K-Harmonic means type clustering algorithm for mixed datasets

Amir Ahmad a,1, Sarosh Hashmi

Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, Saudi Arabia

A R T I C L E   I N F O

Article history:
Received 14 December 2015
Received in revised form 29 February 2016
Accepted 16 June 2016
Available online 29 June 2016

Keywords:
Clustering
Categorical attributes
Numeric attributes
Mixed data
K-Harmonic means clustering

A B S T R A C T

K-means type clustering algorithms for mixed data that consists of numeric and categorical attributes suffer from cluster center initialization problem. The final clustering results depend upon the initial cluster centers. Random cluster center initialization is a popular initialization technique. However, clustering results are not consistent with different cluster center initializations. K-Harmonic means clustering algorithm tries to overcome this problem for pure numeric data. In this paper, we extend the K-Harmonic means clustering algorithm for mixed datasets. We propose a definition for a cluster center and a distance measure. These cluster centers and the distance measure are used with the cost function of K-Harmonic means clustering algorithm in the proposed algorithm. Experiments were carried out with pure categorical datasets and mixed datasets. Results suggest that the proposed clustering algorithm is quite insensitive to the cluster center initialization problem. Comparative studies with other clustering algorithms show that the proposed algorithm produce better clustering results.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Partitioning of data points into different set of groups on the basis of similarity between data points is called clustering [1]. These groups are called clusters. K-means clustering method [2] is a useful clustering method for pure numeric datasets because of its low time complexity. The selection of initial cluster centers is a critical step in this algorithm. The final clusters are strongly related with the initial cluster centers. Selecting initial cluster centers randomly, is a popular method for this purpose. However, stable results are not achieved in different runs because different initial cluster centers are generated in different runs due to random initialization of cluster centers. Several cluster center initialization methods have been suggested. These methods try to compute initial cluster centers [3–6]. With these initial cluster centers, K-means clustering algorithms give stable results. K-Harmonic means (KHM) clustering algorithm tries to overcome this problem in a different way by using a different cost function [7]. Experimental results suggest that clustering results achieved from KHM clustering algorithm are more stable as compared to K-means clustering algorithm with random initialization of cluster centers [7]. K-means clustering algorithm and KHM clustering algorithm can handle only numeric attributes. However, some datasets have both categorical and numeric attributes. Many K-means type clustering methods have been proposed that extend K-means clustering algorithm for mixed datasets [8–11]. However, all these clustering algorithms suffer from cluster center initialization problem. Cluster center initialization problem has not got much attention for mixed datasets. Ji et al. [12,13] propose two cluster center initialization methods for K-means type clustering algorithms for mixed datasets. However, the time complexity of these methods is quadratic with respect to the number of data points; hence they are not suitable for K-means type clustering algorithms which have linear complexity with respect to the number of data points.

In this paper, we suggest a clustering method that can handle mixed data and is robust results to selection of initial cluster centers. Ahmad and Dey [10] propose a distance measure and a center definition for mixed datasets. Using this distance measure and a cluster center definition they propose a clustering algorithm for mixed datasets. This distance measure has also been applied for fuzzy clustering of mixed datasets [14] and subspace clustering of mixed datasets [11]. We propose that this distance measure and a new cluster center definition for the mixed datasets can be used with the cost function of KHM clustering algorithm. The proposed KHM type algorithm can handle mixed datasets and the proposed algorithm gives very stable clustering results. The paper is organized in following way. The related work is presented in Section 2. Our proposed algorithm is described in Section 3. Section 4 has results and discussion. Section 5 discusses conclusion and future work.
2. Related work

In this section, we will discuss K-means type clustering algorithms for numeric datasets, categorical datasets and mixed datasets. We will also discuss the cluster initialization methods for these algorithms.

K-means algorithm [2] is a partitional clustering method. It uses an iterative methodology to cluster the data points into k clusters. K-means clustering algorithms iteratively minimize the following cost function, $\xi$, for $n$ data points:

$$\xi = \sum_{i=1}^{n} |d_i - C_j|^b$$

where $n$ is the number of data points, $C_j$ is the nearest cluster center of data point $d_i$ and $b$ is an integer. For Euclidean distance, $b = 2$.

The means value of each attribute of all the data points in a cluster is used to define the cluster center. Through the iterative partitioning, K-means algorithm minimizes the cost function. This algorithm starts with $k$ selected cluster centers. Data points are assigned to the clusters and then cluster centers are recalculated by using the data points in the cluster. These two steps are repeated until a predefined termination condition is achieved. The final clusters are related with initial cluster centers. Various cluster centers initialization methods have been suggested.

