Journal of King Saud University – Computer and Information Sciences xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Initial seed selection for K-modes clustering – A distance and density based approach

S.A. Sajidha *, Siddha Prabhu Chodnekar, Kalyani Desikan

School of Computing Science and Engineering, VIT, Chennai, India

ARTICLE INFO

Article history:
Received 9 January 2018
Revised 30 April 2018
Accepted 30 April 2018
Available online xxxx

Index Terms:
K modes
Clustering
Classification
Initial seed artefact
Density
Distance

ABSTRACT

Initial seed artefacts play a vital role in proper categorization of the given data set in partitioning based clustering algorithms. Hence, it is important to identify them. We propose a density with distance based method which ensures identification of seed artefacts from different clusters that leads to more accurate clustering results. Our algorithm improves on the search for initial seed artefacts iteratively until the minimum value of the sum of within sum errors, normalized by their data sizes, is ensured. This is because the initial artefacts are selected from different clusters. Here the choice of seed artefacts guarantees a global optimum clustering solution. We have compared our results with random, Wu, Cao and Khan's methods of initial seed artefact selection, to show the efficacy of our method.

1. Introduction

Many research issues revolve around clustering and classification of data artefacts. This is because researchers and industries are interested in identifying hidden patterns across data objects for improving their business. Classification involves training the model and testing of the data artefacts to be classified. Clustering has a greater significance as clustering algorithms learn the nature of the data artefacts and categorize them based on the similarity among the data artefacts. Clustering and classification may not generate the same results as clustering studies the underlying differences in the data artefacts while classification generates a strict model to classify the data.

Categorical data clustering plays a major role in various demanding fields such as businesses (finance, shopping carts etc.) in finding similar customers based on the transactions they have made, epidemiological studies to obtain similar patterns of health and disease conditions in a defined population, to identify corresponding policy decisions and targets for preventive healthcare and social sciences for measuring attitudes and opinions etc.

Clustering of categorical data is more challenging because each data artefact is described by its descriptive properties and each property belongs to a particular domain which is non-numerical in nature and is not ordered in any way. One of the most popular partitioning algorithms for categorical data is K-modes algorithm. This algorithm iteratively updates the cluster to which each data artefact is assigned, based on some distance measure until convergence.

K-means clustering algorithm is suitable for numeric data and Ralambondrainy (1995) extended K-means algorithm to cluster categorical data. The presence or absence of each value across categorical attributes was denoted as 1 or 0 and they were treated as numeric entities in the K-means algorithm. The distance measure used here is same as that for numerical data and this may not give expected results for categorical data.

Huang (1997) presented the K-modes clustering algorithm by introducing a new dissimilarity measure to cluster categorical data. The algorithm replaces means of clusters with modes (most frequent attribute value of an attribute). It uses a frequency based method to update the modes in the clustering process to minimize the cost function which is estimated by computing the normalised sum of within sum errors. This algorithm is experimentally shown to achieve convergence with linear time complexity with respect to the number of data artefacts. Huang (Huang, 1998) also points

* Corresponding author.
E-mail addresses: sajidha.sa@vit.ac.in (S.A. Sajidha), kalyanidesikan@vit.ac.in (K. Desikan).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2018.04.013
1319-1578/

Please cite this article in press as: Sajidha, S.A., et al. Initial seed selection for K-modes clustering – A distance and density based approach. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.04.013

out that in general, the K-modes algorithm is faster than the K-means algorithm because it takes fewer iterations to converge.

Definition 1:. A data artefact $(DA_p, p = 1 \ldots n)$ may be defined with a set of /m/ descriptive attributes A as (DA_p, A_q) where $p = 1 \ldots n$ and $q = 1 \ldots m$.

Definition 2:. Let DA be a set of data artefacts, which are categorized into K sets $D_1, D_2, \ldots, D_K$ where K is the number of clusters and let their associated mode (center) vectors generated by the final clusters be denoted as $\text{ModeVec}(l)_q$, where $l = 1 \ldots K$ and $q = 1 \ldots m$.

The similarity function $\text{sim}(g, h)$ between two categorical data artefacts, g and h may be defined as

$$\text{sim}(g,h) := \sum_{A_q \in A} \text{dist}(g,h,A_q)$$
$$\text{where} \begin{cases} dist(g,h,A_q) := 1 & \text{if } \text{Val}(g,A_q) \neq \text{Val}(h,A_q) \\ dist(g,h,A_q) := 0 & \text{otherwise} \end{cases} \quad (1)$$

This distance measure is a simple matching measure between any two data artefacts, which we apply in our paper.

