International Journal of Machine Learning and Cybernetics (2021) 12:2069–2090
https://doi.org/10.1007/s13042-021-01293-w

ORIGINAL ARTICLE

A rough set based algorithm for updating the modes in categorical clustering

Semeh Ben Salem[1,2,3] · Sami Naouali[1] · Zied Chtourou[3]

Received: 12 November 2019 / Accepted: 5 March 2021 / Published online: 27 March 2021

Abstract
The categorical clustering problem has attracted much attention especially in the last decades since many real world applications produce categorical data. The *k*-mode algorithm, proposed since 1998, and its multiple variants were widely used in this context. However, they suffer from a great limitation related to the update of the modes in each iteration. The mode in the last step of these algorithms is randomly selected although it is possible to identify many candidate ones. In this paper, a rough density mode selection method is proposed to identify the adequate modes among a list of candidate ones in each iteration of the *k*-modes. The proposed method, called Density Rough *k*-Modes (DR*k*-M) was experimented using real world datasets extracted from the UCI Machine Learning Repository, the Global Terrorism Database (GTD) and a set of collected Tweets. The DRk-M was also compared to many states of the art clustering methods and has shown great efficiency.

Keywords Unsupervised learning · Categorical clustering · Rough set theory · K-modes · Uncertainty

1 Introduction

Cluster analysis, also known as unsupervised learning, is a branch of Machine Learning which has multiple applications in various domains, including financial fraud detection, medical diagnosis, image processing, information retrieval and bioinformatics [1]. Clustering is used to unlock initially hidden and undetectable patterns by identifying groupings within the dataset under investigation in an unsupervised manner: no labels (or classes) are initially provided as input parameters. This analysis process based on identifying the clusters without a prior knowledge of the outcomes makes this task more difficult, challenging and prone to errors. Clustering methods can be classified into five types: hierarchical [2, 3], partitional [4–16], density-based [17, 18], grid-based [19] or model-based methods [20]. The aim of cluster analysis is to partition a dataset composed of *N* observations embedded in a *d*-dimensional space into *k* distinct clusters. In this process, data points within the same cluster will have more similar characteristics than observations in other clusters according to a specific measure called distance metric [21].

Clustering qualitative data is problematic due to the lack of its geometric properties. For example, categorical attributes are unordered and it is inappropriate to use traditional numerical distance functions to capture resemblance between these categorical values. To overcome this limitation, many methods were proposed to deal with categorical data types such as the *k*-modes and its variants [4–16]. These methods start with K initial centroids and use the alternating minimization method to solve a non convex optimization problem. In the *k*-modes, the simple matching dissimilarity measure is used to compare two categorical values: the comparison yields a difference of zero for two identical values and one otherwise. However, for all these methods, one main issue is related to the identification of the initial number of clusters [22].

The mode, representing the most frequent patterns (modality in each attribute) in the cluster is randomly

✉ Semeh Ben Salem
semehbensalem0@gmail.com

Sami Naouali
snaouali@gmail.com

Zied Chtourou
ziedchtourou@gmail.com

[1] Science and Technologies for Defense (STD) Laboratory, Military Academy of Fondouk Jedid, Nabeul, Tunisia

[2] Polytechnic School of Tunisia, Rue El Khawarizmi, Al Marsá, B.P. 743, 2078 Tunis, Tunisia

[3] Military Research Center, Aouina Military Base, Cité Taieb Mhiri, 2045 Tunis, Tunisia

selected in the last step of the clustering process in all these methods. However, it is possible to identify more than one mode depending on the modalities' frequency in the attributes. In the *k*-modes type clustering algorithms' iterative process, the mode is also identified when moving from the *i*th iteration to the $(i+1)$th iteration. The selection of this mode is essential and directly influences the formation of the final clusters. Frequently, random modes selection, widely used in the literature, may induce a clustering process to terminate in a locally optimal solution. Thus, this method is not convenient to ensure high performance. On the other hand, tackling the mode's random selection issue during the clustering process was not considered in previous methods and thus, is an interesting issue to consider.

As it is widely known, in reality, the border of the data is hard to partition, as there is often no sharp boundary between the clusters. Most of the proposed formal modeling tools are deterministic and precise which does not fit real world situations that are very often not deterministic and cannot be described precisely. This fact requires integrating uncertainty based models in the clustering process. The first attempts to handle uncertainty were proposed with fuzzy theory such as the fuzzy *k*-means [23], the fuzzy *k*-modes [24] and their variants. In fuzzy clustering, each object can have membership functions to more than one cluster instead of the hard assignment given in the k-modes based methods. However, fuzzy algorithms have the same limitations as hard algorithms, i.e. they require multiple runs with different centroid initializations to ensure stability. Fuzzy methods also need to adjust one control parameter for the membership fuzziness to obtain better solutions, which is a complex task even through extensive experiments. Another attempt to handle uncertainty and avoid the fuzzy sets limitations was by using the Rough Set Theory (RST), introduced by Pawlak [25]. One main reason for the success of the RST is that no additional information is required to start the clustering process, such as thresholds or expert knowledge in a particular domain. In recent years, RST has attracted much attention in a variety of fields [26] such as computer vision [27, 28], biomedical engineering [29] and economy and finance [30].

In this paper, it is proposed to tackle the mode identification in categorical clustering using an uncertainty based model. The proposed method, called Density Rough *k*-Modes (DR*k*-M) aims to select the most appropriate modes in each iteration during the clustering process. This method can be implemented either for the *k*-modes or any of its variants. The rough mode selection permits identifying the most central centroid for each cluster and thus ensures better clustering performance. To better illustrate this notion and the main differences between the DR*k*-M and states of the art methods, Fig. 1 is proposed.

