Available online at www.sciencedirect.com
ScienceDirect
Data & Knowledge Engineering 63 (2007) 879–893
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

MMR: An algorithm for clustering categorical data using Rough Set Theory

Darshit Parmar, Teresa Wu *, Jennifer Blackhurst

Department of Industrial Engineering, PO Box 875906, Arizona State University, Tempe, AZ 85287-5906, USA

Received 5 August 2006; received in revised form 18 April 2007; accepted 29 May 2007
Available online 13 June 2007

Abstract

A variety of cluster analysis techniques exist to group objects having similar characteristics. However, the implementation of many of these techniques is challenging due to the fact that much of the data contained in today's databases is categorical in nature. While there have been recent advances in algorithms for clustering categorical data, some are unable to handle uncertainty in the clustering process while others have stability issues. This research proposes a new algorithm for clustering categorical data, termed Min–Min-Roughness (MMR), based on Rough Set Theory (RST), which has the ability to handle the uncertainty in the clustering process.
Published by Elsevier B.V.

Keywords: Cluster analysis; Categorical data; Rough Set Theory; Data mining

1. Introduction

Cluster analysis is a data analysis tool used to group data with similar characteristics. It has been used in data mining tasks such as unsupervised classification and data summation, as well as segmentation of large heterogeneous data sets into smaller homogeneous subsets that can be easily managed, separately modeled and analyzed [12]. The basic objective in cluster analysis is to discover natural groupings of objects [14]. Cluster analysis techniques have been used in many areas such as manufacturing, medicine, nuclear science, radar scanning and research and development planning. For example, Jiang et al. [13] analyze a variety of cluster techniques for complex gene expression data. Wu et al. [40] develop a clustering algorithm specifically designed to handle the complexities of gene data that can estimate the correct number of clusters and find them. Wong et al. [39] present an approach used to segment tissues in a nuclear medical imaging method known as positron emission tomography (PET). Mathieu and Gibson [26] use cluster analysis as a part of a decision support tool for large-scale research and development planning to identify programs to participate in and to determine resource allocation. Finally, Haimov et al. [8] use cluster analysis to segment radar signals in scanning land and marine objects.

* Corresponding author.
E-mail address: teresa.wu@asu.edu (T. Wu).

0169-023X/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.datak.2007.05.005

A problem with many of the clustering methods and applications mentioned above is that they are applicable for clustering data having numerical values for attributes. Most of the work in clustering is focused on attributes with numerical value due to the fact that it is relatively easy to define similarities from the geometric position of the numerical data. Unlike numerical data, categorical data have multi-valued attributes. Thus, similarity can be defined as common objects, common values for the attributes, and the association between the two. In such cases, the horizontal co-occurrences (common attributes for the objects) as well as the vertical co-occurrences (common values for the attributes) can be examined [40].

A number of algorithms for clustering categorical data have been proposed including work by Huang [12], Gibson et al. [5], Guha et al. [6], Ganti et al. [4], and Dempster et al. [2]. While these methods make important contributions to the issue of clustering categorical data, they are not designed to handle uncertainty in the clustering process. This is an important issue in many real world applications where there is often no sharp boundary between clusters. Recently, there has been work in the area of applying fuzzy sets in clustering categorical data including work by Huang [12] and Kim et al. [16]. However, these algorithms require multiple runs to establish the stability needed to obtain a satisfactory value for one parameter used to control the membership fuzziness.

Therefore, there is a need for a robust clustering algorithm that can handle uncertainty in the process of clustering categorical data. This research proposes a clustering algorithm based on Rough Set Theory (RST). The proposed algorithm, named Min–Min-Roughness (MMR), is designed to deal with uncertainty in the process of clustering categorical data. In addition, the algorithm is implemented and tested with three real world data sets. To compare the algorithm performance in handling uncertainty, Soybean and Zoo data sets are used and the results are compared with fuzzy set theory based algorithms (including *K*-modes, fuzzy *K*-modes and fuzzy centroids). To test the applicability to large scale date sets, the Mushroom data set is used and the results are compared with Squeezer, *K*-modes and LCBCDC, as well as ROCK and a traditional hierarchical algorithm. The contributions of our proposed approach include:

(1) Unlike previous methods, MMR gives the user the ability to handle uncertainty in the clustering process.
(2) Using MMR, the user is able to obtain stable results given only one input: the number of clusters.
(3) MMR has the capability of handling large data sets.

This paper is structured as follows: Section 2 presents an overview of standard clustering methods existing in the literature. In Section 3, the basics of the rough set theory are introduced followed by the proposed MMR algorithm. A synthetic data set is used to illustrate the MMR algorithm. Section 4 discusses the implementation of the algorithm and the results from the application of the algorithm on Soybean, Zoo and Mushroom data sets (from the UCI Machine Learning Repository[1]). In addition, the comparison results are analyzed. Section 5 presents conclusions and identifies future research directions.

2. Literature review

In this section, an overview of methods available in the literature to cluster categorical data is presented. Ralambondrainy [33] proposes a method to convert multiple category 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. Dempster et al. [2] presents a partitional clustering method, called the Expectation-Maximization (EM) algorithm. EM first randomly assigns different probabilities to each class or category, for each cluster. These probabilities are then successively adjusted to maximize the likelihood of the data given the specified number of clusters. Since the EM algorithm computes the classification probabilities, each observation belongs to each cluster with a certain probability. The actual assignment of observations to a cluster is determined based on the largest classification probability. After a large number of iterations, EM terminates at a locally optimal solution. Han et al. [9] propose a clustering algorithm to cluster related items in a market database based on an association rule hypergraph. A hypergraph is used as a model for relatedness. The

[1] Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.