The following is the algorithm for K-modes clustering:

Algorithm 1. K-modes Algorithm

Input: A collection of data artefacts DA with 'm' attributes, to be partitioned into K groups.
Output: Data partitioned into K groups.
1. Randomly initialize the seed artefacts seed_1 where $l = 1 \ldots K$
2. Assign each data artefact DA_p to the closest seed artefact group G_l.
   $G_l = \{DA_p : \min(\text{sim}(DA_p, \text{seed } l))\}$,
   $p = 1 \ldots n$ and $l = 1 \ldots K$
   where $\text{sim}(DA_p, \text{seed}_l)$ is as given in Eq. (1)
3. Compute the mode vector of the currently grouped G_l data artefacts.
   For $l = 1 \ldots K$
   For $q = 1 \ldots m$
   $\text{ModeVec}(l)_q = \{a | a \in A_q \text{ and } \text{freq}(a, A_q) \text{is maximum}\}$
   for all artefacts in G_l
4. Now replace seed_l with the newly computed $\text{ModeVec}(l)$, where $l = 1 \ldots K$
   $\{\text{seed}_l := \text{ModeVec}(l)\}$
5. Repeat steps 2-4 until convergence i.e., until ModeVecdoes not change.
6. Estimate the normalized sum of within sum error (NSWSE) of the obtained clustering solution
   $$\text{NSWSE} = \frac{\sum_{l=1}^{K} \sum_{g_x \in G_l}^{\text{sim}} (\text{seed}_l, g_x)}{|DA|}$$
   where $\text{sim}(\text{seed}_l, g_x)$ is computed as per Eq. (1)

Huang (1998) performed two types of scalability tests on K-modes algorithm. One in terms of the number of clusters for a given number of data artefacts and the other in terms of the number of data artefacts for a given number of clusters for a data set having 500,000 records, which are described by 34 categorical attributes out of which 4 attributes have more than 1000 categories each. He has proved that his algorithm works efficiently in terms of both the number of data artefacts and the number of clusters.

K-modes though known for its simplicity and performance still has certain drawbacks. In principle, K-modes clustering algorithm functions similar to K-means clustering algorithm and hence, suffers from the same drawbacks. It fails to determine the number of clusters, converge to global optimum due to random selection of initial seed artefacts, identify clustering tendency, handle empty clusters, identify outliers etc.

One of the major issues with K-modes algorithm is that it selects the initial seed artefacts randomly. This may lead to selection of artefacts which may fall in the same region and may, hence, not be diverse. We understand that more diverse (selected from different clusters) the initial data artefacts are, the better the clustering results. Random initialization leads to variation in the normalized sum of within sum error of the clustering solutions. Hence, our goal is to find the initial seed artefacts which provide the global minimum of the normalized sum of within sum error of the clustering solution which is the best clustering solution that can be obtained. We propose a novel method to ensure that the initial seed artefacts are more diverse and consequently, generate better clustering results.

In addition to the seed artefacts selection problem, in order to have an effective clustering solution the data sets are supposed to possess a certain pattern. If the patterns are not discernible, we may obtain some clustering solution but it may not lead to an effective inference. Hence, to validate and ensure the correctness of the clustering algorithms most of the researchers have simulated data sets (Milligan and Isaac, 1980).The advantage of this approach is that the structures of constructed data sets can be controlled. Once this is ensured the algorithm is tested for its performance with real data sets. This is one of the most common evaluation methods. We have also followed a similar approach to ensure the correctness of our algorithm.

The paper is organized as follows. In Section 2 we discuss some background work related to our proposed technique. In Section 3 we present our proposed technique for initial seed artefact selection. The experimental results and discussions are presented in Section 4. The conclusion of the paper is presented in Section 5.

2. Related work

Sun et al. (2002) apply Bradly and Fayyad's iterative initial-point refinement algorithm to K-modes clustering. They have used a three step approach where K sets of sub-samples are chosen first depending on the data set size, then each of them are clustered using K-modes with random seeds, all the clustering solutions (cluster modes) obtained for each sub-sample set are brought together into a single set. In the second step the refinement happens, where the combined modes are considered as the data set and are clustered using each of the previous cluster modes as the representative seed artefacts. Distortion between the refined modes and the representative modes is computed and the set with the least distortion is selected as the initial seed data artefacts. Here the sub-samples are randomly selected to represent the actual data which may not generate replicable clustering solution.

Wu et al. (2007) propose a new initialization method which has three phases. The first phase is the sub-sample initialisation phase where K sets of sub-samples are randomly selected from the entire data set where the number of sub-samples is equal to square root of the total number of points. Repeat the following for each of the sub-samples: The density of each data artefact of the sub-sample is estimated. The artefact which is in the most dense region is chosen as the first representative seed artefact. The remaining K-1 (where K is the number of clusters) representative seed artefacts are selected based on the maximum probability of it being selected by computing the product of density of each point and its

Please cite this article in press as: Sajidha, S.A., et al. Initial seed selection for K-modes clustering – A distance and density based approach. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.04.013

minimum distance from the previously selected representative seed artefacts. The value of density of each artefact is boosted by a parameter $\frac{1}{\max(\text{density of all points})}$ and the value of distance is boosted by $\frac{1}{\max(\min(\text{dist}(\text{data artefact,previously selected seed artefacts})))}$. These K representative seed artefacts are stored in a set - $\mathbf{S_1}$. Similarly, $\mathbf{S_2}$ to $\mathbf{S_K}$ sets of representative artefacts are got. The second phase is the refinement phase where all the sets of the representative seed artefacts are considered as a data set and are clustered using each of the representative artefact sets. The third phase is the evaluation phase where the cost of each of the clustering solutions got from the second phase is computed. The representative seed artefact set having the least error of clustering solution is chosen as the initial seed artefacts. Since this method randomly selects the sub-samples, it does not ensure a repeatable clustering solution with global optimum.

