Abstract

Clustering categorical data is relatively difficult than clustering numeric data. In numeric data the inherent geometric properties can be used in defining distance functions between data points. In case of categorical data, a distance or dissimilarity function can't be defined directly. An extension of the classical k-means algorithm for categorical data has been done in [1], where a method of representing a cluster using representatives which are very much similar to means used in k-means algorithm has been proposed together with a new distance measure. In this paper we first propose an alternative representation of categorical data as numeric data making it easier to handle. This technique provides a uniform representation for data points and the cluster representatives. The similarity measure proposed in [2] has been used in this new setting. The algorithm used in [1] has been implemented and tested with this new setting and the results obtained have been reported. Experiments were conducted on two real life data sets, namely, soybean diseases, and mushroom data sets. The clusters obtained in soybean dataset are pure clusters with hundred percent accuracy. In the other dataset also it gives relatively higher accuracy with small errors.
Categorical Data Clustering based on an Alternative Data Representation Technique

- Zengyou He, Xiaofei Xu, Shengchun Deng. A Cluster Ensemble Method for Clustering Categorical Data

Index Terms
Computer Science
Data Mining

Keywords
clustering categorical data cluster representative