Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 93 – 98

Procedia Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

The k-modes algorithm with entropy based similarity coefficient

Ravi Sankar Sangam*, Hari Om

Department of Computer Science & Engineering, Indian School of Mines, Dhanbad and 826004, India

Abstract

Clustering is the process of organizing dataset into isolated groups such that data points in the same are more similar and data points of different groups are more dissimilar. The k-modes algorithm well known for its simplicity is a popular partitioning algorithm for clustering categorical data. In this paper, we discuss the limitations of distance function used in this algorithm with an illustrative example and then we propose a similarity coefficient based on Information Entropy. We analyze the time complexity of the k-modes algorithm with proposed similarity coefficient. The main advantage of this coefficient is that it improves the clustering accuracy while retaining scalability of the k-modes algorithm. We perform the scalability tests on synthetic datasets.

Keywords: data mining; clustering; k-modes; data analysis; unsupervized learing

1. Introduction

Due to the advances in scientific data collection methods, the size and variety of data have tremendous growth. Therefore it is practically impossible to discover the useful information from such raw data by using the traditional database analysis techniques. The data mining techniques are very essential for extracting the useful information from the very large databases. Clustering is an unsupervised classification data mining technique that divides the dataset into homogenous groups based on some resemblance criterion [1].

Many previous works [2-6] focus on only clustering the numerical data. Clustering the categorical data is a very complex task since there exists no natural relative order between the categorical values. The hierarchical clustering algorithms can be used to cluster the categorical data, but their quadratic time complexity hinders its usage. The k-

* Corresponding author.
E-mail address: srskar@gmail.com

doi:10.1016/j.procs.2015.04.066

modes algorithm proposed by Haung [7] is a popular categorical clustering algorithm. It follows the k-means [8] algorithm paradigm and divides the dataset into k number of groups based on the simple matching distance metric. The simple matching distance metric is however not a good measure as it results in poor intra-cluster similarity. Ng et al. [9] have proposed a new dissimilarity coefficient for the k-modes algorithm in which the frequency of the categorical values in current cluster has been considered to calculate the dissimilarity between a data point and a cluster mode. Cao et al. [10] have discussed the k-modes algorithm with rough set based dissimilarity coefficient for biological datasets. The ROCK [11] algorithm measures the similarity between the categorical patterns using the concept of links i.e., the similarity between any two categorical patterns depends on the number of their common neighbors. The aim of this algorithm is to merge the patterns into a group that have relatively large number of links. Unlike the ROCK algorithm, the CATCUS [12] algorithm measures the similarity between the categorical vectors in terms of the support of two categorical attribute values. Here the support of two categorical attributes values is the frequency of these two values present in patterns where the higher support value represents the maximal resemblance. An efficient algorithm, called squeezer, is discussed for clustering categorical data [13]. This algorithm sequentially reads the data points one by one and then it determines either current data point corresponds to a cluster or creates a new cluster based on the similarity coefficient defined in this algorithm. Among afore discussed algorithms, the k-modes algorithm is very efficient due its linear time complexity. As a result, in this paper we propose a new similarity coefficient based on information entropy [14] for the k-modes algorithm that improves the clustering result accuracy while retaining its scalability.

The rest of the paper is organized as follows. In section 2, various notations and the k-modes algorithm are presented. Section 3 presents our proposed scheme and section 4 evaluates its scalability. Finally we draw the conclusion in section 5.

2. Notations and k-modes algorithm

For sake of notations consider that S is a dataset i.e., $S = \{x_1, x_2, \ldots, x_N\}$, where each data point $x_i, 1 \leq i \leq N$, is described by n categorical attributes i.e., $x_i = \{d_1^i, d_2^i, \ldots, d_n^i\}$. The objective of the k-modes algorithm is to divide the dataset S into k isolated groups, called clusters, by minimizing the cost function given below.

$$Cost(W, M) = \sum_{l=1}^{k}\sum_{i=1}^{N} w_{i,l} dis(x_i , m_l) \tag{1}$$

$$\text{subject to } \; 0 \leq w_{i,l} \leq 1, for \;\; 1 \leq i \leq N, \;\; 1 \leq l \leq k \tag{2}$$

where W is an $k \times N$ partition matrix, $m_l \in M$ is a mode vector of categorical attribute set of lth cluster, that can be obtained from the most frequent categorical value of each attribute domain and $dis(\,)$ is a simple matching distance between ith data point and lth cluster, that is defined as follows:

$$dis(x_i, m_l) = \sum_{j=1}^{n} \delta(x_i^j, m_l^j) \tag{3}$$

$$\text{here, } \; \delta(x_i^j, m_l^j) = \begin{cases} 0, & \text{if } x_i^j = m_l^j \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

In the next section we discuss the drawback of simple matching distance metric (3) with the help of an artificial dataset given in Table 1 and then we present our new similarity coefficient.