Cao et al. (2009) compute the density of each data artefact, and the first seed artefact is the data artefact with the maximum average density. The second seed artefact is selected by computing the product of distance of each artefact from the first seed and its corresponding density. The artefact corresponding to the maximum value of the product is chosen as the second seed artefact. The remaining seed artefacts are obtained by computing, the product of minimum distance of each data artefact from the previously selected seed artefacts and its corresponding density and the point with maximum value is selected as the next seed artefact. This process is iteratively performed until the desired number of initial seeds is obtained. Here both the distance and density are given equal weightage.

Bai et al. (2011) find the first seed artefact as the point with maximum density. The remaining seed artefacts are found by splitting the data artefacts into sets based on the unique value of each attribute and finding the potential exemplars based on the most frequent values in each set for each attribute. They compute the possibility of the potential exemplar becoming eligible seed artefacts based on the density, and distance from the previously selected seed artefacts. Here more weightage is given to density and lesser weightage is given to distance. This paper does not recommend how to set the weights for the distance and density, so that efficient searching for initial seed artefacts from different clusters of the clustering solution can be made. The best values of the weights for which the ideal initial seed artefacts are obtained, is also not indicated.

Bai et al. (2012) select an artefact called exemplar, based on both density of the data artefact and its distance from the mode of the data set; it is selected such that it is farthest from the mode of the data set. This is the first exemplar. An exemplar is a representative data artefact and not an actual seed artefact. The eligible candidate set from the first exemplar is obtained as follows. The artefacts are grouped based on their distances from the first exemplar as $S_1 \ldots S_{|A|}$. For each artefact in S_i, compute the frequency of each of the attribute value. Assign S_1 with the computed frequency values to a new set $Q_j,\ j = 1$. The mode of each attribute of the artefacts in Q1 is computed. This forms the first candidate of a candidate set. Remaining candidates are found by performing union of all S_a where $a \subseteq b$ and $b = 2, 3, \ldots |A|$, along with the computed frequency values as above and assigning it to another new set $Q_j, 2 <= j <= |A|$. We now have $Q_{|A|}$ sets, with each set having the frequencies of each of the attribute value for every data artefact. Finally, the mode of each attribute of the data artefacts of each set $Q_j, 2 <= j <= |A|$ is computed, which forms the remaining candidates of the candidate set. An eligible first seed artefact is selected from the candidate set based on the criteria that it is in a dense region, far from the mode of the data set but nearer to the first exemplar. These candidates represent the neighbours of the exemplar, closer a candidate to the exemplar better is the information gained with respect to the exemplar. For identifying the remaining K-1 seed artefacts, an exemplar is found by finding the artefact which has the maximum value of summing up the density of each data artefact with its minimum distance from the previously selected seed artefacts. For each exemplar the corresponding candidate set is computed as stated above. Finally, the seed artefact is selected from the candidate set such that it is in a dense region, at a minimum distance from the previously selected seed artefact and closer to its corresponding exemplar.

Khan and Ahmad (2013) have pointed out the inaccuracies in the cluster metric values obtained by Bai et al. (2012).

Khan and Ahmad (2013) propose a seed selection method which uses three attribute selection methods based on the significance of attributes. The first method is the vanilla approach where all the attributes are considered as significant attributes. The second method is the prominent attribute method where an attribute is considered if the number of unique values of the attributes is lesser than or equal to the required number of clusters. The third method is to identify the most significant attributes by measuring the co-occurrence of its values with the values of other attributes. The best attributes are those whose values are well separated against all the attributes. Even after applying the second and the third methods of attribute selection, considerate amount of reduction in attributes was seen in only 3 out of 7 data sets as mentioned in their paper. The initial seed selection algorithm is applied to the attributes got by using all the above three attribute selection methods. First the number of unique values present in each attribute is computed. This represents the number of clusters present in each attribute. For each of the unique values of an attribute the data set is divided into groups such that the data artefacts having the same value as the unique values of this attribute are assigned to the same cluster. Then the modes (seeds) of each group are obtained. K–modes clustering algorithm is then executed using the modes obtained as the initial seed artefacts. A label 'i' depicting the cluster number is assigned to each data artefact. This is repeated for all the attributes of the data set. Finally, each data artefact is associated with a string got from the labels assigned to it. Artefacts which have the same string are assigned to the same cluster. The frequency of the unique strings generated is computed. If the number of unique strings is equal to the required number of clusters (K), the procedure is stopped. If it is higher, hierarchical clustering is used to merge similar distinct cluster strings into the desired number (K) of clusters. Then the center (mode) of each cluster is calculated. If the number of unique strings is lesser than K, the procedure is repeated by reducing K. By computing the performance evaluation metrics (Accuracy, Precision, Recall) for the clustering solutions obtained for each of the attribute selection methods, the clustering solution with the best performance metrics is chosen as the final clustering solution for the given data set.