Duda and Hart [13] propose a recursive method for computing initial cluster centers. The method is to calculate the initial cluster centers from the clusters for another clustering problem. The initial cluster centers for the cluster problem can be the final cluster centers for the cluster problem plus a data point that is farthest from the nearest cluster center. This way the $k-1$ cluster centers can be used to create $k$ cluster centers. The cluster center for the one-cluster problem is the mean of all data points. The results are extended to get two clusters. The procedure is repeated to get $k$ cluster centers. Bradley and Fayyad [3] suggest a cluster center initialization procedure that uses joint probability density of the data. The algorithm initially selects $J$ small random subsets of the data. The subsets are clustered by K-means clustering method. The data is then clustered $J$ times by using the cluster centers achieved by clustering of these $J$ random subsets of the data. The initial cluster centers are selected from these $J$ clustering results which give minimal value of the cost function. Khan and Ahmad [4] present an algorithm to computer initial cluster centers. The algorithm uses two observations, the first that similar data points have same cluster membership irrespective to the initial cluster centers. Secondly, an individual attribute can be helpful in providing some indication about initial cluster centers. Firstly, cluster centers are computed for individual attributes. It is assumed that each attribute is normally distributed. This information is used to compute the center in the direction of that attribute. The data is then clustered by using K-means clustering algorithm with these centers. Resultant clusters are used to get the cluster centers. Then K-means clustering is run on all the data. The process is done for all the attributes. All the clustering results are converted into strings. Data points with similar strings form clusters. The centers of these clusters are used as initial cluster centers. Erisoglu et al. [6] suggest a method that are based on selecting 2 of the $m$ attributes that best represent the change in the dataset. The distances between data points in these two attributes are used to compute initial cluster centers.

K-means clustering method can handle only numeric attribute as the means is calculated for computing cluster center and the distances (like Euclidean distance) between points and cluster center is calculated. Huang [16] presents the K-modes clustering method for pure categorical datasets. The method replaces means of clusters with modes. Hamming distance is used as a distance measure [16]. Generally, K-modes clustering algorithm runs with random initial cluster centers. With random initialization, clusters are not consistent in different runs. Several cluster center initialization methods have been suggested to overcome this problem [17–22].

Huang proposes two methods for cluster center initialization of K-modes clustering algorithm [16]. The first method suggests that the first $k$ distinct data points be taken as initial cluster centers. In the second method, the frequencies of all attribute values are calculated and the most frequent values are assigned as initial cluster centers. Sun et al. [18] show that combination of Bradley and Fayyad’s algorithm [3] with K-modes clustering algorithm gives more reliable clustering results than random initial cluster centers. He [19] suggests a farthest point heuristics for initialization of cluster centers that uses frequency of attribute values.

Cao et al. [20] present a method for cluster center initialization by using a density function and a distance function. The first cluster center is calculated by this density function. The distance between the probable new cluster center and already computed cluster centers are calculated. These distances along with the density function are used to compute other cluster centers. Khan and Ahmad [21] propose an algorithm for cluster center initialization that combines the results of multiple clustering of a dataset based on different initial cluster centers to get initial cluster centers. Wu et al. [22] propose a cluster centers initialization approach based on a density function. The complexity of this algorithm is quadratic. Random subsampling is suggested to reduce the complexity, however, because of this step consistent clustering results are not guaranteed.

Some of the datasets have both numeric and categorical attributes. K-means clustering algorithm has been extended to handle these mixed datasets [8]. Huang [8] presents a cost function for mixed datasets along with definitions of cluster centers and distances. Means values for numeric attributes and mode values for categorical attributes are used for cluster centers. However, the use of mode as cluster center for categorical attributes results in loss of information. Also, Hamming distance is considered between a cluster center and a data point which is not very accurate for multi-valued categorical attributes. To overcome these problems, Ahmad and Dey [10] present a cost function and a distance measure. A novel definition of a cluster center is also suggested to provide a better representation of clusters [10]. Their clustering results were competitive to other computationally expensive clustering algorithm for mixed datasets [10]. Huang et al. [9] suggest a clustering algorithm for mixed datasets that uses attribute weights in cost function. Ji et al. [23] present a K-prototype clustering method for mixed datasets that combine the cluster definition proposed by Ahmad and Dey [10] with attribute weights proposed by Huang et al. [9]. All these clustering methods use random initialization of cluster centers. The cluster center initialization problem has not been addressed for these algorithms. We found only two algorithms [12,13] in the literature for this purpose. Ji et al. [12,13] propose algorithms for computing cluster centers for K-means type clustering algorithms for mixed datasets. These algorithms use the concept of nearest neighbors and density. The complexity of these algorithms is quadratic which makes the algorithms unsuitable for K-means type clustering algorithms which have linear time complexity.

Almost all the methods that overcome cluster center initialization problem in K-means clustering algorithm or K-mode clustering algorithm compute initial cluster centers so that these clustering algorithms can run with these cluster centers and give single clustering result. However, K-Harmonic means clustering algorithm [7] for numeric attributes gives robust results with random initialization. In other word, this method is very insensitive to the initial cluster centers. Our proposed algorithm uses the philosophy of K-
Harmonic means clustering algorithm, which [7] is discussed in the next section.

2.1. K-Harmonic means (KHM) clustering algorithm

Cluster centers initialization is a problem for K-means clustering algorithm. To overcome this problem Zhang [7] proposed KHM clustering algorithm with a novel cost function. KHM clustering algorithm uses the harmonic means of the distances from data points to the cluster centers in its cost function. KHM clustering algorithm uses the following cost function defined by \(\xi\) for \(n\) data points:

\[
\xi = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} |d_i - C_j|^a}
\]

(2.1)

\(d_i\) is the \(ith\) data point, \(C_j\) is the \(jth\) cluster center. \(k\) is the number of desired clusters. \(a\) is a positive number. The cost function uses the squared Euclidean distance (\(a = 2\)), however, Zhang [7] suggests that generalizing to other distance functions is possible.