In Fig. 1, the input is a categorical dataset composed of *N* observations described by *d* attributes. Although many previous studies considered the issue of the selection of the initial number of clusters [22], these methods do not fall under the scope of this study. In Fig. 1, a description of the clustering process in the k-modes and its variants is given. In the first phase, K initial centroids are randomly selected. These centroids will correspond to the starting point of the partitioning process. However, this random selection may usually lead to incorrect results since elements that are supposed to be part of the same cluster can be selected as centroids and thus put in different clusters. Many initialization methods were proposed in the literature to choose the most representative initial centroids [8, 35–37, 42]. In the second step of the k-modes, the (N-K) remaining observations will

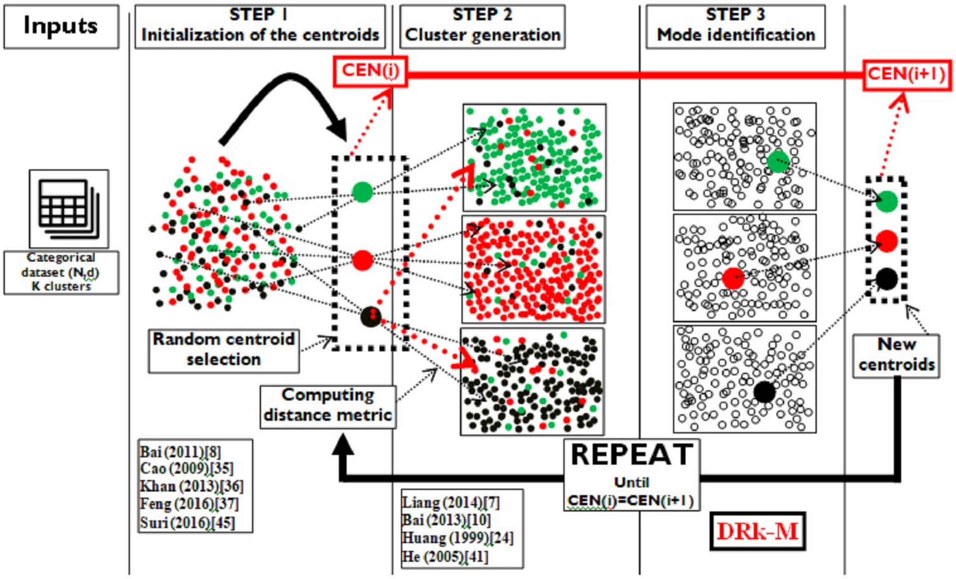


Fig. 1 The k-modes algorithm and its variants compared to the DRk-M

be assigned to the clusters according to the value of similarity computed using the simple matching dissimilarity metric between these points and the centroids. In this step, some variants of the k-modes, proposed using other distance metrics to enhance the algorithm [7, 10, 24, 38]. Once all the observations are assigned to their corresponding clusters, the centroids are updated in the last step of the algorithm and the new mode is computed for each group. The DR*k*-M proposes to tackle this last centroid updating step. As shown in Fig. 1, no previous method was presented in the literature to investigate this issue.

The DRk-M has some contributions and characteristics that can be summarized as follows:

- The DRk-M is a categorical clustering approach. Categorical clustering is a tricky subject since it deals with categorical data characterized by their complex form.
- The random selection of the mode is a critical issue in categorical clustering. Avoiding this limitation permits obtaining more accurate and stable results.
- Since more than one possible mode can be generated in a given cluster in each iteration, the RST can be efficiently used in this context to bring more certainty and accuracy to the final results.
- The convergence, performance and scalability of the DRk-M under the new mode selection method were investigated.

This paper is organized as follows: In Sect. 2, categorical clustering is detailed. Some states of the art categorical clustering methods are given and classified according to their contribution compared to the k-modes. In Sect. 3, the DRk-M is detailed and the building theories and concepts are given. The algorithm and a complexity analysis are also given in this section. In Sect. 4, an experimental analysis is provided using several datasets and the DRk-M is compared to many states of the art algorithms. Finally, in the last section, conclusions and perspectives are provided.

2 State of the art of categorical clustering

2.1 Clustering based on prototypes

A cluster is commonly characterized using a centroid that measures its centrality. Clustering methods based on prototypes are usually called partitional methods. Initially the number of clusters K is required as an input parameter and the centroids are iteratively updated until reaching a stop criterion. One of the prototypes based algorithms proposed in the literature is the *k*-means and its multiple variants [5, 6]. However, because large categorical data sets exist in many applications, such as environmental data analysis [31], market basket data analysis [32], DNA or protein sequence analysis [33], text mining [34], it was not possible to use the k-means in such application types. Thus, developing more appropriate algorithms to handle categorical clustering was an interesting subject in the last twenty years.

Partitional methods, consecutively divide the dataset until finding K clusters with a static configuration where no further splitting is possible. The partitional process can be described as follows:

Let $CEN=\{Cen_1,Cen_2,\ldots,Cen_k\}$ be the set of centroids for the clusters $\mathcal{C}_j$ where $j=1,\ldots,K$ and $X=\{obs_1, obs_2,\ldots,obs_N\}$ the initial dataset with size N. For each observation $obs_i \in X$.

Step 1: assign obs_i to the cluster $\mathcal{C}_j$ where Cen_j verifies: $Cen_j=argmin_{j=1,\ldots,K}\{d(obs_i,Cen_j\}$, where d is a similarity metric.

Step 2 : update the centroid C_j according to a given procedure depending on the data type involved (mean, median, modes).

Step 3: if the stopping criteria are met, the process will end otherwise repeat starting from step 1.

