Journal of Information Technology Research
Volume 11 • Issue 2 • April-June 2018

A Support Based Initialization Algorithm for Categorical Data Clustering

Ajay Kumar, Jaypee University of Engineering & Technology, Guna, India
Shishir Kumar, Jaypee University of Engineering & Technology, Guna, India

ABSTRACT

Several initial center selection algorithms are proposed in the literature for numerical data, but the values of the categorical data are unordered so, these methods are not applicable to a categorical data set. This article investigates the initial center selection process for the categorical data and after that present a new support based initial center selection algorithm. The proposed algorithm measures the weight of unique data points of an attribute with the help of support and then integrates these weights along the rows, to get the support of every row. Further, a data object having the largest support is chosen as an initial center followed by finding other centers that are at the greatest distance from the initially selected center. The quality of the proposed algorithm is compared with the random initial center selection method, Cao's method, Wu method and the method introduced by Khan and Ahmad. Experimental analysis on real data sets shows the effectiveness of the proposed algorithm.

KEYWORDS

Accuracy, Clustering, K-modes, Precision, Recall, Support

1. INTRODUCTION

Clustering is an unsupervised learning technique used to find the group in the data sets. The goal of clustering is to minimize an objective function so that the variance between the data within the group is minimum and maximum from other data points outside the group. Clustering algorithms are used in a variety of scientific areas, including biology (Liang & Wang, 2007), image processing (Shanthi & Bhaskaran,2013), GIS application (Vesanto & Alhoniemi, 2000), business intelligence (Rajagopalan & Krovi, 2002), social networks (Naser & Alshattnawi, 2014), anomaly detection (Moshtaghi et al., 2011) and security (Thuraisingham, 2005).

Many clustering approaches are documented throughout in the literature such as hierarchical based (Jain, 2010), partitional, density-based (Amineh Amini, 2014), model-based and grid-based (Kaufman & Rousseeeuw, 1990). A partitional algorithm like k-means is popular for the clustering large data sets, due to its simplicity, computational speed and ease of implementation. The initial seed value of the k-means algorithm is important since each seed can produce different local optima leading to varying partitions (Arthur & Vassilvitskii, 2007). Good initialization is therefore, critical for finding globally optimal partitions. The k-means algorithm is not able to handle categorical data sets due to which it is hard to find a distance and means between categorical data objects. Several methods are presented in the literature for the categorical data clustering that are based on k-means clustering.

The k-modes (Huang, 1998) algorithm extends the work of k-means algorithm to cater categorical data sets. The main process of these two algorithms is to specify the number of clusters and then select the k initial center, followed by assigning every object to the nearest initial center. The clustering

DOI: 10.4018/JITR.2018040104

results and convergence speed these algorithms depend on the initial centers, so the selection of an initial center is an important issue in both the algorithm. A good initial center can lead to the fast convergence and valid results while poor initial centers lead to the slow convergence (Gan, Ma, & Wu, 2007). Random initialization method is simple and easy to use. However, this may lead to a non-repeatable clustering result. In many applications several scans of the k-mode or k-means algorithm are required to get a meaningful result.

In the paper, the author's proposed a support based initial center selection algorithm for the categorical data set. In the proposed approach, an object is selected from the data set as an initial cluster center, which has the high support value among all the data objects. Other k-1 centers are selected from the data set, which are at maximum distance from the initially selected center and at a greatest distance from each other's. These initially selected centers are then passed to the k-mode algorithm to find the group in the data set. The comparative analysis of the proposed method illustrates the effectiveness of this approach.

The rest of this paper is organized as follows. A short survey on the cluster center initialization for the categorical data is presented in Section 2. Sections 3 discuss the k-modes algorithm. In Section 4, the proposed support based initialization scheme for the k-modes algorithm is presented. Section 5 shows the complete validation and comparative analysis of the proposed method on various categorical data sets with other initialization methods. Finally, Section 6 concludes the work presented in this paper with a discussion.

2. RELATED WORK

To find the natural groups in a categorical data set, various methods are documented in the literature. Guha et al. introduced the ROCK algorithm and used the Jaccard coefficient to compute the distances between the objects (Guha, 2000). The ROCK algorithm clusters objects in an agglomerative way such that the number of links within a cluster can be maximized. The k-mode algorithm on other hands is one of the popular methods (Bai, Liang, Dang & Cao, 2012). It extends the work of k-means algorithm to cluster categorical data.

Clustering result of a k-mode algorithm depends on the initial center; so as to pick out the initial center is an important issue in the k-mode algorithm. The quality of clustering results can be enhanced by selecting initial center values that are much closer to the final solution (Jain & Dubes, 1988). Many initial center selection algorithms are reported in the literature. A commonly used approach called the direct method is used to choose the first k distinct objects as an initial mode. Another approach, called the diverse modes method, it spread the initial modes over whole data set by assigning the most frequent categories to the initial modes (Huang, 1998). These initial modes can lead to a better clustering result, but their time and cost are high.

Barbara, Couto and Li (2002) introduced a max-min distance method to find the k-most dissimilar data objects to be used as an initial cluster centers. However, due to distance as a measure, outliers may be selected as an initial center. Khan and Ahmad (2003) documented a density-based method to find the k initial modes that are used as an initial cluster center from the data set and their time complexity is quadratic with respect to the number of data objects.

Wu, Jiang and Huang (2007) developed a density based method to compute the k initial cluster center but due to the random choice repeatability of the clustering results is not achieved. Cao et al. (2012) presented an algorithm to find the initial modes for the k-mode algorithm. A joint approach of density and distance measure is used to find the initial modes. They assumed that an object from the data set could not represent true clusters hence; they used the information of the neighbors to find the initial modes for the k-mode algorithm. The method is effective and having linear time complexity with respect to the number of data objects.

Khan and Ahmad (2013) developed an algorithm to partition the categorical data into the clusters that correspond to the number of distinct attribute values for Vanilla/ Prominent/ Significant attributes

Diffusion of Innovations in Organisations

Davood Askarany (2005). *Encyclopedia of Information Science and Technology, First Edition (pp. 853-857).*

www.igi-global.com/chapter/diffusion-innovations-organisations/14348?camid=4v1a

IT Outsourcing Practices in Australia and Taiwan

Chad Lin and Koong Lin (2009). *Encyclopedia of Information Science and Technology, Second Edition (pp. 2291-2297).*

www.igi-global.com/chapter/outsourcing-practices-australia-taiwan/13901?camid=4v1a