KHM clustering algorithm has following steps for data having \(n\) data points and \(k\) desired clusters:

1. KHM algorithm starts with random cluster centers.
2. The distances between each data point to all the centers are calculated.
3. The new cluster centers are calculated.

They are calculated in following way:

(i) Calculate \(\phi(d_i, C_j) = |d_i - C_j|\) for \((i = 1..n, j = 1..k)\) (\(\phi(d_i, C_j)\) is the distance between \(ith\) data point, \(d_i\), and \(jth\) cluster center, \(C_j\). Distances between each data point to each cluster center are calculated).

(ii) \(\alpha_i = \frac{1}{\left(\sum_{j=1}^{k} 1/\phi(d_i, C_j)^a\right)}\) \((a\) is a positive number)

(iii) \(q_{ij} = \alpha_i/\phi(d_i, C_j)^{a+2}\)

(iv) \(q_j = \sum_{i=1}^{n} q_{ij}\)

(v) \(p_{ij} = q_{ij}/q_j\)

(vi) \(C_{js} = \sum_{i=1}^{n} p_{ij}d_{is}\)

where \(d_{is}\) is the \(sth\) attribute value of \(ith\) data point, \(d_i\) and \(C_{js}\) is the \(sth\) attribute value of \(jth\) cluster.

Steps 2 and 3 are continued until the cost function stabilizes.

One of the important points of this algorithm is that there is no hard cluster membership \((0, 1)\) of each data point. Each data point can have memberships to each cluster. Experimental results suggest that KHM clustering algorithm is very robust to initial cluster centers and it converges faster than K-means clustering algorithm when the initialization is far from a local optimal. We will use the cost function of KHM clustering algorithm to develop a clustering algorithm for mixed datasets that will be robust to the selection of initial cluster centers. In the proposed algorithm, we will use the distance measure proposed for mixed data [10]. In the next section, we will discuss this distance measure.

2.2. The distance measure for mixed datasets [10]

Ahmad and Dey [10] propose a distance measure that works well for mixed datasets. This distance measure has shown promising results for clustering [10,11,14] and other applications [24]. This distance measure uses the distances between categorical attribute values. Weights of the numeric attributes are also used in this distance measure. The proposed distance measure with \(m\) attributes, in which \(m_r\) attributes are numeric whereas \(m_c\) attributes are categorical is presented in Eq. (2.2). It has two components; the first component computes the distance for numeric attributes and the second component for categorical attributes.

\[
\delta(d_i, C_j) = \frac{m}{\sum_{t=1}^{m} \Omega(d_i^t, C_j^t)}^{2/3}
\]

(2.2)

\(d_i^t\) is the \(ith\) numeric attribute value of data point \(d_i\) and \(d_i^c\) is the \(ith\) categorical attribute value of data point \(d_i\). Here \(C_j^t = (C_{j1}, C_{j2}, ..., C_{jm})\) represents the cluster centre for \(jth\) cluster. \(C_j^c\) represents the mean of \(rth\) numeric attribute for \(jth\) cluster. \(C_j^c\) represents the centre representation for \(jth\) categorical attributes for \(jth\) cluster. \(\Omega\) is the distance between a categorical attribute value and a cluster center for a categorical attribute.

In the first component, \(w_t\) is the weight of the \(rth\) numeric attribute. The second component uses a novel cluster representation and distances between categorical attribute values [10]. One of the most important steps in the algorithm is the calculation of distances between two categorical attribute values. The distance between two attribute values of a categorical attribute is calculated by computing the co-occurrence of these attribute values with attribute values of other categorical attributes. This algorithm has following steps:

1. Discretize all the numeric attributes to make the dataset purely categorical. This step is required to compute the distances between each pair of attribute values of every categorical attribute and the weight of each numeric attribute. We would like to note that while calculating distances the original attribute values are used.

2. For every categorical attribute, distances between every pair of attribute values are calculated. The distance between attribute values, \(x\) and \(y\), of categorical attribute \(A_t\) with respect to attribute categorical \(A_j\), for a given subset \(W\) of attribute \(A_t\) values, is defined by \(\delta(W)\) and computed by using following formula;  

\[
\delta(W) = P(W|x) + P(W|\neg x) - 1
\]

(2.3)

\(P(W|x)\) is the probability estimates of data points with attribute value equal to \(x\) belong to a class contained in \(W\). Probability estimates are computed by computing the frequencies of pairs of attribute values and class values.

\(P(W|\neg x)\) is the probability estimates of data points with attribute value equal to \(y\) belong to a class not contained in \(W\). Probability estimates are computed by computing the frequencies of pairs of attribute values and class values.

The distance, \(\delta(W, x, y)\), between \(x\) and \(y\) with respect to attribute \(A_j\) is given by \(\delta(W, x, y) = P(W|x) + P(W|\neg x) - 1\), where \(\omega\) is the subset of values of attribute \(A_t\) that maximizes the quantity \(P(W|x) + P(W|\neg x) - 1\). This distance is calculated against every other attribute. The mean of these distances, \(\delta(x, y)\), is taken as the distance between \(x\) and \(y\) in the dataset. The similar exercise is done for all the pairs of values of each categorical attribute.

3. The weight of each numeric attribute is computed. This computation has following steps for each numeric attribute;

(i) The discretized numeric attribute is treated as a categorical attribute. The distance between each pair of attribute values is calculated by the method discussed in step 2.