2.2 Categorical clustering: the *k*-modes and its variants

A categorical information system can be described as a quadruple IS = (U, A, V, f), where:

(1) $U=\{x_1,x_2,\ldots,x_N\}$ is the nonempty set of N data points, called a universe;

(2) $A=\{a_1,a_2,\ldots,a_d\}$ is the nonempty set of d categorical attributes;

(3) V is the union of attribute domains, i.e., $V = U_{j=1}^{d} V_{a_j}$, where $V_{a_j} = \left\{a_j^1, a_j^2, \ldots, a_j^{(n_j)}\right\}$ is the value domain of categorical attribute a_j and is finite and unordered, e.g., for any $1 \le p \le q \le n_j$, either $a_j^p = a_j^q$ or $a_j^p \ne a_j^q$. Here, n_j is the number of categories of attribute a_j for $1 \le j \le d$;

The *k*-modes [4] is based on optimizing a cost function given in Eq. (1):

$$\mathrm{P(W,Q)} = \sum_{l=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{d} \omega_{il}\mathcal{D}(\mathrm{obs_{ij}}, \mathrm{cen_{lj}}) \tag{1}$$

Q = $\{cen_1, cen_2, \ldots, cen_K\}$ represents the set of cluster modes and $W=[w_{ji}]$ is a {0,1} matrix that corresponds to the current membership of an observation obs_i to be part of a given cluster. This matrix verifies the rules given in Eqs. (2) and (3):

$$\sum_{j=1}^{K} \mathrm{w_{ji}} = 1 \tag{2}$$

$$0 < \sum_{i=1}^{N} w_{ji} < N \tag{3}$$

The *k*-modes permits clustering categorical datasets according to three modifications:

(i) The simple matching dissimilarity measure $\mathcal{D}$ is considered to evaluate the similarity between obs_j and cen_j

$$\mathcal{D}\left(cen_j, obs_j\right) = \sum_{j=1}^{d} \delta(cen_j, obs_j) \tag{4}$$

where

$$\delta\left(cen_j, obs_j\right) = \begin{cases} 0, \text{ if} cen_j = obs_j \\ 1, \text{ if} cen_j \neq obs_j \end{cases} \tag{5}$$

(ii) The centroids are called modes where a mode of a categorical dataset *U* described by *d* attributes is a vector $Q = [q_1, q_2, \ldots, q_d]$ that minimizes the quantity defined in Eq. (6):

$$L(U,Q) = \sum_{i=1}^{N} D\left(obs_i, Q\right) \tag{6}$$

The mode *Q* represents the set of the most frequent modalities in each attribute.

(iii) using a frequency-based method to update the modes during the clustering process to minimize the cost function.

Many variants of the *k*-modes were proposed to improve its efficiency and scalability according to many perspectives. In the rest of this section, these methods are detailed in order to provide the main research axis used.

2.2.1 Initialization methods

Most partitional categorical clustering methods such as the k-modes and its variants require pre-defining the number of clusters K as well as the selection of the initial centroids which is a great limitation. The initialization of the centroids can have a high impact on the final clustering results and various initializations can lead to several output clusters. Usually, selecting the initial clusters is random which is also problematic since one may select initial centroids that can have similar characteristics. Due to its simplicity, random initialization was widely used. In order to adjust the negative effects of the random initialization, these algorithms generally need to be rerun many times with different initializations [8, 35–37].

In [8], the authors proposed an initialization method based on the density and distance measures. In this method, the initial dataset is split into several subsets based on its attributes. Thus, it becomes possible to discard some data points from the potential set of initial centroids. Then, the most frequent attribute value is spotted in each attribute domain to compose its representative point and generate the centroid. The proposed method's computational cost is $O(2Nm|V| + |V| + mK^2|V|)$ which is linear with respect to the number of data points where $|V| = \sum_{i=1}^{m} n_j$, *m* is the number of categorical attributes and n_j the corresponding modalities. However, the method is not appropriate when the number of clusters is considerable.

In [36], the authors proposed an advanced method to identify prominent attributes that correspond to the dataset's most relevant attributes. A multiple clustering of data based on the attributes is then performed to spot interesting initial centroids. The method performs multiple clustering on different attributes in the original data space and uses distinct attribute values in an attribute as cluster labels. These multiple views provide new insights into the data's hidden structures to find consistent cluster structure and aid in computing better initial centroids. Three approaches were presented to select different attribute spaces that can help in generating different clustering views from the data, namely:

- Vanilla approach: this method considers all the attributes (m) present in the dataset.
- Prominent attributes: only a few attributes may be useful to generate multiple clustering views.
- Significant attributes: a set of attributes generated from the prominent attributes will be retained.

  In [37], the initialization of the centroids was considered from the view of outlier detection. Two different initialization algorithms were proposed: a distance-based called *Ini_Distance* and an entropy-based outlier detection technique called *Ini_Entropy* within the RST. These two distances were used to calculate the degree of outlierness of each object. The complexity of these two methods is given as follows:
- The complexity of the Ini_Distance is $O(m \times N^2)$ which makes it not suitable for large datasets.
- The complexity of the Ini_Entropy is $O(KmN + m^2N)$ which makes it not suitable for high dimensional datasets.

2.2.2 Cost function and distance based methods

In [7], the authors proposed an enhanced version of the *k*-mode by integrating the between cluster similarity terms in the optimization function to compute the individuals' similarity. This term is defined as follows:

$$B(W,Z) = \sum_{l=1}^{K}\sum_{i=1}^{N}\omega_{li}S(Z_l) \tag{7}$$

$S(Z_l)$ denotes the similarity between the lth cluster represented by z_l and other clusters and $\sum_{i=1}^{N}\omega_{li}$ its weight which is the number of objects in the lth cluster. Thus, it becomes possible to simultaneously minimize the within-cluster dispersion and enhance the between-cluster separation. The corresponding objective function is given as follows:

$$F(W,Z,\gamma) = \sum_{l=1}^{k}\sum_{i=1}^{N}\omega_{li}d(z_l,x_i) + \gamma\sum_{l=1}^{k}\sum_{i=1}^{N}\omega_{li}S(z_l) \tag{8}$$

The parameter γ is to maintain a balance between the effect of the within-cluster information and that of the between-cluster information on the minimization process. In their proposal [7], the authors applied this enhancement to three methods: the Huang version of the *k*-modes, the Ng's version of the *k*-modes and the weighted *k*-modes version.

