Knowledge-Based Systems 65 (2014) 60–71

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Hierarchical clustering algorithm for categorical data using a probabilistic rough set model

Min Li [a,*], Shaobo Deng [a,c,d], Lei Wang [a], Shengzhong Feng [b], Jianping Fan [b]

[a] Nanchang Institute of Technology, Nanchang, Jiangxi 330099, PR China
[b] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, PR China
[c] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, PR China
[d] Graduate School of Chinese Academy of Sciences, Beijing 100080, PR China

ARTICLE INFO

Article history:
Received 8 April 2013
Received in revised form 4 April 2014
Accepted 5 April 2014
Available online 18 April 2014

Keywords:
Cluster analysis
Categorical data
Probabilistic rough sets
Distribution approximation precision
Approximation accuracy

ABSTRACT

Several clustering analysis techniques for categorical data exist to divide similar objects into groups. Some are able to handle uncertainty in the clustering process, whereas others have stability issues. In this paper, we propose a new technique called TMDP (Total Mean Distribution Precision) for selecting the partitioning attribute based on probabilistic rough set theory. On the basis of this technique, with the concept of granularity, we derive a new clustering algorithm, MTMDP (Maximum Total Mean Distribution Precision), for categorical data. The MTMDP algorithm is a robust clustering algorithm that handles uncertainty in the process of clustering categorical data. We compare the MTMDP algorithm with the MMR (Min–Min–Roughness) algorithm which is the most relevant clustering algorithm, and also compared it with other unstable clustering algorithms, such as *k*-modes, fuzzy *k*-modes and fuzzy centroids. The experimental results indicate that the MTMDP algorithm can be successfully used to analyze grouped categorical data because it produces better clustering results.

1. Introduction

Clustering analysis refers to dividing a given dataset into similar groups according to given rules, and the technique is widely applied in many domains, such as computer vision, biology, medicine, information retrieval, data mining, and pattern recognition. Many clustering algorithms have been proposed to suit different requirements [1].

Clustering categorical data have attracted much attention from the data mining research community. Huang has developed the *k*-modes [2–4] algorithm by extending the standard *k*-means algorithm with a simple matching dissimilarity measure for categorical data. The simple matching dissimilarity measure between two objects is calculated as the number of attributes whose values do not match. The *k*-modes algorithm then replaces the means of clusters with modes, using a frequency-based method to update the modes in the clustering process in order to minimize the clustering cost function. The *k*-modes algorithm preserves the efficiency of the *k*-means algorithm and displays an advantage in interpreting the clustering results. However, the *k*-modes algorithm generates local optimal solutions based on the initial modes and the processing order of objects in the datasets. Moreover, the *k*-modes algorithm assigns each object in a given dataset into only one cluster, and all of the objects have the same degree of confidence when grouped into a cluster [5]. As we know, in reality, the boundary of the data is hard to partition, as there is often no sharp boundary between clusters. Thus, Huang and Ng have introduced the fuzzy *k*-modes algorithm, a generalized version of the *k*-modes algorithm [6], which assigns membership degrees to data in different clusters. However, the clustering result of the fuzzy *k*-modes algorithm is still influenced by the initial values of modes and the processing order of objects in the datasets. Furthermore, fuzzy *k*-modes need to adjust one control parameter of membership fuzziness to obtain better solutions. In the applications, it is not clear how to find out the optimal parameters and their values are often selected on the basis of the decision makers' previous knowledge of the domain and their intuition or the proposed criteria.

Another popular approach for handling uncertainty is rough set theory, which was introduced by Pawlak [7]. It is a type of symbolic machine learning technology for categorical value information systems [8,21]. One reason for the success of rough set theory is that no additional information is required for data analysis, such as thresholds or expert knowledge in a particular domain [9]. In recent years, rough set theory has attracted much attention

* Corresponding author. Tel.: +86 13607002079.
E-mail address: liminghuadi@hotmail.com (M. Li).

http://dx.doi.org/10.1016/j.knosys.2014.04.008
0950-7051/

in some of the clustering literature. For example, Chen and Wang [10] present an improved clustering algorithm on the basis of rough set and Shannon's Entropy theory. Lingras and West [11] introduce a rough k-means clustering algorithm and apply it to the analysis of student web access logs at Saint Mary's University, Canada. Maji and pal [12] describe a clustering algorithm, rough-fuzzy c-medoids, to select the most informative bio-bases. Cao et al. [13] present a framework for clustering categorical time-evolving data based on rough membership function and sliding window technique. On the basis of the idea of biological and genetic taxonomy and rough membership function, Cao et al. [14] propose a new dissimilarity measure for the k-modes algorithm. The above-mentioned algorithms have either convergence flaws or stability ones. Chen et al. present a rough set-based hierarchical clustering algorithm for categorical data [15], but the time complexity of this algorithm is as $O(mn^3)$, where n is the number of objects and m the number of attributes.

A main approach of rough set-based data clustering is that the clustering dataset is mapped as the decision table, and this approach can be performed by introducing a class attribute. Thus, it is of primary importance for this approach to select from many candidates in a database one attribute that can best partition the objects. Mazlack et al. [16] propose two techniques to select the partitioning attribute: Bi-Clustering (BC), based on balanced/unbalanced bi-valued attributes, and Total Roughness (TR), based on the average of the accuracy of approximation precision in the rough set theory. Herawan et al. [17] bring out a technique, Maximum Dependency Attributes (MDAs), for selecting the partitioning attribute by taking into account the dependency of attributes of the dataset. Qin et al. [18] put forward a Novel Soft Set (NSS) approach to select the partitioning attribute. However, all of these papers [16–18] have not presented the concrete clustering algorithm nor have they evaluated the practical effect of their techniques for clustering categorical data.