In this paper, we propose a method to identify initial seed artefacts which are selected from highly dense regions such that they are well separated from each other. The identified seed artefacts are fed into K- modes clustering. We use both distance and density of each data artefact to select the eligible seed artefacts. To ensure that the seed artefacts are well separated and do not fall in the same region as the previously selected seed artefacts, our algorithm uses a separation parameter. This parameter provides a higher weightage to the index component of distance of the data artefacts from the previously selected seed artefacts and a proportionately lesser weightage for the index component of density. The weights assigned to the index component of distance range from $1/2, 2/3, 3/4 \ldots 9/10$ and the weights assigned to the index component of density range from $1/2, 1/3, 1/4 \ldots 1/10$. Based on the nature of the given data artefacts, one or more of the weightage values may generate seed artefacts which are well separated and ensure that only one set of artefacts are closer to the final

Please cite this article in press as: Sajidha, S.A., et al. Initial seed selection for K-modes clustering – A distance and density based approach. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.04.013

centroids. This yields the best and global minimal cost for the clustering solution as compared to random K-modes clustering algorithm.

3. Proposed seed selection method

The first step in our method is to find the density of each data artefact. Next we find the mode of the data set. The point which is in the densest region and far from the mode is chosen as the first seed artefact. The remaining seed artefacts are selected such that they are selected from a dense region and are ensured that they are well separated from the previously selected initial seed artefacts. This ensures that the seed artefacts are selected from different possible clusters. This is achieved by assigning weights to the index components of both the distance and density of each data artefact, such that, higher weights varying from $1/2, 2/3, 3/4, \ldots, 9/10$ are assigned to index component of distance and corresponding lower weights varying from $1/2, 1/3, 1/4, \ldots, 1/10$ are given to the index component of density, such that ratio between these two weights increases uniformly. This ensures that the initial seed artefacts obtained are sufficiently well separated such that they are obtained from the anticipated clusters of the clustering solution.

Algorithm 2. Proposed Seed Selection Algorithm

Input: A collection of 'n' data artefacts DA with 'm' attributes.
Output: K initial seed artefacts.

Step 1. For each data artefact DA_i where $i = 1 \ldots n$
  For each data artefact DA_j where $j = 1 \ldots n$
    Compute $M_{ij} = \text{sim}(DA_i, DA_j)$
    where n = no. of data artefacts, and $\text{sim}(DA_i, DA_j)$ is computed using Eq. (1).

Step 2: Compute the distance vector (DV) of the data set

$$\text{DistVec}(i) = \sum_{j=1}^{n} M_{ij}$$

Step 3: Compute the density vector of the data set

$$\text{DensVec}(i) = -\text{DistVec}(i)/n$$

Step 4: Sort density vector DensVec with the corresponding indices of the artefacts in decreasing order and store it as SortDens.

Step 5: Find mode vector of the data set
  $\text{ModeVect}(i) = \{a | a \in A_q \text{ and } \text{freq}(a, A_q) \text{ is maximum}\}$
  where $\text{freq}(a, A_q)$ is the number of times a appears in attribute A_q and $q \in 1 \ldots\ldots m$ where m is number of attributes of data set

Step 6: Find distance of all data artefacts from mode vector
  For each data artefact DA_i where $i \in 1 \ldots n$
    $\text{ModeDist}(i) = \text{Dist}(DA_i, \text{ModeVec})$

Step 7: For each data artefact compute the density and distance from the mode and store it as DensModeDist (DMD)
  For each data artefact DA_i where $i = 1 \ldots n$
    $\text{DensModeDist}(i) = \text{DensVec}(i) + \text{ModeDist}(i)$

Step 8: For α from 1 to 9
  Step 8.1: Find the index of the first Initial Seed Artefact (ISA)
    $ISA_1 = \text{index\_of\_artefact}[\max(\text{DMD})]$
  Step 8.2: Loop l from 2 to K
    Step 8.2.1: For $i = 1 \ldots n$
      For $h = 1 \ldots .l - 1$
        Get $d_{ih} = M[i, ISC_h]$ where $M = M_{ij}$ is computed from Step1
    Step 8.2.2: For $i = 1 \ldots n$
      $S_i := \min(d_{ih})$
    Step 8.2.3: Sort S_i with the corresponding indices of the concepts in non-increasing order to get the SortDist (SD)
    Step 8.2.4: For each data artefact DA_i where $i = 1 \ldots n$
      Dens_Dist_New_Index (DDNI(i)) =
      $\frac{\alpha}{1+\alpha} *$ (index of DA_i from Sort Dist(DA_i))+
      $\frac{1}{1+\alpha} *$ (index of DA_i from Sort Dens(DA_i))
    Step 8.2.5: Obtain the index of the minimum value in DDNI
      $\text{ind\_NDDI} = \text{index}[\min(\text{DDNI})]$
      $ISA_j := \text{ind\_DDNI}$
  Step 8.3: Retrieve the seed concepts (Actual_Seed) using their index
    values in $ISA_1, ISA_2, \ldots, ISA_K$

In Step 1 the distance matrix is computed which is used to estimate the density of each data artefact (Step 3) and also used in extracting the distance of the artefacts from the previously selected seed artefacts (Step 8). This is computed only once and referred until the desired number of seed artefacts is identified. In step 8.2.5 we vary α from 1 to 9 to compute the weights assigned to the index components of distance and density. Consequently, weights ranging from 0.5 to 0.9 are assigned to index component of distance while corresponding lower weights are assigned to index component of density of the data artefacts maintaining a uniform increase in the ratio between the two weights. This enables our proposed algorithm in identifying well separated seed artefacts.

