A Modified K-Modes Clustering Algorithm

Conference Paper · December 2013

DOI: 10.1007/978-3-642-45062-4_7

CITATION
1

READS
457

2 authors, including:

Vandana Bhattacharjee
Birla Institute of Technology, Mesra
82 PUBLICATIONS **424** CITATIONS

SEE PROFILE

A Modified *K*-Modes Clustering Algorithm

Partha Sarathi Bishnu and Vandana Bhattacherjee

Deptartment of Computer Science & Enggineering, Birla Institute of Technology, Ranchi, Jharkhand, 834001, India
psbishnu@gmail.com, bhattacherjeev@yahoo.co.in

Abstract. The main aim of this paper is to apply a new dissimilarity measure, and handle boundary data properly in K-Modes clustering thereby increasing the clustering efficiency. Moreover our proposed algorithm identifies the outlier data efficiently.

Keywords: Data mining, clustering, K-Modes algorithm, outlier.

1 Introduction

To overcome the drawback of partitional clustering techniques like K-Means, K-Medoid algorithms which are popular clustering algorithms for numeric only datasets, Huang, [1] suggested K-Modes clustering algorithm to handle categorical data. Software cost estimation, bioinformatics, and computer networks are some of the areas of application for the K-Modes clustering algorithm [4]. There are some issues which directly affect the quality of the clusters using K-Modes algorithm: first, the initial selection of the cluster centers [1], [2], [6], second, the similarity or dissimilarity measures between the two non numeric data [3], [4], [5], third, proper handling of the boundary data [7] and finally outlier detection. In this paper our main objectives are threefold; first, we introduce a new dissimilarity metric, second, we introduce a new technique to handle boundary data properly and third, identification of the outlier data to handle the second, third, and fourth issue respectively. We have compared our technique with three existing techniques suggested by Huang, 1998 (K-Modes 1)[1], Huang and Ng, 1999, (K-Modes 2) [7] and He et al., 2005 (K-Modes 3) [3] with help of real datasets from UCI machine learning repository.

The outline of this paper is as follows: in section two we describe our proposed dissimilarity measure, and a method to handle boundary data and outliers. In section three we explain the experiments which we carried out. In section four we present the results and this is followed by conclusions in section five.

2 Proposed *K*-Modes Clustering Algorithm

2.1 *K*-Modes Clustering Algorithm

First we discuss the simple K-Modes clustering algorithm for the sake of completeness [1], [4]. A categorical dataset (O) is defined by a set (D) of attributes

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 60–66, 2013.

$I_1, I_2, ..., I_d$where d is the dimension of the dataset. Each attribute I_e is described by a set of categorical values $DOM(I_e)$ which are finite and unordered. For $a, b \in DOM(I_e)$, either $a = b$ or $a \neq b$. A data $o_i \in O$, where $i = 1, 2, ...n$ and n is the number of data, can be represented as a conjunction of attribute value pairs $[I_1 = a_1] \wedge [I_2 = a_2] \wedge ... \wedge [I_d = a_d]$, where $a_e \in DOM(I_e)$. We use ε to represent the missing value. Moreover, if $o_i, o_j \in O$, and let $o_i = [a_1, a_2, ..., a_d]$ and $o_j = [b_1, b_2, ..., b_d]$ then we write $o_i = o_j$ if $a_e = b_e$, for $1 \leq e \leq d$ and the relation $o_i = o_j$ does not mean that they are the same data, but rather that the two data have equal values on attributes (I_e), where $1 \leq e \leq d$ [4].

