Knowledge-Based Systems 26 (2012) 120–127

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A dissimilarity measure for the *k*-Modes clustering algorithm

Fuyuan Cao [a], Jiye Liang [a,*], Deyu Li [a], Liang Bai [a], Chuangyin Dang [b]

[a] *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*
[b] *Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong, China*

ARTICLE INFO

Article history:
Received 1 December 2010
Received in revised form 20 July 2011
Accepted 20 July 2011
Available online 27 July 2011

Keywords:
Categorical data clustering
k-Modes algorithm
Rough membership function
Dissimilarity measure
Genetic taxonomy

ABSTRACT

Clustering is one of the most important data mining techniques that partitions data according to some similarity criterion. The problems of clustering categorical data have attracted much attention from the data mining research community recently. As the extension of the *k*-Means algorithm, the *k*-Modes algorithm has been widely applied to categorical data clustering by replacing means with modes. In this paper, the limitations of the simple matching dissimilarity measure and Ng's dissimilarity measure are analyzed using some illustrative examples. Based on the idea of biological and genetic taxonomy and rough membership function, a new dissimilarity measure for the *k*-Modes algorithm is defined. A distinct characteristic of the new dissimilarity measure is to take account of the distribution of attribute values on the whole universe. A convergence study and time complexity of the *k*-Modes algorithm based on new dissimilarity measure indicates that it can be effectively used for large data sets. The results of comparative experiments on synthetic data sets and five real data sets from UCI show the effectiveness of the new dissimilarity measure, especially on data sets with biological and genetic taxonomy information.

1. Introduction

The widespread use of computer and information technology has made extensive data collection in business, manufacturing and medical organizations a routine task. This explosive growth in stored data has generated an urgent need for new techniques that can transform the vast amounts of data into useful knowledge. Data mining is, perhaps, most suitable for this need [1].

In data mining, clustering is a widely used technique that partitions a data set consisting of *n* points embedded in an *m*-dimensional space into *k* distinct clusters such that the data points within the same cluster are more similar to each other than to data points in other clusters. Essentially, clustering is performed according to the similarity or dissimilarity among objects. The similarity or dissimilarity between two objects is generally based on difference in corresponding attribute values. In a clustering algorithm, the similarity or dissimilarity between objects is usually measured by a distance function. The smaller the distance, the more similar the two objects are considered to be. The most commonly used distance function is the Minkowski metric that includes the Euclidean distance and the Manhattan distance as special cases. However, Minkowski metric is only for numeric data, and it becomes difficult to capture this notion for categorical attributes. Therefore, the computation of similarity or dissimilarity between categorical data objects in unsupervised learning is very important.

Roughly speaking, the current approaches to similarity or dissimilarity measures of categorical values can be classified into the following four categories.

1.1. Simple matching approaches

Simple matching is a common approach in which comparison of two identical categorical values yields a difference of *zero* while comparison of two distinct categorical values yields a difference of *one*. The idea of simple matching has been utilized in many categorical clustering algorithms in [2–4] including the *k*-Modes algorithm and its variants, such as the *k*-Modes algorithm [5], fuzzy *k*-Modes algorithm [6], fuzzy *k*-Modes algorithm with fuzzy centroid [7], and *k*-prototype algorithm [8]. However, simple matching often results in clusters with weak intrasimilarity [9], and disregards the similarity hidden between categorical values [10]. A valuable dissimilarity measure is introduced for *k*-Modes clustering algorithm by Ng et al. [9], that extends the standard simple matching approach by taking account of the frequency of mode components in the current cluster.

1.2. Co-occurrence approaches

Gibson et al. [11] pointed out that the similarity of two categorical values refers to their co-occurrence with a common value or a

[*] Corresponding author.
E-mail addresses: cfy@sxu.edu.cn (F. Cao), ljy@sxu.edu.cn (J. Liang), lidy@sxu.edu.cn (D. Li), sxbailiang@126.com (L. Bai), mecdang@cityu.edu.hk (C. Dang).

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.07.011

set of values. Some algorithms based on the idea of co-occurrence of categorical value are proposed. ROCK [12] uses the concept of a *link* to measure the similarity between categorical patterns. A measure $Link(p_i, p_j)$ is defined as the number of common neighbors between two patterns p_i and p_j for ROCK. The objective of the algorithm is to group together patterns that have a relatively large number of links. CACTUS [13] defines the similarity between patterns by looking at the support of two attribute values, which is the frequency of two values appearing in patterns together. The higher the support is, the more similarity the two attribute values are. Based on the co-occurrence probability of two categorical values, a distance metric is presented by Ahamad and Dey [14] for mixed numeric and categorical data clustering. The significance of an attribute towards the clustering process is also hidden in this distance metric.