In [38], a new dissimilarity measure is defined for the *k*-modes. The measure is based on the idea that the similarity between a data object and cluster mode, is directly proportional to the sum of relative frequencies of the common values in mode. Formally, the new dissimilarity measure is:

$$d(X_i,Q_l) = \sum_{j=1}^{m}\phi(x_{ij},q_{lj}) \tag{9}$$

where

$$\phi(x_{ij},q_{lj}) = \begin{cases} 1 - f_r(A_j = q_{lj}|X_l)\,(x_{ij} = q_{lj}) \\ 1\,(x_{ij} \neq q_{lj}) \end{cases} \tag{10}$$

Note that $f_r(A_j = q_{lj}|X_l)$ is the frequency of qlj in cluster X_l.

2.2.3 Uncertainty based methods

One of the first attempts proposed to handle uncertainty was by using the fuzzy sets theory. As an extension of the fuzzy k-means, the fuzzy *k*-modes [24] were proposed and many variants were also developed [10]. In this algorithm, each pattern or object can have membership functions to all clusters rather than having a strict membership to exactly one cluster. In the fuzzy *k*-modes, the objects of the universe U will be put in k clusters by finding W and Z that minimize the objective function given in Eq. (11):

$$F(W,CEN) = \sum_{l=1}^{K}\sum_{i=1}^{N}\omega_{li}^{\alpha}d(cen_l,obs_i) \tag{11}$$

subject to

$$\begin{cases} \omega_{\mathrm{li}}\epsilon[0,1],\ 1 \leq \mathrm{l} \leq \mathrm{K},\ 1 \leq \mathrm{i} \leq \mathrm{N} \\ \sum_{l=1}^{\mathrm{K}}\omega_{\mathrm{li}} = 1, 1 = l = K \\ 0 < \sum_{i=1}^{\mathrm{N}}\omega_{\mathrm{li}} < N, 1 \leq \mathrm{i} \leq \mathrm{N} \end{cases}$$

$\alpha \in [1, +\infty[$ is the fuzzy index; $W = [w_{li}]$ is a $K \times N$ real matrix, w_{li} is the membership degree of obs_i to the l^{th} cluster; $CEN=\{cen_1, cen_2, \ldots, cen_k\}$, $cen_l=[cen_{l1}, cen_{l2}, \ldots, cen_{lm}]$ is the lth cluster prototype with categorical attributes a_1, $a_2, \ldots, a_m$;

$d(cen_l, obs_i)$ is the simple matching dissimilarity measure as defined by Huang.

The method finds fuzzy cluster modes when a simple matching dissimilarity measure is used for categorical objects.

In [10], the authors proposed a fuzzy categorical clustering algorithm where the fuzzy *k*-modes' objective function was modified by adding a between-cluster information term. This consideration permitted simultaneously minimizing the within-cluster dispersion and enhancing the between-cluster separation. To obtain the modified objective function's local optimal solutions, the corresponding update formulas of the membership matrix and the cluster prototypes were derived. In their methods, the authors integrated the within-cluster and between-cluster information to update the membership matrix and cluster prototypes, which can effectively produce clustering results with high within-cluster similarity and low between-cluster similarity.

By assigning confidence to objects in different clusters, the clusters' core and boundary objects can be decided. This provides more useful information when dealing with boundary objects. However, the final fuzzy clustering outputs are still influenced by the mode initialization and the processing order of the objects in the datasets. Furthermore, these types of methods need to adjust one control parameter of membership fuzziness. In the applications, it is not clear how to find out the optimal parameters. Their values are often selected based on the decision makers' previous knowledge of the domain and their intuition or the proposed criteria.

On the other hand, RST, proposed by Pawlak since 1980, has received considerable attention in the computational intelligence literature since its development. It was used to develop clustering algorithms to handle uncertainty. The main advantage of RST based clustering methods compared to fuzzy clustering is that they don't require any domain expertise to assign the fuzzy membership.

In [39], the authors proposed the information-theoretic dependency roughness (ITDR), taking into account the information-theoretic attributes dependencies degree of categorical-valued information systems. In [26], the authors proposed the Total Mean Distribution Precision (TMDP) to select the partitioning attribute based on probabilistic RST. Using this technique and the concept of granularity, a new hierarchical clustering algorithm, called Maximum Total Mean Distribution Precision (MTMDP), for categorical data was developed. The MTMDP searches the best clustering attribute among the set of available features. It takes into account the mean distribution precision of all attributes and determines the further clustering node by considering the cohesion degree of all nodes. This consideration is a more reasonable method compared to previous methods proposed for RST clustering [40].

In [41], the authors proposed an algorithm based on fusing rough set and fuzzy set theories. The proposed rough fuzzy clustering method was used sequentially to integrate different measures to enhance the clustering performance. Thus, pure classified, semi rough and pure rough points are identified. After that, the Random Forest can be used in an incremental manner to classify these semi and pure rough points using pure classified points to yield better clustering results.

In [42], the authors addressed the issue of outlier detection as an initialization method to select the best centroids when starting the clustering process. The uncertainty regarding the clustering process is addressed by considering a soft computing approach based on rough sets. Accordingly, the modified clustering algorithm incorporates the lower and upper approximation properties of rough sets.

Most of the clustering approaches based Rough Set consider two techniques: (*i*) introducing a decision attribute based on which the dataset will be divided to partition the objects [39, 40, 43, 44] or, (*ii*) evaluating the lower, upper and quality of approximations of the a dataset [42, 45].