(ii) The means distance of all the pairs of attribute values is used as the weight of the numeric attribute.

4. Cluster centers are computed and then the distances between every data point and cluster centers are calculated. In the representation of a cluster center, a numeric attribute is presented by the means of member data points, whereas, a new definition was presented for categorical attributes. The cluster center for cluster \(C\) for
where $\theta_{i,k,c}$ is related to data points which have the $k$th attribute value for the $i$th attribute (it is assumed that $i$th attribute can have $p$ distinct values). We can compute $\theta_{i,k,c}$ in following way.

If $k$th attribute value $= x$,

$$\theta_{i.k,c} = \sum_{i=1}^{n} \eta(x, C) d_i \text{ (for cluster C)}$$  

where $\eta(x, C) = p_{ij}$ for data points which have $i$th attribute value $= x$, $\eta(x, C) = 0$ for data points which have $i$th attribute value $\neq x$. $p_{ij}$ is calculated as suggested in KHM clustering algorithm (discussed in Section 2.1).

3. The proposed K-Harmonic means (KHM) type clustering algorithm for mixed datasets

We use the distance measure discussed in previous section [10] with the cost function [7] of KHM clustering algorithm to propose a clustering algorithm for mixed datasets. The proposed algorithm uses the following cost function, $\zeta$, for $k$ number of desired clusters.

$$\zeta = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \delta(d_i, C_j)^a}$$

where $C_j$ is the nearest cluster center of data point $d_i$, and $\delta(d_i, C_j)$ is a distance measure proposed by Ahmad and Dey [10] with modified centers. $a$ is a positive number. One of the main differences between the proposed algorithm and the algorithm proposed by Ahmad and Dey [10] is the computation of the cluster centers. Ahmad and Dey [10] use hard cluster memberships for computing cluster centers. However, in the proposed algorithm similar to KHM clustering algorithm, soft memberships of data points will be used to compute the cluster centers. For the numeric attributes, the computation will be similar as KHM clustering algorithm; however, for categorical attributes the computation will be done by the following procedure.

In the proposed algorithm, for a categorical attribute, the cluster center is described by the proportional distribution of categorical values in a cluster. Using the definition of the center of cluster in fuzzy clustering algorithms for categorical datasets [25, 26], the center for a cluster $C$ (the $j$th cluster for $i$th categorical data) is represented as:

$$\theta_{i,j.k,c} = \frac{\sum_{i=1}^{n} \eta(x, C) p_{ij} \text{ (for cluster C)}}{N_c}$$

where $\eta(x, C) = p_{ij}$ for data points which have $i$th attribute value $= x$, $\eta(x, C) = 0$ for data points which have $i$th attribute value $\neq x$. $p_{ij}$ is calculated as suggested in KHM clustering algorithm (discussed in Section 2.1).

### Table 1

<table>
<thead>
<tr>
<th>Toy Data</th>
<th>Attribute F</th>
<th>Attribute G</th>
<th>$p_{ij}$ for cluster C, the $j$th cluster is C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\alpha$</td>
<td>4.0</td>
<td>0.10</td>
</tr>
<tr>
<td>2</td>
<td>$\alpha$</td>
<td>8.1</td>
<td>0.20</td>
</tr>
<tr>
<td>3</td>
<td>$\beta$</td>
<td>9.0</td>
<td>0.30</td>
</tr>
<tr>
<td>4</td>
<td>$\alpha$</td>
<td>5.0</td>
<td>0.25</td>
</tr>
<tr>
<td>5</td>
<td>$\beta$</td>
<td>3.0</td>
<td>0.15</td>
</tr>
</tbody>
</table>

A toy dataset will be used to show the calculation of a cluster center and distance between a data point and a cluster center. There are 5 data points represented by two attributes in the dataset. Attribute $F$ is a categorical attribute which has two attribute values $\alpha$ and $\beta$. The attribute $G$ is a numeric attribute. The calculation is shown for the center of cluster $C$.

For attribute value $\alpha$, the first attribute value of attribute $F$ (the first attribute), $\theta_{11,c}$ is calculated in following manner, $\theta_{11,c} = 0.10 + 0.20 + 0.25 = 0.55$

Whereas For attribute value $\beta$, the second attribute value of attribute $F$ (the first attribute), $\theta_{12,c}$ is calculated in following manner,

$\theta_{12,c} = 0.30 + 0.15 = 0.45$

For numeric attributes, Euclidean distance with the weight of each numeric attribute is used for computing distances between data points and cluster centers. But for categorical attributes, the proposed center and the distances between attribute values are used. The distance between an attribute value $X$ and center of cluster $C$ in the direction of attribute $F$ can be calculated as (using Eq. (2.5))

$$\Omega(X, C) = \theta_{11,c} \delta(X, A_1) + \theta_{12,c} \delta(X, A_2) + \theta_{13,c} \delta(X, A_3)$$

where $A_i$ is the $i$th attribute value of $i$th categorical attribute.

