Chapter 14

Cluster Analysis with Various Algorithms for Mixed Data

Abha Sharma
Maulana Azad National Institute of Technology, India

R. S. Thakur
Maulana Azad National Institute of Technology, India

ABSTRACT

Analyzing clustering of mixed data set is a complex problem. Very useful clustering algorithms like k-means, fuzzy c-means, hierarchical methods etc. developed to extract hidden groups from numeric data. In this paper, the mixed data is converted into pure numeric with a conversion method, the various algorithm of numeric data has been applied on various well known mixed datasets, to exploit the inherent structure of the mixed data. Experimental results shows how smoothly the mixed data is giving better results on universally applicable clustering algorithms for numeric data.

INTRODUCTION

The fast growth in real world data has generated need for new techniques that can convert huge datasets into useful information and data mining is one which has a power to do this transformation. Information retrieval, medical diagnosis, financial fraud, image processing, bioinformatics are various applications of data mining. According to Han et al. (2001) proposes clustering is one of the important unsupervised learning which is the process of grouping a set of objects such that the objects in the same groups are similar but are very dissimilar with objects in other clusters.

Many clustering algorithms have been developed for numeric, categorical and mixed data, special attention has been paid to mixed data analysis, where data objects are neither pure categorical, nor pure numeric and very useful for real world datasets from statistics to psychology. According to MacQueen (1967) k-means clustering is a very fast and simple method for clustering high-dimensional numerical data.

Sharma et al. (2014) discuss in their paper that the absence of inherent ordering of data the clustering of mixed data has became a challenge along with the high dimensionality of mixed data. So it is not a simple and straight forward to develop clustering technique of high quality as well as the functions, models, algorithms, methods which applied on the numeric datasets are not applicable to the mixed datasets.

DOI: 10.4018/978-1-5225-0536-5.ch014
Cluster Analysis with Various Algorithms for Mixed Data

Hsu et al. (2008) discusses that usually real world datasets are mixed contain both numeric and categorical attributes. However, most existing clustering algorithms assume all attributes are either numeric or categorical, examples of which includes the k-means, Huang (1997) proposed k-modes, again Huang et al. (1999) discovered fuzzy k-modes, He et al. (2012) proposed the TGCA algorithm, Barbara et al. (2002) proposed COOLCAT, and Deng et al. (2010) proposed G-ANMI algorithms. When mixed data are encountered, most of them usually exploit transformation approaches to convert one type of the attributes to the other and then apply traditional single-type clustering algorithms. However, in most cases, transformation scheme may result in loss of information, leading to undesired clustering outcomes.

Uptill now there has been some research carried out for directly dealing with mixed data. Cheessman and Stutz (1996) proposed the approach AutoClass for mixed data, which exploited mixture model and Bayesian method to deal with mixed data. Li et al. (2002) presented the Similarity-Based Agglomerative Clustering (SBAC), which is a hierarchical agglomerative algorithm. The SBAC algorithm adopts the similarity measure defined by Goodall et al. (1966) to evaluate similarities among data objects. Hsu et al. (2006) Clustering Algorithm based on the Variance and Entropy (CAVE) for clustering mixed data. The CAVE algorithm needs to build a distance hierarchy for every categorical attribute, and the determination of the distance hierarchy requires domain expertise. Hsu et al. (2011) extended self-organizing map to analysis mixed data. In their method, the distance hierarchy can be automatically constructed by using the values of class attributes. Chatzis (2011) proposed the KL-FCM-GM algorithm, which is based on the assumption that data deriving from clusters are in the Gaussian form and designed for the Gauss Multinomial distributed data. David et al. (2012) presented SpectralCAT to deal with mixed data SpectralCAT system, the numeric values were transformed to the categorical ones. Huang (1997) presented a k-prototypes algorithm, which integrated the k-means with k-modes methods to partition mixed data. Jinchao et al. (2012) proposed the fuzzy k-prototypes which took into account the fuzzy nature of data objects.

The sections of this chapter are as follows: Background, Experimental analysis, Conclusion, Future work, and References.

BACKGROUND

The k-means algorithm which is the base of all the clustering algorithms and implemented in various areas. But the algorithm has major drawback of not handling the type of data other than numeric. Huang (1997) proposed k-prototype algorithm which combines the k-modes and the k-means algorithm using two types of distance measures Euclidian distance and match mismatch measure respectively, therefore it can handle mixed numeric and categorical datasets.

One more big challenge for clustering pure categorical data Sharma et al. (2012, 2016) or mixed data is to preprocess the mixed data or categorical data into pure numeric such as in binary or any other formats. In this work the transformation of mixed data into pure numeric data has been done using first step of TMCM algorithm Ming-Yi et al. (2010) to use the power of existing pure numeric algorithm such as k-means, single linkage, fuzzy c means, Stephen (1994) proposed subtractive with fuzzy C-means all algorithms is made for numeric data. According to Yinghua et al. (2013) to convert mixed data into binary formats is simple and fast to process but lack of resemblance a similarity information of categorical values, therefore slightly not relevant.
Related Content

Challenges in Clinical Data Linkage in Australia: Perspective of Spinal Cord Injury
[www.igi-global.com/article/challenges-in-clinical-data-linkage-in-australia/171402?camid=4v1a](www.igi-global.com/article/challenges-in-clinical-data-linkage-in-australia/171402?camid=4v1a)

[www.igi-global.com/article/a-multi-objective-ensemble-method-for-class-imbalance-learning/197439?camid=4v1a](www.igi-global.com/article/a-multi-objective-ensemble-method-for-class-imbalance-learning/197439?camid=4v1a)

Study and Analysis of Delay Factors of Delhi Metro Using Data Sciences and Social Media: Automatic Delay Prediction System for Delhi Metro

Conceptual View on Healthcare Digitalization: An Extended Thematic Analysis
[www.igi-global.com/article/conceptual-view-on-healthcare-digitalization/197440?camid=4v1a](www.igi-global.com/article/conceptual-view-on-healthcare-digitalization/197440?camid=4v1a)