The selection of the most appropriate centroids when initializing the clustering process has been considered in many types of research since it may heavily impact the final results resulting from the partitioning procedure. However, this issue was only considered in the algorithm's first step and not in all the consecutive iterations within the process. Even when executing the updating of the modes in each iteration, multiple modes can be proposed. It is crucial to identify the most appropriate when to consider instead of automatically using the random selection method.

3 The DR*k*-M paradigm

3.1 The clustering model of the DR*k*-M

Let $IS=(U,A,V,f)$ be a categorical information system, and P a subset of the descriptive attributes A of the universe U ($P \subseteq A$). The objective of the clustering is to find the set of observations $OBS=\{obs_1, obs_2, \ldots, obs_N\}$ and the set of centroids $CEN=\{cen_1, cen_2,\ldots,cen_K\}$ that minimize the same cost function given in Eq. (1). The same constraints given in Eqs. (2) and (3) will also be considered. The DR*k*-M will implement the simple matching dissimilarity measure defined in Eqs. (4) and (5) during the assignments step of the observations to their closest clusters [4, 24].

The process of optimization can be described as follows:

- **Step 1.** Choose K distinct objects $cen_1, cen_2,\ldots,cen_K$ from the universe U as the initial set of modes ($t=1$) $CEN^{(t=1)}=\{cen_1,cen_2,\ldots,cen_K\} \in U^k$. Determine $W^{(1)}$ such that F $(W,cen^{(1)})$ is minimized.
- **Step 2.** Determine $cen^{(t+1)}$ such that $F(W^{(t)},CEN^{(t+1)})$ is minimized. If $F(W^{(t)},CEN^{(t+1)})=F(W^{(t)},CEN^{(t)})$, then stop.
- **Step 3.** Determine $W^{(t+1)}$ such that $F(W^{(t+1)},CEN^{(t+1)})$ is minimized. If $F(W^{(t+1)},CEN^{(t+1)})=F(W^{(t)},CEN^{(t+1)})$, then stop; otherwise set $t=t+1$ and go to step 2.

The set of observations is represented by the matrix $\bar{W}=[\bar{\omega}_{li}]$ according to Theorem 1.

Theorem 1 *Considering a fixed set of initial centroids* $\bar{Z}=\{cen_1, cen_2, \ldots, cen_h, \ldots,cen_k\}$, *the optimization problem defined in Eq.* (1) *is minimized by defining the matrix* $\bar{W}=[\bar{\omega}_{li}]$ *according to the following equation*:

$$\bar{\omega}_{li} = \begin{cases} 1, & \text{if} D\left(c\bar{e}n_l, obs_i\right) \le D\left(c\bar{e}n_h, obs_i\right) \text{ for } 1 \le h \le k \\ & 0, \quad \text{otherwise} \end{cases}$$

D stands for the simple matching dissimilarity measure. The $\bar{W}=[\bar{\omega}_{li}]$ matrix is a NxK binary matrix that reports whether an observation is part of a given cluster or not. To better illustrate the notion of this matrix, let's consider the example given in Fig. 2.

In the given example, the dataset is composed of seven observations to be put in three clusters. Once the clustering

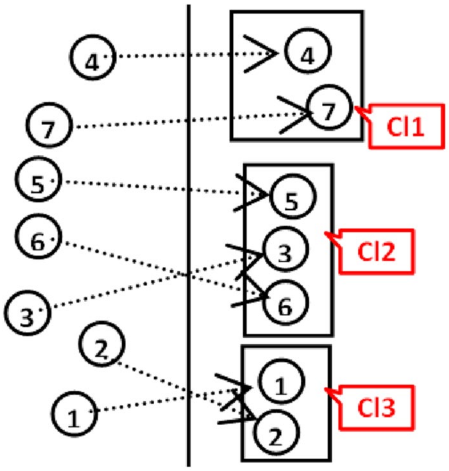


Fig. 2 W matrix with assigning observations to clusters

Table 1 Assignment of observations to clusters

		obs1	obs2	obs3	obs4	obs5	obs6	obs7	|cl*l*|
W=	cl_1	0	0	0	1	0	0	1	2
	cl_2	0	0	1	0	1	1	0	3
	cl_3	1	1	0	0	0	0	0	2
								N	7

is terminated, the *W* matrix will be generated as given in Table 1. If an observation obs_i belongs to the cluster Cl_l then $W_{il}=1$ otherwise $W_{il}=0$. The W matrix is used to determine to which cluster, each observation belongs.

Theorem 2 *let* $cen_{lj}=[cen_{l1}, cen_{l2},...,cen_{ld}]$ *be the mode of the lth* $(1 \leq l \leq K)$ *cluster and* $V_{a_j}=\left\{ a_j^{(1)}, a_j^{(2)}, \ldots, a_j^{(n_j)} \right\}$ *the domain of attributes* a_j*where* $|a_j|=n_j$ *be where* $(1 \leq j \leq d)$. *For a given object* $obs_i=[obs_{i1}, obs_{i2},\ldots,obs_{id}]$.

$F(W,Z) = \sum_{l=1}^{K} \sum_{i=1}^{N} \omega_{\mathrm{li}} \mathrm{D}_d\left(\mathrm{cen_l}, \mathrm{obs_i}\right)$ *is minimized if and only if* $\mathrm{cen_{lj}} = \mathrm{a_j^{(r)}}$

where $a_j^{(r)} \in V_{a_j}$ satisfies:

$$\left|\left\{ \omega_{\mathrm{li}} \mid \mathrm{obs_{ij}} = \mathrm{a_j^{(t)}}, \omega_{\mathrm{li}} = 1 \right\}\right| \leq \left|\left\{ \omega_{\mathrm{li}} \mid \mathrm{obs_{ij}} = \mathrm{a_j^{(r)}}, \omega_{\mathrm{li}} = 1 \right\}\right|$$

where $1 \leq t, r \leq \mathrm{n_j}$ for $1 \leq \mathrm{j} \leq d$

In other words, according to theorem 2, the quantity $\left|\left\{ \omega_{\mathrm{li}} \mid \mathrm{obs_{ij}} = \mathrm{a_j^{(r)}}, \omega_{\mathrm{li}} = 1 \right\}\right|$ should be maximized.