Algorithm 3. Initial Seed Artefact Based Best Clustering Solution.

Input: Actual Initial Seed Artefacts(ISA) from Algorithm2
Output: Best clustering solution

Step 1: For $\alpha = 1 \ldots 9$
  Step 1.1: Initialize the seeds in K-modes(Algorithm1) using seed artefacts from ISA set obtained from Algorithm 2 and execute K-modes algorithm on the data artefacts to be clustered.
  Step 1.2 : Find distance between the cluster center and all the data artefacts in the cluster and store it in SSEvector[α]
  Step 1.3: Find the sum of distances between the cluster centers and its corresponding ISAs' and store it in DBSCvector[α]

Step 2: From the SSEvector find α corresponding to the minimum SSEvalue.
  If there is a tie then find α corresponding to minimum value in DBSCvector.
    If the tie still persists select maximum α value.
  Let this α be finalAlpha.

Step 3: Clustering solution corresponding to finalAlpha value will be the best clustering solution for the given data artefacts.

The nine sets of initial seed artefacts corresponding to the nine values of α got from Algorithm 2 are given as input to Algorithm 1 where K-modes Algorithm is executed. The cost function is estimated for each of the clustering solutions using Algorithm 3. The final clustering solution is selected based on one of the following criteria:-

Please cite this article in press as: Sajidha, S.A., et al. Initial seed selection for K-modes clustering – A distance and density based approach. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.04.013

1. Solution having minimum NSWSE (cost value).
2. Sometimes two or more α values may lead to the same minimum cost when a few initial seed artefacts are misplaced. Our goal is to find all the initial seed data artefacts which fall in different clusters and this in turn produces the best solution. In this case we estimate the error between the seed artefacts and the centre of the clusters for every initial seed artefact set corresponding to different α values. The set that corresponds to minimum error, which shows that the obtained seed artefacts are well separated and are closer to the centroids of the clusters, is chosen.
3. When there is a tie in criteria 2 we select the initial seed artefact set obtained using the highest α value.

Table 1
Synthetic data artefacts.

A1	A2	A3	A4	CLASS
J	X	L	A	C1
H	X	M	A	C1
H	Z	L	C	C1
H	X	L	A	C1
H	X	O	A	C1
I	Y	N	B	C2
I	Y	M	C	C2
I	Z	M	B	C2
J	Y	L	B	C2
I	X	M	B	C2
J	Z	M	C	C3
I	Z	L	C	C3
J	Z	N	C	C3
J	X	N	C	C3
J	Z	N	B	C3

3.1. Illustration for synthetic data set

We illustrate our proposed algorithm for our synthetic data set (classified into 3 categories) in Table 1 with 4 attributes and 15 instances.

We clearly see that the mode of the data set is J X L C. The density of all the data artefacts is computed and ranked from highest to lowest as given in Table 2. To compute the density of each data artefact the distance between each artefact and all the other artefacts is calculated. The first seed artefact is the one which has maximum density and farthest from the mode. Hence, the sum of density of each data artefact with its corresponding distance from the mode (Mode Distance) is computed and the artefact which has the maximum sum is chosen as the first seed artefact as in Table 3. Consequently, the eighth data artefact belonging to class C2 is chosen as the first seed artefact.

Distance of all points from first seed artefact is computed and they are ordered in decreasing order as in Table 4.

We find the next seed artefact by computing a new index by applying weights $\frac{\alpha}{1+\alpha}$ to the index component of distance and $\frac{1}{1+\alpha}$ to index component of density of each data artefact using the following formulae: –

$$\begin{aligned}&\text{Dens\_Dist\_New\_Index}(\text{DDNI}(i)) :\\ &= \alpha/(1+\alpha) * (\text{index of } DA_i \text{ from Sort Dist}(DA_i))\\ &\quad + 1/(1+\alpha) * (\text{index of } DA_i \text{ from Sort Dens}(DA_i))\end{aligned} \tag{2}$$

This is elaborated in Table 5.

Table 2
Artefacts ordered based on density.

Data artefact	Index	Density
11	1	−1.03333
12	2	−1.06667
13	3	−1.08333
14	4	−1.08333
8	5	−1.1
10	6	−1.1
1	7	−1.11667
3	8	−1.11667
15	9	−1.11667
2	10	−1.2
4	11	−1.2
7	12	−1.23333
9	13	−1.23333
6	14	−1.31667
5	15	−2

Table 3
Sum of Density and Mode Distance.

Data artefact	Density	Mode Distance	Density + Mode Distance
1	−1.11667	1	−0.1167
2	−1.2	3	1.8
3	−1.11667	2	0.88333
4	−1.2	2	0.8
5	−2	3	1
6	−1.31667	4	2.68333
7	−1.23333	3	1.76667
8	**−1.1**	**4**	**2.9**
9	−1.23333	2	0.76667
10	−1.1	3	1.9
11	−1.03333	2	0.96667
12	−1.06667	2	0.93333
13	−1.08333	2	0.91667
14	−1.08333	1	−0.0833
15	−1.11667	0	−1.1167

Table 4
Data artefacts ordered based on distance from first seed artefact.