3. Proposed method

Consider the artificial dataset as shown in Table 1. The dataset has been partitioned into three clusters c_1, c_2 and c_3 with their respective modes m_1, m_2 and m_3. We see that each data point in the dataset has been described by three categorical attributes. Suppose we want to assign the data point $x_{10}[\mathrm{A}, \mathrm{X}, \mathrm{R}]$ to any one of these clusters. The distance metric (3) as given below.

$dis(x_{10}, m_1) = 0 + 0 + 1 = 1, dis(x_{10}, m_2) = 1 + 0 + 0 = 1$ and $dis(x_{10}, m_3) = 1 + 0 + 1 = 2$.

According to these values we can assign data point x_{10} to either cluster c_1 or c_2, i.e., we cannot determine appropriate cluster for that data point. However, it is more appropriate to assign data point x_{10} to cluster c_2 since it maximizes the inter-cluster criterion.

Table 1. Artificial dataset

Cluster c_1			Cluster c_2			Cluster c_3		
d_1	d_2	d_3	d_1	d_2	d_3	d_1	d_2	d_3
A	X	P	B	X	R	A	X	W
A	X	Q	C	X	R	D	X	P
B	X	P	B	X	R	D	X	P
$m_1[\mathrm{A}, \mathrm{X}, \mathrm{P}]$			$m_2[\mathrm{B}, \mathrm{X}, \mathrm{R}]$			$m_3[\mathrm{D}, \mathrm{X}, \mathrm{P}]$		

As a result based on information entropy [14], we propose a new similarity coefficient for k-modes algorithm. Let Y be a random variable with possible states $y_1, y_2, \dots, y_k$ and $P(y_l)$ is the probability of lth state of Y. The entropy of Y, denoted by $En(Y)$, is given as follows:

$$En(Y) = -\sum_{l=1}^{k} P(y_l) \log(P(y_l)) \tag{5}$$

If Y is uniformly distributed, then the entropy (5) returns maximum value i.e., maximum uncertainty or minimum information. We can see from the Fig. 1 that the categorical value 'X' of d_2 attribute has uniform distribution and hence it minimizes the distance between two clusters. In contrast to this, the categorical value 'R' of d_3 attribute has minimal distribution and accordingly it minimizes the intra-cluster distance and maximizes the distance between cluster c_2 and other two clusters. Based on the foregoing discussion, we define a new similarity coefficient between a data point and a cluster mode as follows:

$$sim(x_i, m_l) = \sum_{j=1}^{n} \phi(x_i^j, m_l^j) \tag{6}$$

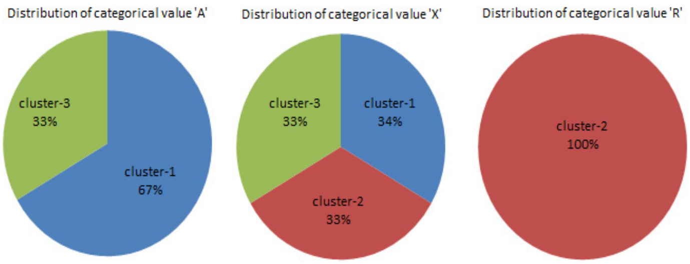

Fig.1. Distribution of categorical values between the clusters

$$\text{here, } \phi\left(x_i^j, m_l^j\right) = \begin{cases} 1 - \dfrac{1}{\log(k)} \times En\ (m_l^j), \text{ if } x_i^j = m_l^j \\ 0, \text{ otherwise} \end{cases} \tag{7}$$

We see that the similarity coefficient (6) has defined on the scale [0, n], where the similarity values 0 and n indicate minimum and maximum similarity, respectively. Now, the similarity between the data point x_{10} and the cluster c_1, denoted by $sim(x_{10}, m_1)$ is calculated as follows:

$$sim(x_{10}, m_1) = 1 - \left(\frac{1}{\log(3)} \times - \left(\frac{2}{3}\log\left(\frac{2}{3}\right) + \frac{1}{3}\log\left(\frac{1}{3}\right)\right)\right) + 1 - \left(\frac{1}{\log(3)} \times - \left(\frac{3}{9}\log\left(\frac{3}{9}\right) + \frac{3}{9}\log\left(\frac{3}{9}\right) + \frac{3}{9}\log\left(\frac{3}{9}\right)\right)\right) + 0 = 0.421$$