The similar calculation will be done for all the categorical attributes. The total distance is calculated by using Eq. (2.2)

For the dataset given Table 1, the distance between data point 1 (attribute value $\alpha$ for categorical attribute $F$) and center of cluster $C$ in the direction of the categorical attribute $F$ is calculated as:

$$\Omega(\alpha, C) = \theta_{11,c} \delta(\alpha, \alpha) + \theta_{12,c} \delta(\alpha, \beta) + \theta_{13,c} \delta(\alpha, \beta)$$

$$\Omega(\alpha, C) = 0.55 \delta(\alpha, \alpha) + 0.45 \delta(\alpha, \beta)$$

The distance between data point 1 (attribute value 4.0 for categorical attribute $G$) and center of cluster $C$ in the direction of the numeric attribute $G$ is calculated as:

Centre = $4.0 \times 0.10 + 8.1 \times 0.20 + 9.0 \times 0.30 + 5.0 \times 0.25 + 3.0 \times 0.15 = 6.42$

Distance = $\sqrt{w_C \times (4.0 - 6.42)^2}$

where $w_C$ is the significance of the attribute $G$.

Eqs. (3.5) and (3.6) are combined by using Eq. (2.2) to calculate the distance between data point 1 and cluster center $C$. KHM clustering algorithm has two steps; the computation of cluster centers and the computation of cluster membership by using the distances between data points and cluster centers. In our algorithm, we use cluster centers as discussed above and the distance measure proposed by Ahmad and Dey [10]. Remaining steps of KHM clustering algorithm are followed as discussed in Section 2.1. The proposed KHM type clustering (C) algorithm for mixed (M) datasets (D) (KHMCMDO) is presented in Fig. 3.
Ahmad_and_Dey_Clustering_Algorithm
Input – A mixed dataset with have \( n \) number of data points described by \( m \) attributes.
Output - \( k \) clusters

1. Normalize all numeric attributes.
2. Discretize all numeric attribute.
3. For each categorical attribute
   (i) Compute the distances between each pair of attribute values.
4. For each numeric attribute
   (i) Consider the discretized attribute as categorical attribute and compute the distances between each pair of attribute values.
   (ii) The average of all the distances (between each pair of attribute values) is taken as the weight of the numeric attribute.

Take the original data with normalized numeric attributes. Assign data points to \( k \) clusters randomly.
Repeat steps \( A \) and \( B \)
(A) Calculate cluster centers for all the clusters (suggested in Section 2.2).
(B) Each data point is assigned to its closest cluster by using the distance measure (discussed in Section 2.2).
Until no data point changes cluster membership or a pre-defined number of runs are reached.

end.

Fig. 1. The clustering algorithm proposed by Ahmad and Dey [10].

3.1. The complexity of KHMCMD algorithm

The proposed clustering algorithm has two parts: the computation of distances between every pair of attribute values for categorical attributes and discretized numeric attributes, and iteration process in which we calculate distances between all data points \((n \) data points with \( m \) attributes\) to all the cluster center. \(O(m^2n + m^2S^3)\) steps are required to compute the distances between every pair of attribute values for each attribute where \( S \) is the means of numbers of attribute values of categorical attributes. For each iteration, distances between each data point and each cluster center are calculated. \(O(nKm_1 + nKm_2S)\) steps are required for this calculation. For \( i \) iterations, the complexity of KHMCMD is \(O(m^2n + m^2S^3 + un(Km_1 + Km_2S))\). This suggests that KHMCMD has linear computational complexity to number of data points.

4. Results and discussion

The proposed initialization method was tested on three pure categorical dataset and three mixed datasets taken from UCI Machine Learning Repository [27]. The information about these datasets are presented in Table 2. In our experiments, Equal-width discretization [28] was used to discretize the numeric attributes for the proposed algorithm and the algorithm proposed by Ahmad and Dey [10]. The number of bins was set to 10. Data points of each dataset were distributed in predefined classes. The class information was taken as ground truth. The clustering results were compared against the ground truth. When clusters are more than classes, the clusters were assigned to a class such that the clustering accuracy was the maximum. Zhang [7] shows that \( a \geq 2 \) gives good results. \( a = 4 \) was selected for all the experiments. We used two measures to compare the results of clustering: Average Accuracy and Standard Deviation. Assume that a dataset contains \( K \) classes and total \( n \) data points, let \( c_i \) \((i = 1,..K)\) be the data points correctly assigned to class \( c_i \), \((i = 1,..K)\) then Accuracy \((AC)\) in\% is defined as:

\[
AC = \frac{\sum_{i=1}^{K} c_i}{n} \times 100
\]

If an algorithm is executed \( T \) times, then Average Accuracy in\% is calculated as:

\[
Average\ Accuracy\ (in\%) = \mu = \frac{1}{T} \sum_{i=1}^{T} AC_i
\]

\(AC_i\) is the Accuracy in\% in ith run.
And, Standard Deviation is defined as:

\[
Standard\ Deviation = \sqrt{\frac{1}{T} \sum_{i=1}^{T} (AC_i - \mu)^2}
\]

The Average Accuracy in\% exhibits how good the cluster results are against the ground truth. The higher value is desirable. The maximum value of Average Accuracy in\% can be 100. Standard Deviation shows the consistency of the clustering results in different runs. A low value of Standard Deviation is desirable. The minimum value of Standard Deviation is 0. As different runs of clustering algorithm start with different initial cluster centers, the low value of Standard Deviation (SD) shows that clustering algorithm is quite robust to the selection of initial cluster centers.