Parmar et al. [19] propose the (Min–Min–Roughness) MMR algorithm, which is a "purity" rough set-based hierarchical clustering algorithm for categorical data. The MMR algorithm determines the clustering attribute by MR (Min–Roughness) criterion. The main advantages of the MMR algorithm are as follows: (1) it is capable of handling the uncertainty in the clustering process; (2) it is a robust clustering algorithm because it enables the users to obtain stable results by only one input: the number of clusters; (3) it has the capability of handling large datasets.

Top-down hierarchical clustering algorithms are characterized by an irreversible splitting process. This means the splitting strategy and the approach of further determining the clustering node are crucial, which directly affects the final clustering results. Therefore, two key steps of the MMR algorithm lie in searching the partitioning attribute and determining the leaf node to be selected for further clustering. Concretely, the MMR algorithm employs the concept of roughness to search the partitioning attribute and choose the leaf node with more objects for further splitting. However, the roughness cannot reflect the discernibility power to the boundary objects. In addition, the MMR algorithm chooses the leaf node with more objects for further splitting, thus possibly generating undesirable clustering results. These are the two major drawbacks of the MMR algorithm.

Inspired by the MMR algorithm, in this paper we propose a new algorithm for clustering categorical data, called MTMDP (Maximum Total Mean Distribution Precision) algorithm, which is based on probabilistic rough set theory. Except for maintaining all the advantages of MMR algorithm, the MTMDP algorithm has noticeable improvements in two key steps. First, the MTMDP algorithm searches the clustering attribute by taking into account the mean distribution precision of all attributes, which is better than the MR (Min–Roughness) criterion. Second, the MTMDP algorithm determines the further clustering node by considering the cohesion degree of all nodes, which is a more reasonable method compared with the method used in the MMR algorithm. The experimental results on real-life datasets indicate that the MTMDP algorithm can be successfully used in analyzing grouped categorical data because the MTMDP algorithm produces better clustering results.

The structure of the remainder of this paper is as follows. Section 2.1 presents some basic notions related to rough set theory. Section 2.2 reviews the MMR algorithm. Section 3.1 introduces the concept of the probabilistic rough set. Section 3.2 introduces the MTMDP algorithm followed by two examples for illustrative purposes. Section 4 presents our experimental results. Section 5 concludes the paper and identifies future research directions.

2. Related works

In this section, some basic notions are briefly reviewed. In Section 2.1, we provide the basic concepts of rough set theory such as the categorical information system and accuracy of approximation. Then, in Section 2.2, we review the most relevant clustering algorithm, MMR.

2.1. Basic concepts

In general, the structural data can be stored in a table, where each row represents facts about an object. A data table is also called an information system. More formally, a categorical information system (IS) is usually expressed in the following form: $IS = (U, A, V, f_a)_{a\in A}$, where

$U = \{x_1, x_2, \ldots, x_n\}$ is a set of finite and nonempty objects, called the universe;

A is a nonempty finite set of attributes;

V is a set of values$=\cup_{a\in A}V_a$, where V_a is the domain (value set) of the attribute a, and it is finite and unordered;

f is an information function denoted by $f{:}U \times A \to V$, which specifies attribute value of $x \in U$.

With any subset of attributes $B \subseteq A$, there is an indiscernible relation $Ind(B) = \{(x, y) \in U \times U | \forall a \in B, f_a(x) = f_a(y)\}$, where $f_a(x)$ and $f_a(y)$ denote the values of objects x and y under the condition attribute a, respectively. This indiscernible relation $Ind(B)$ divides the universe U into a family of disjoint classes, which are denoted by $U/Ind(B) = \{X_1, X_2, \ldots, X_S\}$, where X_i is an equivalence class induced by $Ind(B)$, $i = 1, 2, \ldots, S$. Obviously, any two objects belonging to the same equivalence class $X_i \in U/Ind(B)$ are indiscernible according to attribute set B.

For an arbitrary set $X \subseteq U$, X can be approximated using only the information that is contained within attribute set B by constructing two unions of elemental sets $\underline{B}(X) = \{x \in U | [x]_B \subseteq X\}$ and $\bar{B}(X) = \{x \in U | [x]_B \cap X \neq \phi\}$, where $\underline{B}(X)$ and $\bar{B}(X)$ are called B-lower and B-upper approximations of X in IS. These definitions state that object $x \in \underline{B}(X)$ certainly belongs to X, whereas object $x \in \bar{B}(X)$ could belong to X. Obviously, there is $\underline{B}(X) \subseteq X \subseteq \bar{B}(X)$. X is said to be definable if $\underline{B}(X) = \bar{B}(X)$. Otherwise, X is said to be rough.

Definition 1. Let X be any subset of U. The accuracy of approximation of X with respect to $B \subseteq A$ is defined as [21]:

$$\alpha_B(X) = \frac{|\underline{B}(X)|}{|\bar{B}(X)|}, \tag{1}$$

where $|\bullet|$ denotes the cardinality of the set. Obviously, $0 \leqslant \alpha_B(X) \leqslant 1$. If $\alpha_B(X) = 1$, then $\underline{B}(X) = \bar{B}(X)$. The B-boundary of X is empty, and X is crisp with respect to B. If $\alpha_B(X) < 1$, then $\underline{B}(X) \subset \bar{B}(X)$. The B-boundary of X is not empty, and X is rough with respect to B.