Proof of theorem 2

For a given matrix W, we have:

$$F(W,Z) = \sum_{\mathrm{l=1}}^{\mathrm{K}} \sum_{\mathrm{i=1}}^{\mathrm{N}} \omega_{\mathrm{li}} \mathrm{D_d}\left(\mathrm{cen_l}, \mathrm{obs_i}\right) = \sum_{\mathrm{l=1}}^{\mathrm{K}} \sum_{\mathrm{i=1}}^{\mathrm{N}} \sum_{j=1}^{d} \omega_{li} \mathrm{D}_{a_j}\left(\mathrm{cen_{lj}}, \mathrm{obs_{ij}}\right) = \sum_{\mathrm{l=1}}^{\mathrm{K}} \sum_{\mathrm{j=1}}^{d} \sum_{i=1}^{N} \omega_{li} \mathrm{D}_{a_j}\left(\mathrm{cen_{lj}}, \mathrm{obs_{ij}}\right) = \sum_{\mathrm{l=1}}^{\mathrm{K}} \sum_{\mathrm{j=1}}^{d} \gamma_{lj}$$

where $\gamma_{lj} = \sum_{i=1}^{N} \omega_{li} \mathrm{D}_{a_j}\left(\mathrm{cen_{lj}}, \mathrm{obs_{ij}}\right)$. Thus minimizing $F(W,Z)$ corresponds to minimizing γ_{lj}. Besides, $\gamma_{lj} = \sum_{i=1}^{N} \omega_{li} \mathrm{D}_{a_j}\left(\mathrm{cen_{lj}}, \mathrm{obs_{ij}}\right) = n - \left|\left\{ \omega_{\mathrm{li}} \mid \mathrm{cen_{lj}} = \mathrm{obs_{ij}}, \omega_{\mathrm{li}} = 1 \right\}\right|$ and thus minimizing γ_{lj} corresponds to maximizing $\left|\left\{ \omega_{\mathrm{li}} \mid \mathrm{cen_{lj}} = \mathrm{obs_{ij}}, \omega_{\mathrm{li}} = 1 \right\}\right|$ where n represents the cardinality of the dataset.

In terms, in theorem 2, minimizing the cost function $F(W,Z)$ corresponds to minimizing all the inner sums of the quantity γ_{lj} that are nonnegative and independent. The inner sum is minimized iff every term $n - \left|\left\{ \omega_{\mathrm{li}} \mid \mathrm{cen_{lj}}, \mathrm{obs_{ij}}, \omega_{\mathrm{li}} = 1 \right\}\right|$ is minimal which requires maximizing the cardinality of the sets where $\mathrm{cen_{lj}} = \mathrm{obs_{ij}}$.

Theorem 3 *The new proposed clustering method with the considered dissimilarity measure converges in a finite number of iterations.*

Proof theorem 3

Only a finite number of possible cluster modes *CEN={cen*$_1$*, cen*$_2$*, ...,cen*$_K$*}* can be defined. We then show that each final mode can have only one occurrence in the clustering process. This case corresponds to the last iteration and the stop criterion of the DRk-M.

If not, then there exist two distinct iterations $t_1 \neq t_2$ such that the centroids are equal $\mathrm{CEN}^{(\mathrm{t_1})} = \mathrm{CEN}^{(\mathrm{t_2})}$. According to the first Theorem, the proposed clustering algorithm using the simple matching dissimilarity measure computes the minimizers $W^{(t_1)}$ and $W^{(t_2)}$ for $CEN = CEN^{(t_1)}$ and $CEN = CEN^{(t_2)}$ for these two iterations, respectively which implies that: $F\left(W^{(t_1)}, CEN^{(t_1)}\right) = F\left(W^{(t_1)}, CEN^{(t_2)}\right) = F\left(W^{(t_2)}, CEN^{(t_2)}\right)$. In the other hand, the sequence $F\left(W^{(t)}, CEN^{(t)}\right)$ generated with the DRk-M using the simple matching dissimilarity measure is strictly decreasing which is not compatible with the previous result.

Table 2 Illustrative example of the generation of the candidate modes in a categorical dataset

	obs1	obs2	obs3	obs4	obs5	obs6	obs7
a1	a	a	c	d	a	d	d
a2	e	f	e	g	f	g	h
a3	l	n	k	l	l	m	l
a4	x	z	y	y	z	x	y

3.2 The mode of a categorical cluster

In order to illustrate the notion of the mode, an example is provided in Table 2. Let's consider a cluster composed of seven observations $\{obs_1, obs_2, obs_3, obs_4, obs_5, obs_6, obs_7\}$ and described by four categorical attributes $\{a_1, a_2, a_3, a_4\}$:

According to Table 2, the modes corresponding to that cluster are given as follows:

- The domain DOM of the attribute a_i corresponding to the set of the modalities taken by each attribute are given as follows: DOM $(a_1) = \{a,c,d\}$, DOM $(a_2) = \{e,f,g,h\}$, DOM $(a_3) = \{l, n, k,m\}$ and DOM $(a_4) = \{x,z,y\}$
- The most frequent value of a_1 could be a or d with three occurrences.
- The most frequent value of a_2 could be e, f or g with two occurrences.
- The most frequent value of a_3 is l with four occurrences.
- The most frequent value of a_4 is y with three occurrences.