1.3. Probabilistic approaches

Conceptual clustering algorithms in [15,16] for handling data with categorical values use conditional probability estimate to define relations between groups or clusters. The *category utility* (CU) measure [17] defines a probability matching strategy to measure the usefulness of a class in correctly predicting feature values, and the idea is also adopted in the system COBWEB [18] and its derivatives [19,20]. AUTOCLASS [21] assumes a classical finite mixture distribution model on the data and uses Bayesian method to derive the most probable class distribution for the data with some prior information. Wong et al. proposed a discrete-valued data clustering algorithm DECA [22], which has been used in bimolecular data clustering [23]. Chiu et al. [24] proposed a distance measure for dealing with mixed-type attributes in large database. Their techniques are derived from a probabilistic model in which the distance between two clusters is equivalent to the decrease in the log-likelihood function as for merging. An entropy-based categorical data clustering algorithm COOLCAT [25] finds a set of initial clusters, and then incrementally adds patterns to the clusters according to the criterion that minimizes the expected entropy of the clusters.

1.4. Distance hierarchy approaches

Distance hierarchy [10,26–28] extends the concept of hierarchy [29] by associating each link with a weight representing a distance to facilitate the computation of distance between categorical values. However, such an approach needs domain experts to incorporate knowledge, e.g., the general-specific relationship, for facilitating further mining of clustering results.

Other similarity or dissimilarity measures for categorical data clustering algorithms include Gower's similarity coefficient [30], Goodall's similarity measure [31,32], and Gowda's dissimilarity measure [33].

Rough set theory introduced by Pawlak [34] is a kind of symbolic machine learning technology for categorical value information systems with uncertainty information [35–37]. In recent years, rough set theory has received a great deal of attention in some of the clustering literature. Parmar et al. [38] proposed a new algorithm min-min-roughness (MMR) for clustering categorical data based on rough set theory, which has the ability to handle the uncertainty in the clustering process. By defining outlying partition similarity based on the concept of rough set, outliers on the key attribute subset rather than on the full dimensional attribute set of data set can be mined [39]. Using the notion of rough membership function from rough set theory, Jiang et al. [40] defined the rough outlier factor for outlier detection. Chen and Wang [41] presented an improved clustering algorithm based on rough set and Shannon's Entropy theory. Herawan et al. [42] proposed a new technique called maximum dependency attributes for selecting clustering attribute based on rough set theory by taking into account the dependency of attributes of the database. Cao et al. [43] proposed a framework for clustering categorical time-evolving data based on rough membership function and sliding window technique.

In this paper, the limitations of simple matching dissimilarity measure and Ng's dissimilarity measure are revealed using some illustrative examples. Based on the idea of biological and genetic taxonomy, we introduce a new rough membership-based dissimilarity measure between two objects by taking into account the distribution of attribute values in the universe. Furthermore, the dissimilarity measure between a mode of a cluster and an object is given by improving Ng's dissimilarity measure. The proposed dissimilarity measure is utilized in the *k*-Modes algorithm, the algorithm convergence is proved and the corresponding time complexity is analyzed as well. The scalability and clustering effectiveness of the *k*-Modes algorithm with the proposed dissimilarity measure are demonstrated on synthetic data sets and five standard data sets downloaded from the UCI Machine Learning Repository [44], respectively.

The organization of the rest of this paper is as follows. In Section 2, two kinds of new dissimilarity measures, between two objects and between a mode and an object, for the *k*-Modes algorithm are defined. Convergence and time complexity of the *k*-Modes algorithm with the proposed measure are analyzed in Section 3. In Section 4, experimental results on the synthetic data sets and five real data sets demonstrate the scalability and effectiveness of the *k*-Modes algorithm based on the new dissimilarity measure by comparison with other dissimilarity measures. Section 5 concludes the paper.

2. New dissimilarity measures for the *k*-Modes algorithm

In this section, we first review some basic concepts of rough set theory, such as categorical information system, indiscernibility relation and rough membership function. Then, a new dissimilarity measure between two objects is defined based on rough membership function. Furthermore, a new dissimilarity measure between the mode of a cluster and an object is introduced for the *k*-Modes algorithm.

The data is assumed to be in a table, where each row (tuple) represents facts about an object. A data table is also called an information system. Objects in the real world are sometimes described by categorical information system.

Definition 1. Formally, a categorical information system is a quadruple $IS = (U, A, V, f)$, where:

U, the nonempty set of objects, called the universe;
A, the nonempty set of attributes;
V, the union of all attribute domains, i.e., $V = \bigcup_{a \in A} V_a$, where V_a is the domain of attribute a and it is finite and unordered;
f: $U \times A \rightarrow V$, a mapping called an information function such that for any $x \in U$ and $a \in A$, $f(x, a) \in V_a$.

Definition 2. Let $IS = (U, A, V, f)$ be a categorical information system and $P \subseteq A$, a binary relation $IND(P)$, called indiscernibility relation, is defined as:

$$IND(P) = \{(x, y) \in U \times U |\ \forall\ a \in P,\ f(x, a) = f(y, a)\}.$$

Informally two objects are indiscernible in the context of a set of attributes if they have the same values for those attributes. $IND(P)$ is an equivalence relation on U and $IND(P) = \bigcap_{a \in P} IND(\{a\})$.