One main obstacle is the efficiency of the optimization algorithms like Genetics.

In addition, He et al. [28] propose TCSOM algorithm for clustering binary data by extending traditional self-organizing map (SOM). The same authors also propose Squeezer algorithm [29]. Squeezer is a threshold based one-pass algorithm which is also