4.1. The comparative study against the clustering algorithm proposed by Ahmad and Dey [10]

The clustering algorithm proposed by Ahmad and Dey [10] has shown good results for various pure categorical datasets and mixed datasets. Hence, we used that clustering algorithm for the comparative study. We call this algorithm KMCMD (K-Means type clustering algorithm for mixed datasets). Our proposed clustering algorithm, KHMCMD, uses the concepts of KMCMD [10], this is another reason
for taking that clustering algorithm for comparative study. Random initialization of cluster centers was used for both the algorithms. Experiments were carried out with the number of desired clusters as the number of classes. Experiments were also carried out with 3, 5, 7 number of clusters. There was no specific reason for selecting these numbers. The experiments were done to study the performance of two clustering algorithms at different number of clusters. With each clustering setting, our proposed KHMCMD algorithm and KMCMD [10] algorithm were run 10 times and Average Accuracy in% and Standard Deviation are presented.
KHMCMD

Input – A mixed dataset with have \( n \) number of data points described by \( m \) attributes.

Output: K clusters

5. Normalize all numeric attributes.
6. Discretize all numeric attribute.
7. For each categorical attribute

Calculate the distances between every pair of attribute values.

8. For each numeric attribute

(i) Consider the discretized attribute as categorical attribute and compute the distances between every pair of attribute values.

(ii) The average of all the distances (between every pair of attribute values) is taken as the weight of the numeric attribute.

9. Take the original data with normalized numeric attributes. Assign data points to \( \lambda \) clusters randomly. Compute the center of the cluster by using the method suggested in Section 3 with hard membership.

Run following steps till the data points don’t change their clusters or a predefined number of iterations have been finished.

(i) Compute the distance between each data point to all the cluster centers as discussed in Section 3.

(ii) The nearest cluster center is used to allocate data points to a cluster. This step is used only to check whether data points are changing their clusters or not.

(iii) Compute the membership of each data point to each cluster using the method suggested in K-Harmonic mean clustering algorithm (discussed in Section 2.1).

(iv) Compute the centers as discussed in Section 3.

Fig. 3. KHMCMD clustering algorithm.

Table 2
Description of the datasets used in the experiments.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Data points</th>
<th>Attributes</th>
<th>Classes</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>Categorical</td>
<td>699</td>
<td>9</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>Categorical</td>
<td>32</td>
<td>56</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Mushroom</td>
<td>Categorical</td>
<td>8124</td>
<td>22</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Heart-1</td>
<td>Mixed</td>
<td>303</td>
<td>8</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Heart-2</td>
<td>Mixed</td>
<td>270</td>
<td>7</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Credit</td>
<td>Mixed</td>
<td>690</td>
<td>8</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3
Average Accuracy (in%) and Standard Deviation for Breast Cancer dataset for KHMCMD and KMCMD with varying number of clusters.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of Clusters</th>
<th>Average Accuracy (in%)</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>KHMCMD</td>
<td>2</td>
<td>96.3</td>
<td>0.00</td>
</tr>
<tr>
<td>KMCMD</td>
<td>3</td>
<td>96.1</td>
<td>0.10</td>
</tr>
<tr>
<td>KHMCMD</td>
<td>3</td>
<td>96.3</td>
<td>0.00</td>
</tr>
<tr>
<td>KMCMD</td>
<td>5</td>
<td>96.3</td>
<td>0.12</td>
</tr>
<tr>
<td>KHMCMD</td>
<td>7</td>
<td>97.2</td>
<td>0.39</td>
</tr>
<tr>
<td>KMCMD</td>
<td>7</td>
<td>96.3</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Table 4
Average Accuracy (in%) and Standard Deviation for Lung Cancer dataset for KHMCMD and KMCMD with three clusters.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of Clusters</th>
<th>Average Accuracy (in%)</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>KHMCMD</td>
<td>3</td>
<td>61.1</td>
<td>0.028</td>
</tr>
<tr>
<td>KMCMD</td>
<td>3</td>
<td>52.6</td>
<td>0.052</td>
</tr>
</tbody>
</table>

results show that Average Accuracy of KHMCMD is quite similar to KMCMD. Considering Standard Deviation, KHMCMD means always exhibited a lower value. This shows that the proposed K-Harmonic clustering algorithm is producing consistent clustering results.

4.1. Breast cancer dataset

This is a pure categorical dataset. Missing attribute values were replaced by the mode of that attribute. We executed KMCMD algorithm [10] and the proposed KHMCMD algorithm with varying number of clusters (2, 3, 5 and 7). With each cluster setting, both algorithms were executed ten times. Table 3 and Fig. 4 show the results in the form of Average Accuracy and Standard Deviation.

4.1.2. Lung cancer dataset

This is a pure categorical dataset. Missing attribute values were replaced by the mode of the attribute. The number of clusters was taken as three which is equal to the number of classes. As the number of data points was small (32), we did not do experiments with more number of clusters. Table 4 and Fig. 5 show the Average Accuracy (in%) and Standard Deviation. The results indicate that
KHMCMD performed better than KMCMD due to higher accuracy and much lower standard deviation.