Let $o_i, o_j \in O$, be two categorical data represented by $[a_1, a_2, ..., a_d]$ and $[b_1, b_2, ..., b_d]$, respectively. The distance (simple matching dissimilarity) between o_i and o_j is defined as $dis(o_i, o_j) = \sum_{e=1}^{d} \delta(a_e, b_e)$ where

$$\delta(a_e, b_e) = \begin{cases} 0 & : \quad a_e = b_e \\ 1 & : \quad a_e \neq b_e \end{cases} \tag{1}$$

In K-Modes the objective of clustering categorical dataset into $K(<< n)$ clusters is to search W and M that minimize the cost function $F(W, M)$, where $F(W, M) = \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ki} dis(m_k, o_i)$ (2) , subject to $w_{ki} \in \{0, 1\}$, $1 \leq k \leq K$ and $1 \leq i \leq n$ (3), $\sum_{k=1}^{K} w_{ki} = 1$, $1 \leq i \leq n$ (4) and , $0 < \sum_{i=1}^{n} w_{ki} < n$, $1 \leq k \leq K$ (5), where $W = [w_{ki}]$ is a k-by-n $\{0, 1\}$ matrix and $m_k \in M$, $1 \leq k \leq K$, is the k^{th} cluster mode.

2.2 New Dissimilarity Measure

We introduce a new dissimilarity measure ($dist$) for the K-Modes clustering algorithm so as to minimize the cost function $F_{BB}(W, M)$, where $F_{BB}(W, M) = \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} dist(m_k, o_i)$ subject to conditions as in (3), (4), and (5).

Definition 1. *Let $o_i, o_j \in O$ be two categorical data represented by $[a_1, a_2, ..., a_d]$ and $[b_1, b_2, ..., b_d]$, respectively. Let $dist(o_i, o_j)$ be the distance between two objects. The distance is defined as follows: $dist(o_i, o_j) = 1 - \frac{\alpha_{ij}}{\beta_{ij}}$, where $\alpha_{ij} = \sum_{e=1}^{d} \chi(a_e, b_e)$, where,*

$$\chi(a_e, b_e) = \begin{cases} 1 & : \quad a_e = b_e \\ 0 & : \quad a_e \neq b_e \end{cases} \tag{6}$$

and, $\beta_{ij} = 2d - \alpha_{ij}$, where d is the dimension of the data. Here, $\frac{\alpha_{ij}}{\beta_{ij}}$ measures the similarity between two objects as the degree of relative overlap between the two objects. α_{ij} counts the number of attributes for which the values in o_i, o_j match while β_{ij} counts those which do not.

Theorem 1. *Let M be fixed and consider the problem: $min_w F_{BB}(W, M)$, subject to (3) to (5). The minimizer M is given by*

$$w_{ki} = \begin{cases} 1, & : \quad dist(m_k, o_i) \leq dist(m_h, o_i), 1 \leq h \leq K, \\ 0, & : \quad otherwise. \end{cases}$$

Theorem 2. *Let* $m_k = [m_{k1}, m_{k2}, ..., m_{kd}]$ *be the mode of the* k^{th} *cluster and the domain* ζ_{a_j} *of attributes* a_j *be* $\{a_j^1, a_j^2, ..., a_j^{n_j}\}$, $1 \leq j \leq d$ *and* n_j *is the cardinality of domain of* a_j. *Denote arbitrary object* o_i *by* $[a_{i1}, a_{i2}, ..., a_{id}]$. *Then* $F_{BB}(W, M) = \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ki} dist(m_k, o_i)$ *is minimized if and only if* $m_{kj} = a_j^r$, *where,* $a_j^r \in \zeta_{a_j}$ *satisfies:* $|\{w_{ki} | o_{ij} = a_j^r, w_{ki} = 1\}| \geq |\{w_{ki} | o_{ij} = a_j^t, w_{ki} = 1\}|$, $1 \leq t \leq n_j$, $1 \leq j \leq d$.

Proof. For a given W, $F_{BB}(W, M) = \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ki} dis(m_k, o_i) = \sum_{k=1}^{K} \psi_k$. Note that all the inner sums ψ_k of $F_{BB}(W, M)$ are non negative and independent. Then minimizing $F_{BB}(W, M)$ is equivalent to minimizing each inner sum. By definition 1, when $m_{kj} = a_j^t$ we have $\psi_k = \sum_{i=1}^{n} w_{ki} dist(m_k, o_i) = \sum_{i=1}^{n} w_{ki}(1 - \frac{\alpha_{ki}}{\beta_{ki}}) = N_k - \sum_{i=1}^{n} \frac{\alpha_{ki}}{\beta_{ki}})$, where N_k is the number of data in k^{th} cluster. Since N_k is nonnegative, ψ_k is minimized if $\frac{\alpha_{ki}}{\beta_{ki}}$ is maximized for $1 \leq i \leq n$. Since $\frac{\alpha_{ki}}{\beta_{ki}}$ is the similarity (or degree of relative overlap) between mode m_k and object o_i, it is maximized when $t = r$ i.e., $m_{kj} = a_j^r$. The result follows. □