Thus, for the cluster given in Table 2, six candidate modes could be defined and are provided as follows:

$$\mathrm{Q}_1 = [\mathrm{a, e, l, y}], \mathrm{Q}_2 = [d, f, l, y], Q_3 = [a, g, l, y], Q_4 = [d, e, l, y], Q_5 = [a, f, l, y], Q_6 = [d, g, l, y]$$

In the original version of the k-modes, the mode is selected randomly and, to our knowledge, no previously defined method was provided to identify the most appropriate mode which is considered as a restriction that limits the performance of the clustering method.

3.3 Uncertainty using the Rough Set Theory (RST)

The RST can deal with imperfect, vague and imprecise data based on the notion of indiscernibility between the observations. In this context of study, it is used to identify the most suitable modes among a list of candidate ones. With any rough set, a pair of precise sets, called the lower approximation and upper approximations, is associated. The lower approximation consists of all objects which surely belong to the set and the upper approximation contains all objects which possibly belong to the set [40, 46].

Definition 1 The indiscernibility relation.

Let $IS = (U,A,V,f)$ be a categorical information system and $\mathrm{B} \subseteq \mathrm{A}$ a subset of attributes, a binary relation $IND(\mathrm{B})$, called indiscernibility relation between two observations obs_i and obs_j of U is defined as:

$\mathrm{IND(B)} = \{(\mathrm{obs_i,obs_j}) \in \mathrm{U \ x \ U} \mid \forall \ \mathrm{a} \in \mathrm{B}, \mathrm{f(obs_i,a)} = \mathrm{f(obs_j,a)}\}$.

Thus, it is possible to consider that for every subset of attributes B selected from A, an indiscernibility relation can be generated. In other words, two observations are indiscernible in the context of a set of attributes if they have the same values for those attributes.

Definition 2 The lower approximation.

The lower approximation of a subset $X \subseteq U$ and $\mathrm{B} \subseteq A$ denoted $B_*(X)$ or $\underline{\mathrm{B(X)}}$ is defined as follows:

$$B_*(X) = \bigcup_{x \in U} \{x : [x]_B \subseteq X\}$$

Definition 3 The upper approximation.

The upper approximation of a subset $X \subseteq U$ and $B \subseteq A$ denoted $\mathrm{B}^*(\mathrm{X})$ or $B(\bar{X})$ is defined as follows:

$$B^*(X) = \bigcup_{x \in U} \{x : [x]_B \cap X \neq \emptyset\}$$

In order to better understand the notion of indiscernibility in a dataset, an example is provided in Table 3 related to the Covid-19 pandemic. Covid-19 is a strain of coronavirus that first broke out in Wuhan, China in December 2019 and has since become a global pandemic. The dataset

Table 3 Classification dataset for the Covid infection

Patient	Fever	Fatigue	Cough	Sneezing	Aches and pains	Sore throat	Headache	Covid/ Not Covid
Patient01	Yes	Yes	No	Yes	Yes	Yes	Yes	Covid
Patient02	Yes	Yes	No	Yes	Yes	Yes	Yes	Not Covid
Patient03	No	No	Yes	No	Yes	Yes	No	Covid
Patient04	Yes	Yes	Yes	Yes	Yes	No	Yes	Not Covid
Patient05	No	Yes	Yes	Yes	Yes	Yes	No	Covid

corresponds to seven attributes used as descriptive features and symptoms depicting the Covid-19. The dataset is composed of five patients and the last column corresponds o the labels of whether the patient is affected or not by the Covid-19.

In the example provided in Table 3, all the data were collected for various patients according to the symptoms that can either depict a *Not Covid* or a *Covid-19* illness. According to the table, two classes can be identified: the first one represents patients identified as infected by the Covid virus = {$patient_{01}$,$patient_{03}$,$patient_{05}$} and the second one represents patients identified as not infected by the Covid virus = {$patient_{02}$, $patient_{04}$}. Normally, it is expected that two users having the same profile will be classified into the same class which is not correct in this case since $patient_{01}$ and $patient_{02}$ do not follow this rule. According to definition 1, {$user_{01}$, $user_{02}$} are said to be indiscernible (similar) in view of the available set of attributes. According to definitions 2 and 3, the lower and upper approximations can be defined as follows:

- The lower approximation of the concept {Covid} = {$patient_{03}$, $patient_{05}$}
- The upper approximation of the concept {Covid} = {$patient_{01}$, $patient_{02}$, $patient_{03}$, $patient_{05}$}
- The lower approximation of the concept {not Covid} = {$patient_{04}$}
- The upper approximation of the concept {not Covid} = {$patient_{02}$,$patient_{04}$}

3.4 The rough modes

The DRk-M can be seen as a generalization of the Huang's definition of the mode. The approach is based on defining the list of all candidate modes, then generate a sub list that represents the most potentially modes and called rough modes.

Definition 4 Let *OBS = {obs*$_1$*, obs*$_2$*, …, obs*$_N$*}* be a set of categorical objects composed of *N* observations described by *d* categorical attributes A_1, A_2, …, A_d. A rough mode of *OBS = {obs*$_1$*, obs*$_2$*, …, obs*$_N$*}* is a set of vectors *Q = [q*$_1$*, q*$_2$*, …, q*$_d$*]* that minimize the quantity:

$$D(OBS,Q) = \sum_{i=1}^{N} d\left(obs_i, Q\right) \tag{12}$$

d is the simple matching dissimilarity measure.

The rough mode is the closest element to all the observations of the cluster. Minimizing the previous quantity is a key issue to determine the rough modes.

Theorem 4 *The function D(OBS,Q) is minimized if and only if*:

$$fr(Aj = qj|Cl_j) \geq fr(Aj = ckj|Cl_j)$$

for $q_j \neq c_{kj}$ for all $j = 1, \ldots, d$ where $f_r(A_j = c_{kj}|Cl_j) = \frac{n_{c_{kj}}}{N}$ corresponds to the relative frequency of the k^{th} category c_{kj} in attribute A_j and $n_{c_{kj}}$ is the number of objects having the k^{th} category c_{kj} in attribute A_j.