### 4.1.3. Mushroom dataset

This is a pure categorical dataset. Missing values were replaced by the mode of the attribute. We executed KMCMD and KHMCMD algorithms on this dataset with varying number of clusters (2, 3, 5 and 7). With each cluster setting, both the algorithms were executed ten times. Table 5 and Fig. 6 show the results in the form of Average Accuracy (in%) and Standard Deviation. The results show that the Average Accuracy of KHMCMD is quite similar to KMCMD for two settings (5 and 7 clusters) whereas with two settings (2 and 3 clusters) KHMCMD performed better than KMCMD. Considering Standard Deviation, KHMCMD exhibited superior consist performance than KMCMD by always resulting in a much lower Standard Deviation. The lower Standard Deviation suggests that KHMCMD is not affected much by initial cluster centers.

### 4.1.4. Heart-1 dataset

This is a mixed dataset. We executed KHMCMD and KMCMD algorithms on this dataset with varying number of clusters (2, 3, 5 and 7). Table 6 and Fig. 7 show the results. These results show that Average Accuracy of KHMCMD is quite similar to that of KMCMD. Considering Standard Deviation, KHMCMD exhibited good performance than KMCMD by mostly resulting in a much lower Standard Deviation. That shows the robustness of KHMCMD for the different initial cluster centers.

### 4.1.5. Heart-2 dataset

This is a mixed dataset. We executed KHMCMD and KMCMD algorithms on this dataset with varying number of clusters (2, 3, 5 and 7). With each cluster setting, both the algorithms were executed ten times. Table 7 and Fig. 8 show the clustering results. The results show that KHMCMD exhibited similar or higher Average Accuracy than KMCMD. Considering Standard Deviation, KHMCMD exhibited good performance than KMCMD by always resulting in a much lower Standard Deviation value except in case of 7 clusters where it showed a slightly higher Standard Deviation value.

![Fig. 4. Plots for Average Accuracy (in%) and Standard Deviation for Breast Cancer dataset for KHMCMD and KMCMD with varying number of clusters.](image-url)

![Fig. 5. Plots of Average Accuracy (in%) and Standard Deviation for Lung Cancer dataset for KHMCMD and KMCMD with three clusters.](image-url)

**Table 5**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of Clusters</th>
<th>Average Accuracy (in%)</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>KHMCMN</td>
<td>2</td>
<td>89.32</td>
<td>0.02</td>
</tr>
<tr>
<td>KMCMD</td>
<td></td>
<td>81.16</td>
<td>0.04</td>
</tr>
<tr>
<td>KHMCMN</td>
<td>3</td>
<td>86.93</td>
<td>0.07</td>
</tr>
<tr>
<td>KMCMD</td>
<td>5</td>
<td>73.86</td>
<td>2.47</td>
</tr>
<tr>
<td>KMCMD</td>
<td>7</td>
<td>89.46</td>
<td>2.15</td>
</tr>
<tr>
<td>KMCMD</td>
<td></td>
<td>89.50</td>
<td>0.02</td>
</tr>
<tr>
<td>KMCMD</td>
<td></td>
<td>91.38</td>
<td>3.06</td>
</tr>
</tbody>
</table>

**Table 6**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of Clusters</th>
<th>Average Accuracy (in%)</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>KHMCMN</td>
<td>2</td>
<td>84.09</td>
<td>0.15</td>
</tr>
<tr>
<td>KMCMD</td>
<td></td>
<td>83.89</td>
<td>0.15</td>
</tr>
<tr>
<td>KHMCMN</td>
<td>3</td>
<td>80.46</td>
<td>0.28</td>
</tr>
<tr>
<td>KMCMD</td>
<td>5</td>
<td>80.73</td>
<td>1.33</td>
</tr>
<tr>
<td>KMCMD</td>
<td>7</td>
<td>81.72</td>
<td>0.50</td>
</tr>
<tr>
<td>KMCMD</td>
<td></td>
<td>80.79</td>
<td>1.77</td>
</tr>
<tr>
<td>KMCMD</td>
<td></td>
<td>81.85</td>
<td>0.57</td>
</tr>
<tr>
<td>KMCMD</td>
<td></td>
<td>79.34</td>
<td>2.02</td>
</tr>
</tbody>
</table>

**Table 7**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of Clusters</th>
<th>Average Accuracy (in%)</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>KHMCMN</td>
<td>2</td>
<td>81.63</td>
<td>0.33</td>
</tr>
<tr>
<td>KMCMD</td>
<td></td>
<td>80.74</td>
<td>1.20</td>
</tr>
<tr>
<td>KHMCMN</td>
<td>3</td>
<td>78.74</td>
<td>1.27</td>
</tr>
<tr>
<td>KMCMD</td>
<td>5</td>
<td>78.59</td>
<td>3.22</td>
</tr>
<tr>
<td>KMCMD</td>
<td></td>
<td>80.15</td>
<td>1.50</td>
</tr>
<tr>
<td>KMCMD</td>
<td>7</td>
<td>77.41</td>
<td>2.13</td>
</tr>
<tr>
<td>KMCMD</td>
<td></td>
<td>80.00</td>
<td>1.64</td>
</tr>
<tr>
<td>KMCMD</td>
<td></td>
<td>79.19</td>
<td>1.40</td>
</tr>
</tbody>
</table>
The results show the KHMCMD produces more consistent results as compared to KMCMD.