Data artefact	Index	Distance
1	1	4
4	2	4
5	3	4
14	4	4
2	5	3
3	6	3
9	7	3
13	8	3
6	9	2
7	10	2
11	11	2
12	12	2
15	13	2
10	14	1

Table 5
New Index based on density and distance.

Data artefact	DDNI
1	**2.5**
2	6.25
3	6.5
4	4.25
5	6
6	10.25
7	10.5
9	7.75
10	12
11	8.5
12	9.5
13	6.75
14	4
15	12

Please cite this article in press as: Sajidha, S.A., et al. Initial seed selection for K-modes clustering – A distance and density based approach. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.04.013

The artefact which has the minimum value of DDNI is chosen as the next seed artefact. Here the first data artefact belonging to class C1 is chosen as the second seed artefact. The next seed artefact is chosen by computing the distances of all the data artefacts from the previously selected seed artefacts and choosing the minimum of the two (Min_dist) as in Table 6.

The minimum distances are then arranged in non-increasing order as shown in Table 7.

Table 8 shows the values obtained by applying Eq. (2).

The third seed artefact is the one which has minimum DDNI (13th data artefact of class C3). Our algorithm was executed on the synthetic data artefacts for different values of α from 1...9 and the best clustering solution with minimum NSWSE was obtained when $\alpha = 3$. Hence, from the illustrated example we see that our proposed algorithm works effectively in identifying the initial seed artefacts based on the distance from the previously selected seed artefacts and its corresponding density. Our proposed algorithm ensures repeatability of the clustering solution. This is presented in Table 9. We obtained the clustering solutions for our synthetic data set by executing our proposed seed selection algorithm with K-modes algorithm. We also obtained the clustering solutions using random K-modes algorithm (6 runs). Here we observe that in all the six runs, our proposed seed selection algorithm ensures that the data artefacts are assigned to the same cluster in every run thereby ensuring repeatability of the clustering solution. In the random K-modes algorithm, since random sets of seeds are chosen in every run the data artefacts are assigned to different clusters in each run and hence, we see that repeatability cannot be ensured.

Table 6
Distance from previously selected seed artefacts.

Data artefact	Distance from Seed one	Distance from Seed two	Min_dist
2	3	2	2
3	3	3	3
4	4	1	1
5	4	2	2
6	3	4	3
7	2	4	2
9	4	2	2
10	2	3	2
11	2	3	2
12	2	3	2
13	3	3	3
14	4	2	2
15	3	3	3

Table 7
Sorted Min_dist.

Index	Data artefact	Min_dist
1	3	3
2	6	3
3	13	3
4	15	3
5	2	2
6	5	2
7	7	2
8	9	2
9	10	2
10	11	2
11	12	2
12	14	2
13	4	1

Table 8
New index based on density and distance.

Data artefact	DDNI
2	6.25
3	2.75
4	12.5
5	8.25
6	5
7	8.25
9	9.25
10	8.25
11	7.75
12	8.75
13	**3**
14	10
15	5.25

3.2. Comparison of time complexities

Table 10 compares the time complexities of our proposed cluster initialization algorithm with the three competing initialization methods of Cao et al. (2009), Khan and Ahmad (2013) and Wu et al. (2007).

In Wu et al. (2007) a time complexity of O(n) is achieved when $\sqrt{n}$ random data artefacts are considered. In Khan and Ahmad (2013) O(n) is achieved only under certain conditions as specified in their paper. Our proposed algorithm computes the distance matrix to ascertain the density of all the data artefacts in step 1 of algorithm 2. This step is performed only once and referred to in step 8.2.1 of algorithm 2 until the ideal initial seed artefacts are found. Hence, time complexity of our algorithm turns out to be $O(n^2)$. Nevertheless, we see from our experimental results discussed in the next section that our proposed algorithm outperforms the existing methods for most of the data sets.

4. Experimental results and discussion

The objective of finding initial seed artefacts is to ensure repeatability of the clustering solution with global minimum cost. This was ensured every time our proposed algorithm was executed for the data sets. The accuracy of our clustering solution was estimated by dividing the sum of within-cluster similarity by the total number of data artefacts.

All experiments were conducted on a DELL laptop with an Intel (R) Core(TM) i-5-6200U CPU @2.30 GHz 2.40 GHz and 8 GB of main memory with 64bit OS.

We carried out our experiment on 6 pure categorical data sets from the UCI Machine Learning Repository (Bache and Lichman, 2013) and compared our results with those obtained using random K-modes algorithm and the algorithms proposed by Wu et al., Cao et al. and Khan et al. Description of experimental data sets is given in Table 11. Breast cancer data set is the largest of the data sets used in our experiment with 699 artefacts and lung cancer data set is the smallest with 32 artefacts. We have implemented our algorithm using R version: 3.4.1, R Studio version: 1.1.383 and OS: Windows 10.

4.1. Data Pre-processing

The basic assumption for handling missing values is that all the data artefacts belonging to the same class in the data set generally have the same value for a particular attribute. So, the technique followed to handle the missing values was as follows: for a data artefact DA_i belonging to class L_k having a missing value for attribute A_q, the most frequently occurring value or mode, M_q of

Table 9
Repeatability of clustering solution.