2.3 Proper Handling of Boundary Objects

The data o_i should be assigned to cluster number k_q if $dis(o_i, m_k) < dis(o_i, m_l)$, $1 \leq l, q \leq K$ and $l \neq q$. Now if $dis(o_i, m_k) = dis(o_i, m_l)$, then o_i is termed as boundary data [7]. It is conventional to assign o_i to cluster number k_q or k_l, which ever comes first. However, in this paper we suggest a novel approach to assign proper cluster number for a boundary data. To handle the proper assignment of boundary data we use one of the criteria, these being the average distance ($AvgDist$), total distance ($Dist$), and size (of q^{th} cluster, c_q) (n_q). The distances are computed as $AvgDist_{iq} = \frac{Dist_{iq}}{n_q}$, where $Dist_{iq} = \sum_{j=1}^{n_q} dist(o_i, o_{qj})$, $o_{qj} \in c_q$. The boundary data a_i should be assigned to q^{th} cluster if $AvgDist_{iq} < AvgDist_{il}$, $1 \leq l, q \leq K$, $l \neq q$. Further, in a rare situation when $AvgDist_{iq} = AvgDist_{il}$, data o_i should be assigned to q^{th} cluster if $Dist_{iq} > Dist_{il}$ or if $n_q > n_l$. In the rarest of situation where all these values happen to coincide and the data is not detected as outlier then we can assign it randomly to any of q or l.

2.4 Outlier Detection

An outlier is a data which is far away (very different) from the rest of the data.

Definition 2. *Let* $o_i = [a_{i1}, a_{i2}, ..., a_{id}]$ *and* $o_j = [b_{j1}, b_{j2}, ..., b_{jd}]$ *be two data objects, where* $1 \leq j \leq n, i \neq j, o_i, o_j \in O$. *Then* o_i *is outlier data if* $a_{ie} \neq b_{je}$, $1 \leq j \leq n, 1 \leq e \leq d$, $i \neq j$.

Since none of the attributes of an outlier data o_i matches with any attribute of other data, neither simple technique nor boundary data assignment technique will help to assign cluster number and data will be floating among clusters. So we assign that data point o_i as outlier data and remove it from the dataset.

Definition 3. *Let* $o_i = [a_{i1}, a_{i2}, ..., a_{id}]$ *then* o_i *is outlier data if* $AvgDist_{ij} = 1$, *for all* $1 \leq j \leq K$.

Example 1. Let cluster 1 consist of three data $[a, a, b; a, a, c; a, a, d]$, where mode of cluster 1 is $m_1 = [a, a, b]$ and cluster 2 consists of four data $[a, a, e; a, c, b; a, c, b; a, a, f]$ where mode of cluster 2 is $m_2 = [a, a, b]$ (assume modes are same by chance). Consider another data $o_p = [a, c, d]$, which is not part of any cluster and is a boundary data because of the same distance $dist(o_p, m_1) = dist(o_p, m_2) = 1 - \frac{1}{2*3-1} = 0.8$ from the mode 1 and mode 2. To assign the cluster number we calculate the $AvgDist_{p1}(0.7)$, and $AvgDist_{p2}(0.65)$. Here, $AvgDist_{p2} < AvgDist_{p1}$, so data $[a, c, d]$ should be assigned to second cluster. Furthermore the data $o_i = [x, y, z]$ (initially denoted as border data) is the outlier data because no attribute of o_i is matching any attribute of with any other data and $AvgDist_{i1} = AvgDist_{i2} = 1$.