Table 2. k-modes algorithm with new similarity coefficient

```
Data: Dataset S, #of Clusters k
Result: Dataset S has been divided into k non overlapping clusters
Initialize the variable old_modes as an $k \times m$ empty array
Choose randomly k different data points from dataset S as initial modes and assign $[m_1, m_2, \dots, m_k]$ to k × m array variable new_modes
for i=1 to N do
    for l=1 to k do
        Calculate the similarity between ith data point and lth mode vector using similarity coefficient (6) and assign that data point to
        appropriate cluster whose cluster mode vector is closer to it and update mode vector of corresponding cluster and also find the
        distribution of mode categories between clusters using Equation (5) ;
    end;
end;
while old_modes ≠ new_modes do
    old_modes = new_modes;
    for i=1 to N do
        for l=1 to k do
            Calculate the similarity between ith data point and lth mode vector using similarity coefficient (6) and assign that data
            point to appropriate cluster whose cluster mode vector is closer to it and update mode vectors of corresponding two clusters
            and also find the distribution of mode categories between clusters using Equation (5) ;
        end;
    end;
    if old_modes = new_modes then
        break;
    endif;
end;
```

The similarity between the data point x_{10} and other two clusters can be calculated similarly, i.e., $sim(x_{10}, m_2) = 1$ and $sim(x_{10}, m_1) = 0$. Based on the calculated values the data point x_{10} has maximal similarity on the cluster c_2 and hence it is assigned to that cluster. The k-modes algorithm with the proposed similarity coefficient is presented in Table 2. The time complexity of the k-modes algorithm with new similarity coefficient is calculated as follows. To find the similarity between a data point and a cluster mode the time complexity is $O(n)$ and for all the clusters it is $O(nk)$. Therefore, the computational time complexity of initial allocation of data points to different clusters is $O(nkN)$. To update the mode vectors of all the cluster it takes $O(nN)$. Suppose it takes total I number of iterations to break the while loop (see pseudo code) then the total time complexity of k-modes algorithm is $O(InkN)$. We can see that the computational time complexity of k-modes algorithm is linear with respect to the size of dataset.

4. Results

In this section,we evaluate the scalability of our proposed scheme. We have implemented our method in C++ language and executed on Intel i7 processor (3.40 GHz) with 2GB memory running on windows 7 operating system. We have compared the execution time of our proposed method with the original k-modes algorithm. First, we have evaluated the scalability with respect to the dataset size as shown in Fig. 2. It may be noted that in this scalability test the dataset size has been varied from 10000 records to 50000 records and we have been fixed the dataset dimensionality to 15 i.e., $n = 15$ and number of clusters to 3 i.e., $k = 3$. From the Fig. 2, we can see that both the techniques have consumed almost same amount of time and have almost linear performance with respect to dataset size.