Data Artefact	Proposed seed selection + K-modes	Random K-modes											
	In all 6 runs categorisation	Run 1		Run 2		Run 3		Run 4		Run 5		Run 6	
	Size 5,5,5	Size	3,6,6	Size	5,5,5	Size	3,6,6	Size	6,5,4	Size	5,6,4	Size	4,8,3
	Cluster Number												
1	3	3		3		3		1		1		3	
2	3	3		3		3		1		1		1	
3	3	3		3		3		1		1		1	
4	3	3		3		3		1		1		1	
5	3	3		3		3		1		1		1	
6	1	1		1		1		1		3		2	
7	1	1		1		1		2		3		2	
8	1	2		1		2		3		2		2	
9	1	1		1		1		2		3		3	
10	1	2		1		2		3		2		2	
11	2	2		2		2		2		2		2	
12	2	2		2		2		3		2		2	
13	2	2		2		2		2		2		2	
14	2	3		2		3		2		1		3	
15	2	2		2		2		3		2		2	

attribute$A_{q,}$ for all data artefacts belonging to class L_k was found. The missing value in DA_i for attribute A_q was replaced with M_q.

4.2. Comparison and performance evaluation metric

The quality of clustering results was evaluated and comparisons were made using the performance metrics used by Wu et al. (2007). Let us assume that the given data set contains K classes. For evaluating a clustering method, let $corr_L$ denote the number of data artefacts which are correctly assigned to class L_l, let wor_L denote the number of data artefacts which are wrongly assigned to class L_l, and let rej_L be the number of data artefacts that are incorrectly rejected from class L_l, thus precision (PR), recall (RE) and accuracy (AC) are defined as follows

$$PR = \frac{\sum_{L=1}^{K}\left(\frac{corr_L}{corr_L + wor_L}\right)}{K}$$

$$RE = \frac{\sum_{L=1}^{K}\left(\frac{corr_L}{corr_L + rej_L}\right)}{K}$$

$$AC = \frac{\sum_{L=1}^{K}(corr_L)}{N}$$

Table 10
Comparison of time complexities.

Initialisation method	Order of Complexity
Wu et al.	O(cn), where c can be between 2 to $n^{0.5}$
Cao et al.	$O(nmK^2)$
Khan et al.	$O(nm + rKm^2n + nlogn)$
Proposed method	$O(n^2)$

4.3. Clustering results

We present here the results of K-modes clustering using the initial seed artefacts obtained using our proposed method. We compared our results with random K-modes and methods of Wu et al., Cao et al. and Khan et al. and the results are presented in Tables 12a & 12b.

From the accuracy, precision and recall metrics provided in Tables 12a & 12b, we see that our proposed method outperforms random K-modes for soybean, lung cancer, breast cancer and congressional votes data sets except for dermatology and zoo data sets.

From Table 12a, we see that the proposed method is on par with Wu and Cao's methods and outperforms Khan's approach for soybean data set. For lung cancer data set our method outperforms in all the three metrics when compared to Wu and Cao's algorithms, and the accuracy and the recall metrics of Khan outperforms our proposed method. For breast cancer data set the proposed method outperforms all the comparative methods in all the 3 metrics. With respect to congressional vote data set on comparing our method with the random method and Khan's method, we find that our method fairs better than both these methods in all the three metrics.

From Table 12b for dermatology data set when comparing the proposed method with Khan's method, we found that our method does not perform well as can be seen from all the three performance metrics. For zoo data set we find that our method shows

Table 11
Description of experimental data sets.

Data set	Instances	No. of categorical variables	Partitions	Missing values
Soybean	47	35	4	No
Lung Cancer	32	56	3	Yes
Breast Cancer	699	9	2	Yes
Congressional Vote Data	435	16	2	Yes
Dermatology	366	33 + 1 Numerical(categorized to 10 categories)	6	Yes
Zoo	101	16	7	No

Please cite this article in press as: Sajidha, S.A., et al. Initial seed selection for K-modes clustering – A distance and density based approach. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.04.013

Table 12a
Comparative study.

Data set	Measure	Random K- modes	Wu	Cao	Khan	Best α value	Proposed
Soybean	AC	0.8644	**1**	**1**	0.9574	**2**	**1**
	PR	0.8999	**1**	**1**	0.9583		**1**
	RE	0.8342	**1**	**1**	0.9705		**1**
Lung Cancer	AC	0.5210	0.5000	0.5000	0.5000	**1**	**0.5625**
	PR	0.5766	0.5584	0.5584	**0.6444**		0.5766
	RE	0.5123	0.5014	0.5014	0.5168		**0.5764**
Breast Cancer	AC	0.8364	0.9113	0.9113	0.9127	**7**	**0.9428**
	PR	0.8699	0.9292	0.9292	0.9318		**0.9509**
	RE	0.7743	0.8773	0.8773	0.8783		**0.9229**
Congressional Vote	AC	0.4972	–	–	0.8506	**9**	**0.8667**
	PR	0.503	–	–	0.8484		**0.8659**
	RE	0.5031	–	–	0.8672		**0.8859**

Table 12b
Comparative study.