2.5 Proposed K-Modes Clustering Algorithm

Next we give a formal algorithm to explain our proposed algorithm (Algorithm 1) as follows:

```
Algorithm 1: Proposed K-Modes Algorithm
Input: O, K
Output: C (clusters)
Step1: randomly select initial modes from given dataset O;
Step2: (re) assign cluster number to each data by calculating
the distance (dissimilarity) between the data and the modes;
Step3: identify and label boundary data;
Step4: handle each boundary data separately to assign the cluster
number and identify the outlier data and remove the outlier data
from the dataset;(from 2^nd iteration outlier data will not appear)
Step5: (re)identify the modes;
Step6: repeat steps 2 to 5 till convergence criteria is satisfied;
```

The time complexity of the step 1 is $O(1)$ and steps 2 and 5 is $O(ntKd)$, where t is the number of iteration. The time complexity to handle boundary data and outlier data is $O(xntd)$, where x ($x \ll n$) is the number of boundary data and $O(ynd)$where y is the number of outlier data and it is identified in first iteration only i.e. when $t = 1$ (step 4) respectively.

3 Experiments

3.1 Experimental Analysis

To evaluate our proposed algorithm we have used four real datasets from UCI machine learning repository (http://archive.ics.uci.edu/ml/), namely Lung cancer data (LD), Breast cancer data (BD), Zoo data (ZD), and Soybean data (SD).

The number of data are 32, 699, 101, and 47 respectively. The dimensions are 57, 11, 18, 35 respectively. We eliminate attributes or data with any missing values. No real datasets consist of any outlier data, so we have used two synthetic datasets (syn1 and syn2) consisting of 100 and 500 data respectively and the dimensions are 10 and 15 respectively. We randomly add five outlier data to each synthetic dataset. More over to show the scalability of our proposed algorithm we execute all the algorithms on different dimensions and sizes (Figure 1a and Figure 1b). To achieve our first objective of introducing a new dissimilarity measure for categorical data, we execute all but step 3 and step 4 of our proposed algorithm (algorithm 1), and name this as K-Modes 4 i.e., we are not handling boundary and outlier data in K-Modes 4. Finally, we execute all our steps with new dissimilarity measure as well as handling of boundary and outlier data and name it as K-Modes 5 in Table 1. For all the data sets LD, BD, ZD and SD the number of clusters are as per the number of classes present i.e., 3, 2, 7, and 4 respectively. All the algorithms were executed 100 times and the best result of accuracy was reported. All the algorithms iterate 30 times (we set this value as convergence criteria). Efficiency is expressed in terms of execution time. To evaluate the efficiency of the K-Modes clustering algorithms we use three evaluation parameters namely, $accuracy = \frac{\sum_{i=1}^{K} m_i}{n}$, $precision = \frac{\sum_{i=1}^{K} \frac{m_i}{n_i}}{K}$, and $recall = \frac{\sum_{i=1}^{K} \frac{m_i}{m_i+p_i}}{K}$. Where, m_i is the number of data that are correctly assigned to i^{th} cluster, n_i is the total number of data present in the i^{th} cluster, p_i is the data incorrectly rejected from the i^{th} cluster [4], [7]. The experiments were conducted on a PC with an Intel Celeron processor, (1.30 GHz), and 256 MB RAM running the Windows XP operating system. All the K-Modes algorithms (K-Modes 1 [1], K-Modes 2 [7], K-Modes 3 [3], K-Modes 4 and K-Modes 5 (proposed)) have been coded in Octave -3.2.4.