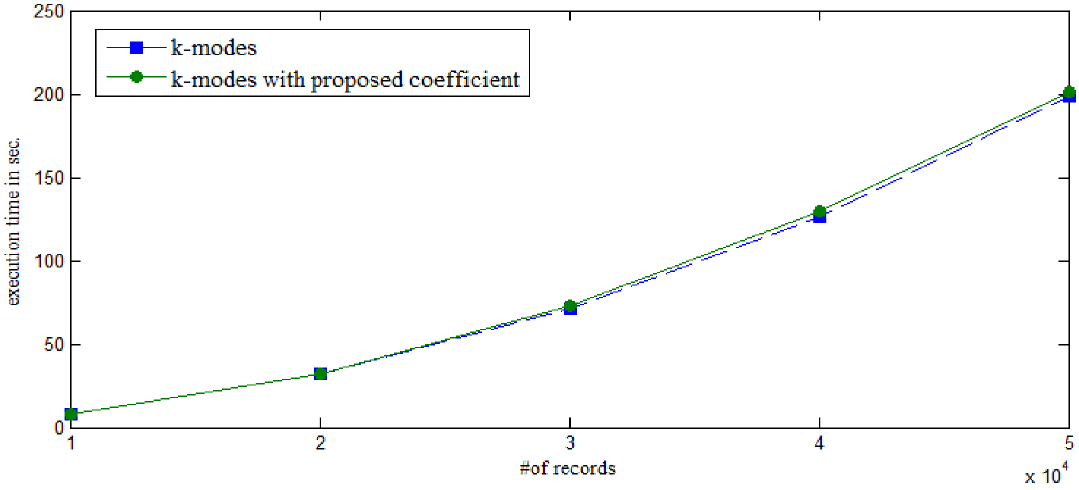

Fig. 2. Scalability of k-modes algorithm with respect to dataset size

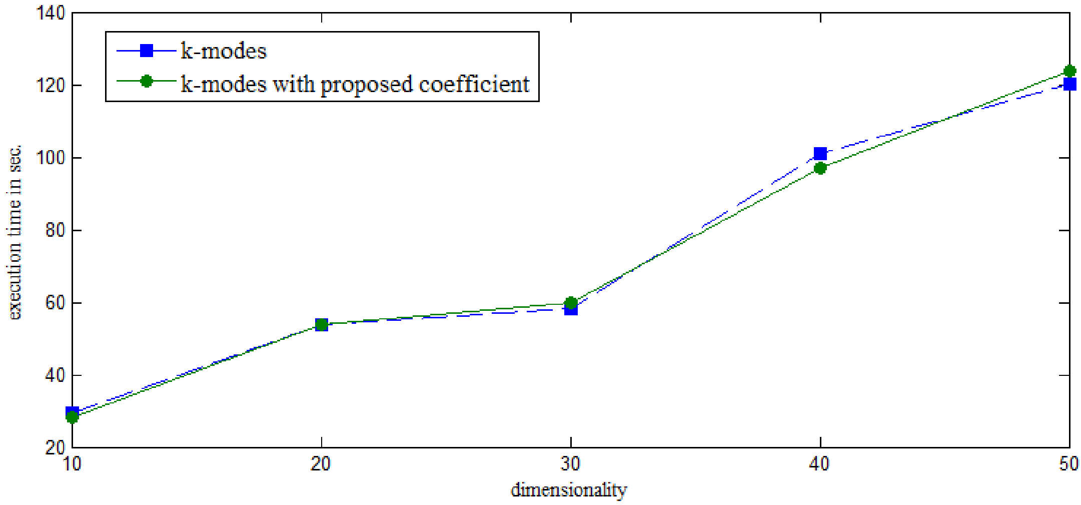

Fig. 3. Scalability of k-modes algorithm with respect to dataset size

In second scalability test, we have fixed the dataset size to 20000 records, number of clusters as 3, i.e., $k = 3$ and dataset dimensionality has been varied from 10 to 50. From Fig. 3, we see that both the techniques consume almost same amount of time and have the linear performance with respect to dimensionality. Based on the above results we conclude that our similarity coefficient maintains the scalability of the original k-modes algorithm. However, the illustrative example discussed in previous section shows that our scheme gives better cluster accuracy results since it considers clustering criterion i.e., both high intra-cluster similarity and low inter-cluster similarity.

5. Conclusion

Clustering categorical data is a complex task since there is no natural order among the categorical values. The k-modes algorithm is a popular clustering algorithm in this regard since it is linearly scalable with respect to the dataset size. In this paper we have discussed the limitations of simple matching distance metric used in this algorithm with an illustrative example and proposed a new similarity coefficient based on the information entropy. The time complexity analysis of our scheme and the experimental results on synthetic datasets shows that it is linearly scalable with dataset size.

References

1. Han J, Kamber M, Pei J. *Data Mining: Concepts and Techniques*. 3rd ed. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA;2011.
2. Karaboga, Dervis, Celal Ozturk. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Applied Soft Computing* 2011; 11(1): 652-657.
3. Zhang X, Jiaqi L, Yu D, Tingjie Lv. A novel clustering method on time series data. *Expert Systems with Applications* 2011; 38(9): 11891-11900.
4. Galluccio, Laurent, Olivier M, Pierre C, Mark K, Alfred O. Clustering with a new distance measure based on a dual-rooted tree. *Information Sciences* 2013;251: 96-113.
5. Yu S, Tranchevent L-C, Liu X, Glanzel W, Suykens JAK, De Moor B, Moreau Y. Optimized data fusion for kernel k-means clustering. *Pattern Analysis and Machine Intelligence*, *IEEE Transactions on* 2012; 34(5): 1031-1039.
6. Xie J, Shuai J, Weixin X, Xinbo G. An efficient global K-means clustering algorithm. *Journal of computers* 2012; 6(2): 271-279.
7. Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery* 1998; 2(3): 283-304.
8. MacQueen, J. et al. Some methods for classification and analysis of multivariate observations. *In Proceedings of the 5th berkeley symposium on mathematical statistics and probability*, California, 1967; p. 281–297.
9. Ng M.K.,Li M.J, Huang J.H, He Z. On the impact of dissimilarity measure in k-modes clustering algorithm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 2007; 29(3): 503-507.
10. Cao F, Liang J, Li D, Bai L, Dang C. A dissimilarity measure for the k--Modes clustering algorithm. *Knowledge-Based Systems* 2012, 26: 120-127.
11. Guha S, Rastogi R, Shim K. Rock: a robust clustering algorithm for categorical attributes, *in: Proceedings of the IEEE International Conference on Data Engineering*, Sydney, Australia, 1999, p. 512–521.
12. Ganti V, Gehrke J,Ramakrishnan R. Cactus-clustering categorical data using summaries, *in: Proceedings of the International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, 1999, p. 73–84.
13. Zengyou H, Xu X, and Deng S. Squeezer: an efficient algorithm for clustering categorical data. *Journal of Computer Science and Technology* 2002; 17(5): 611-624.
14. Shannon, Claude E. Communication Theory of Secrecy Systems. *Bell system technical* journal 1949; 28(4): 656-715.