Data set	Measure	Random K-modes	Wu	CAO	Khan	Best α value	Proposed
Dermatology	AC	0.2523	–	–	**0.7372**	**9**	0.6885
	PR	0.2697	–	–	**0.7909**		0.6293
	RE	0.2954	–	–	**0.757**		0.5823
Zoo	AC	0.8324	0.8812	0.8812	**0.8911**	**1**	0.7327
	PR	0.8433	**0.8702**	**0.8702**	0.7224		0.6077
	RE	0.6576	0.6714	0.6714	**0.7716**		0.6402

a lower value for accuracy, precision and recall metric compared to the three methods.

Hence, it can be inferred from the experimental analysis that our method out performs the other methods when the data artefacts are approximately uniformly distributed across the classes which is not the case for dermatology and zoo data sets.

5. Conclusion

The main objective of our proposed algorithm in this paper is to identify initial seed artefacts for K-modes algorithm. We ensure that the seed artefacts are well separated such that they are selected from different clusters and are from dense regions. A clear illustration of our proposed method using 15 synthetic data artefacts was made to explain the working of our method. We have considered six benchmark data sets for the experimental evaluation of our proposed method vis-à-vis the methods of Wu et al., Cao et al. and Khan et al. Clustering solutions were obtained, by considering all the attributes for all the data sets. We have used three performance metrics to evaluate the performance of our algorithm, which is shown in Tables 12a & 12b. Out of 6 data sets that we have considered our proposed method outperforms in 4 data sets which proves that our algorithm is effective in identifying initial seed artefacts thus ensuring repeatability of the clustering solution with optimal cost. Our proposed method, considers the density estimation of all the data artefacts which requires the computation of distances of each data artefact from all the data artefacts, hence the time taken by our proposed method as compared to other approaches is more. Albeit, our approach yields better clustering solutions for a majority of data sets as compared to the existing approaches.

6. Future work

From our exploration of previous research work, we have understood that efficient clustering using K-means and K-modes clustering algorithms are dependent on the initial seed artefacts. Hence, we have proposed a new approach for identifying initial seed artefacts for both numerical data using K-means (Azimuddin and Desikan, 2017) and categorical data using K-modes as mentioned in this paper. We see that many areas such as finance and health sector generate mixed data having both numerical and categorical data artefacts. In this light, we propose to extend our initial seed selection technique for mixed data sets having both numerical and categorical attributes, by introducing a novel distance measure to perform the clustering task.

Acknowledgements

The authors are very grateful to the anonymous reviewers and editor. Their many helpful and constructive comments and suggestions helped us to significantly improve this work.

References

Azimuddin, Sajidha Syed, Desikan, Kalyani, 2017. A simple density with distance based initial seed selection technique for K means algorithm. CIT J. Comput. Inf. Technol. 25 (4), 291–300. https://doi.org/10.20532/cit.2017.1003605.

Bache, K., Lichman, M., 2013. UCI machine learning repository, http://archive.ics.uci.edu/ml.

Bai, Liang, Liang, Jiye, Dang, Chuangyin, 2011. An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. Knowl.-Based Syst. 24 (6), 785–795. https://doi.org/10.1016/j.knosys.2011.02.015.

Bai, Liang, Liang, Jiye, Dang, Chuangyin, Cao, Fuyuan, 2012. A cluster centers initialization method for clustering categorical data. Expert Syst. Appl. 39 (9), 8022–8029. https://doi.org/10.1016/j.eswa.2012.01.131.

Cao, F., Liang, J., Bai, L., 2009. A new initialization method for categorical data clustering. Expert Syst. Appl. 36, 10223–10228. https://doi.org/10.1016/j.eswa.2009.01.060.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Disc. 2, 283–304. https://doi.org/10.1023/A:1009769707641.

Huang, Z., 1997. A fast clustering algorithm to cluster very large categorical data sets in data mining. In Research issues on data mining and knowledge discovery.

Khan, Shehroz S., Ahmad, Amir, 2013. Cluster center initialization algorithm for K-modes clustering. Expert Syst. Appl. 40 (18), 7444–7456. https://doi.org/10.1016/j.eswa.2013.07.002.

Milligan, G.W., Isaac, P., 1980. A study of variable standardization. J. Classif. 5 (2), 181–204. https://doi.org/10.1007/BF01897163.

Please cite this article in press as: Sajidha, S.A., et al. Initial seed selection for K-modes clustering – A distance and density based approach. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.04.013

Ralambondrainy, H., 1995. A conceptual version of the k-means algorithm. Pattern Recogn. Lett. 16 (11), 1147–1157. https://doi.org/10.1016/0167-8655(95)00075-R.

Sun, Ying, Zhu, Qiuming, Chen, Zhengxin, 2002. An iterative initial-points refinement algorithm for categorical data clustering. Pattern Recogn. Lett. 23, 875–884. https://doi.org/10.1016/S0167-8655(01)00163-5.

Wu, S., Jiang, Q., Huang, J.Z., 2007. A new initialization method for clustering categorical data. In: Proceedings of the 11th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining PAKDD'07. Springer-Verlag, Berlin, Heidelberg, pp. 972–980. https://doi.org/10.1007/978-3-540-71701-0_109.

Please cite this article in press as: Sajidha, S.A., et al. Initial seed selection for K-modes clustering – A distance and density based approach. Journal of King Saud University – Computer and Information Sciences (2018), https://doi.org/10.1016/j.jksuci.2018.04.013