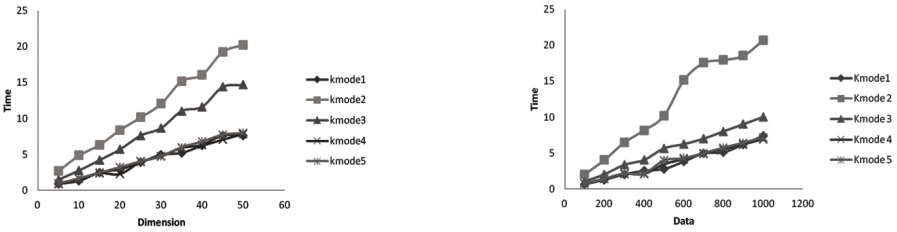
Table 1. Experiments on Real Datasets

Dataset	Evaluation	K-Modes 1	K-Modes 2	K-Modes 3	K-Modes 4	K-Modes 5
Lung Cancer	Accuracy	0.9687	0.9687	0.9687	0.9687	1
	Precision	0.9666	0.9666	0.9761	0.9761	1
	Recall	0.9743	0.9743	0.9629	0.9629	1
	Time	3.5001	8.9062	8.4531	3.5469	4.4219
Breast Cancer	Accuracy	0.9399	0.9799	0.8446	0.9399	0.9985
	Precision	0.9580	0.9725	0.9038	0.9580	0.9989
	Recall	0.9128	0.9847	0.7738	0.9128	0.9765
	Time	8.6875	19.047	15.344	8.8901	19.3570
Zoo	Accuracy	0.8514	0.8811	0.8514	0.9108	0.9306
	Precision	0.8464	0.8623	0.8397	0.9191	0.9285
	Recall	0.9644	0.8507	0.8642	1	0.9642
	Time	7.3125	35.344	10.469	8.1406	19.500
Soybean	Accuracy	1	1	1	1	1
	Precision	1	1	1	1	1
	Recall	1	1	1	1	1
	Time	4.5938	13.0611	11.0942	3.7188	4.9219

3.2 Results

We have evaluated the clustering efficiency of our algorithms with other K-Modes algorithms. The performance of K-Modes 5 algorithm in terms of accuracy and

precision is the best for all datasets and recall is the second best only for zoo data. The performance of K-Modes 4 is second best to K-Modes 5 and better than all the other algorithms.To evaluate the scalability performance of our proposed algorithm we compare execution time with other existing algorithms. For breast cancer and zoo data the execution time of K-Modes 5 is high but it is reasonably low for the other two datasets. From the plot (Figure 1a and Figure 1b) it is seen that the times for our algorithm K-Modes 4 and K-Modes 5 fall in the lowest category (using synthetic datasets). It is also seen that the algorithms are scalable in terms of dimension and size of the data. Moreover our proposed algorithm (K-Modes 5) detected correctly all the outlier data from the synthetic datasets syn1 and syn2, and hence our algorithm shows the potentiality to identify the outlier data.



(a) Dimension versus Time (b) Data versus Time

Fig. 1. Scalability Performance

4 Conclusions

In this paper we have suggested a new dissimilarity measure and techniques for handling border data and outlier data to increase the K-Modes clustering quality. The results show the effectiveness of our proposed K-Modes algorithm by comparing with other popular existing K-Modes algorithms upon real as well as synthetic datasets.

References

1. Huang, Z.: Extensions to the K-Means algorithm for clustering large data sets with categorical values. Data Mining Knowledge Discovery 2(3), 283–304 (1998)
2. Cao, F., Liang, J., Bai, L.: A new initialization method for categorical data clustering. Expert Systems with Applications 36, 10223–10228 (2009)
3. He, Z., Deng, S., Xu, X.-F.: Improving K-Modes algorithm considering the frequencies of attribute values in mode. In: Hao, Y., Liu, J., Wang, Y.-P., Cheung, Y.-M., Yin, H., Jiao, L., Ma, J., Jiao, Y.-C. (eds.) CIS 2005. LNCS (LNAI), vol. 3801, pp. 157–162. Springer, Heidelberg (2005)
4. Ng, M.K., Li, M.J., Huang, J.Z., He, Z.: On the impact of dissimilarity measure in K-Modes clustering algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(3), 503–507 (2007)

5. Cao, F., Liang, J., Li, D., Bai, L., Dang, C.: A dissimilarity measure for the K-Modes clustering algorithm. Knowledge-Based Systems 26, 120–127 (2012)
6. Bai, L., Liang, J., Dang, C.: An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. Knowledge-Based System 24, 785–795 (2011)
7. Huang, Z.X., Ng, M.K.: A fuzzy k-modes algorithm for clustering categorical data. IEEE Transactions on Fuzzy Systems 7(4), 446–452 (1999)