4.1.6. Credit dataset
This is a mixed dataset. We executed KHMCMD and KMCMD algorithms on this dataset with varying number of clusters (2, 3, 5 and 7). With each cluster setting, both the algorithms were executed ten times. Table 8 and Fig. 9 show the average accuracy in % and SD values. These results show that the Average Accuracy of KHMCMD is quite similar to KMCMD, with all settings. Considering Standard Deviation, KHMCMD exhibited good performance than KMCMD by always resulting in a lower Standard Deviation value. Lower values of Standard Deviation show the robustness of KHMCMD against cluster centers initialization problem.

4.1.7. Summary of the results
We present the results for all the datasets in Table 9. For better comparative study, results are presented only for the cases in which the number of clusters is same as the number of classes.
The results suggest that in these cases KHMCMD performs better than KMCMD in terms of higher Average Accuracy and lower Standard Deviation. In other words, KHMCMD can produce accurate and consistent clustering results.

### 4.2. Comparison with K-modes cluster center initialization approaches for categorical datasets

K-modes clustering algorithm is a popular algorithm used for clustering categorical datasets [16]. As discussed in Section 2, it also suffers for cluster centers initialization problem. Various other cluster center initialization approaches have been proposed for K-modes algorithm like, Cao et al. [20], Wu et al. [22] and Khan and Ahmad [21]. We compared the performance of K-modes algorithm with various cluster center initialization approaches against the performance of our proposed algorithm. For K-modes with random initialization method and the proposed KHMCMD the average results of ten runs are presented. Results are shown in Table 10. It can be observed that the proposed approach outperforms other methods. It is noted that other than random initialization and Wu et al. [22] methods, all other initialization methods presented in the study give same clustering results in different runs, whereas the proposed method may not give unique clustering result.

### 4.3. The comparative study with other mixed data clustering algorithms for mixed datasets

We also compared the clustering results with other mixed data clustering algorithm (K-prototype algorithm [8], Similarity-Based Agglomerative Clustering [29], Improved K-prototype clustering algorithm [23]) and Ji et al. [13] clustering initialization method when the number of clusters was two. Hung's clustering algorithm [8] was implemented and clustering results are presented for this algorithm. We ran this algorithm ten times and average results are presented. Results for other clustering results are taken from literature [13,23,29], if available in the literature. Results are presented in Table 11. Results suggest that the proposed method performed better than other clustering algorithms for mixed datasets.

### 4.4. Discussion

As described above, we executed KHMCMD and KMCMD algorithms on six datasets. Three datasets were pure categorical datasets and three were mixed datasets. Ideally, a K-means type clustering algorithm which is quite robust to the cluster center initialization problem should have high Average Accuracy and low Standard Deviation for different runs with different initial cluster centers. It can be noted that generally KHMCMD algorithm exhibits similar or better accuracy against KMCMD algorithm. KHMCMD also shows low Standard Deviation in almost all the cases. These results suggest that KHMCMD is quite insensitive to cluster center initialization problem.

We also compared the results with other clustering algorithms for categorical datasets and mixed datasets. KHMCMD performed better than all the algorithms in term of clustering accuracy. This shows that KHMCMD give accurate clustering results.

### 5. Conclusion and future work

K-means type clustering algorithms for mixed datasets suffers from cluster center initialization problems as different initialization cluster centers give different clustering results. K-Harmonic means clustering algorithm try to address the cluster center initialization problem. K-Harmonic means clustering algorithm is quite insensitive to the choice of initial cluster centers [7]. However, the original K-Harmonic means clustering algorithm works only for numeric

---

### Table 10
Performance of the proposed algorithm against other cluster center initialization algorithms for categorical datasets. Results are presented in clustering accuracy in%. The bold number in a row shows the best performance for the dataset.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>83.64</td>
<td>91.13</td>
<td>91.13</td>
<td>91.27</td>
<td>96.30</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>52.10</td>
<td>50.00</td>
<td>50.00</td>
<td>50.00</td>
<td>61.10</td>
</tr>
<tr>
<td>Mushroom</td>
<td>72.31</td>
<td>87.54</td>
<td>87.54</td>
<td>88.15</td>
<td>88.66</td>
</tr>
</tbody>
</table>

### Table 11
The performance of the proposed KHMCMD against other mixed data clustering algorithms. (- shows results not available). The results are presented in clustering accuracy in%. The bold numbers in a row shows the best performance for the dataset.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart-1</td>
<td>61.3</td>
<td>75.2</td>
<td>82.6</td>
<td>80.8</td>
<td>84.1</td>
</tr>
<tr>
<td>Heart-2</td>
<td>60.8</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>81.6</td>
</tr>
<tr>
<td>Credit data</td>
<td>58.4</td>
<td>55.5</td>
<td>77.9</td>
<td>79.4</td>
<td>85.6</td>
</tr>
</tbody>
</table>
datasets. In this paper, we extended the K-Harmonic means clustering algorithm to mixed datasets. Distance measure proposed by Ahmad and Dey [10] and a new definition of cluster center are used with the K-Harmonic means clustering cost function. Clustering results with pure categorical datasets and mixed datasets show that the proposed clustering algorithm is quite insensitive to the choice of initial cluster centers. The results also suggest that the proposed clustering algorithm give better clustering results against other clustering algorithms. In future other implementations of K-Harmonic means clustering [30,31] may also be examined to achieve more optimized performance.

References

[16] Z. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining, DMKD (1997) 0.