In other words, the *D(OBS,Q)* quantity is minimized by considering the most frequent modalities in each attribute to compose the rough mode.

Proof of Theorem 4

let $f_r(A_j = c_{kj}|Cl_j) = \frac{n_{c_{kj}}}{N}$ be the relative frequency of the k^{th} category c_{kj} in attribute A_j, where N is the total number of observations of the dataset and $n_{c_{kj}}$ the number of objects having the category c_{kj}.

we have

$$\sum_{i=1}^{N} d\left(obs_i, Q\right) = \sum_{i=1}^{N}\sum_{j=1}^{d} \delta\left(x_{ij}, q_j\right) = \sum_{j=1}^{d}\left(\sum_{i=1}^{N} \delta\left(x_{ij}, q_j\right)\right) = \sum_{i=1}^{d} N - n_{ij}$$

where $\delta\left(x_{ij}, q_j\right)$ corresponds to the simple matching dissimilarity measure and thus can take either 1 or 0 with a sum maximum value of N. n_{ij} represents the number of cases where $x_{ij} = q_j$. Thus, minimizing $\sum_{i=1}^{N} d\left(obs_i, Q\right)$ corresponds to minimizing

$$\sum_{i=1}^{d} N - n_{ij} = \sum_{i=1}^{d} N\left(1 - \frac{n_{qj}}{N}\right) = \sum_{i=1}^{d} N\left(1 - f_r\left(A_j = q_j|Cl_j\right)\right)$$

Because $N\left(1 - f_r\left(A_j = q_j|Cl_j\right)\right) \geq 0$ for $1 \leq j \leq d$, $\sum_{i=1}^{N} d_1\left(obs_i, Q\right)$ is minimized if and only if every $N\left(1 - f_r\left(A_j = q_j|Cl_j\right)\right)$ is minimal. Thus, $f_r\left(A_j = q_j|Cl_j\right)$ must be maximal.

Theorem 4 is used to compute the rough modes that correspond to the list of all possible modes within the cluster.

Definition 5 The rough upper and lower approximations

Let *S = (U, A, V, f)* be an information system, let B be any subset of attributes A and let X be any subset of observations U. The B rough-upper approximation of X, denoted by $\overline{B_R}(X)$ and B rough lower approximation $\underline{B_R}(X)$, are defined respectively as follows:

$$\overline{B_R}(X) = \bigcup_{x \in Q} \{B(x) : B(x) \cap X \neq \emptyset\} \text{ and } \underline{B_R}(X) \bigcup_{x \in Q} \{B(x) : B(x) \subseteq X\}$$

In all cases, the rough mode is a vector that contains the most frequent modalities in each attribute of the cluster observations. It may be an element of the cluster or a synthetic one generated during the process.

To select the best mode in the set of potential centroids, we don't only consider the distance between objects, but also the average density of the modes. If the distance between the object and the already existing cluster centers is the only considered factor, it is possible that outlier is taken as a new cluster center. Similarly, if the density of the object is only taken into account, it is utmost possible that many cluster centers can be located in the surrounding of one center. To avoid these potential problems, the distance between objects with the density of the object will be combined together to measure the possibility of an object to be a cluster center.

To better illustrate the notion of rough mode, Fig. 3 is given.

In Fig. 3, the clustering process of the DRk-M is provided. The DRk-M propose to investigate the step where the modes are updated in each iteration of the process which corresponds to the third step of the process. The simple matching dissimilarity measure is used as a distance metric and no centroid initialization in the first step is incorporated. In step 3, the DRk-M considers that more than only one mode is identified. This number can vary from a cluster to another. The mode with the highest density value will has a high probability to be selected as a centroid for that cluster and thus put in the upper approximation.

3.5 The proposed algorithm

The algorithm is described as follows:

```
Input: a CIS a categorical information system with $N$ observations and $d$ dimensions.
         K: the number of clusters
         $cen_i$ ($1 \leq i \leq$K) the set of initial centroids
         $\sigma$ the density threshold
         $\tau$ the distance threshold
Output: $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$ a set of $K$ clusters.
BEGIN
STEP 1: Randomly choose $K$ initial centroids $CEN$={ $cen_1$ ,…,$cen_j$,… $cen_K$ }
STEP 2: Observations cluster assignment.
            for each $cen_j$ ($1 \leq j \leq K$)
                for each $obs_i$ ($1 \leq i \leq N$)
                    Compute $\mathcal{D}(obs_i, cen_j)$ where $\mathcal{D}(cen_j, obs_j)$ corresponds to the simple matching dissimilarity metric
              end;
              for each $obs_i$ where $F(obs_i, cen_j)$ is minimal
                    $\mathcal{C}_j \leftarrow obs_i$
            end;
STEP 3: Computing the rough mode
           for each iteration $t$
               for each cluster $\mathcal{C}_j$
                   generate $\{candidate\_modes\}$
                    for each $can\_mode_p$ in $\{candidate\_modes\}$
                         Compute $density\ (can\_mode_p)$ where
                              $Dens(can\_mode_p) = \frac{|\{obs_j \in \mathcal{C}_j | D(can\_mode_p, obs_j) \leq \tau\}|}{|\mathcal{C}_j|}$
                                  If $Dens(can\_mode_p) \geq \sigma$ then $can\_mode_p \in \{lower\ approximations\}$
                                  otherwise $can\_mode_p \in \{upper\ approximations\}$.
                    end;
                   $rough\_mode_j \leftarrow random\ (\{lower\_approximations\})$
               end;
            end;
           If $CEN_{(t)}=CEN_{(t+1)}$ then stop;
           else
            CEN $\leftarrow rough\_mode_j$
           go to STEP 2.
END
```

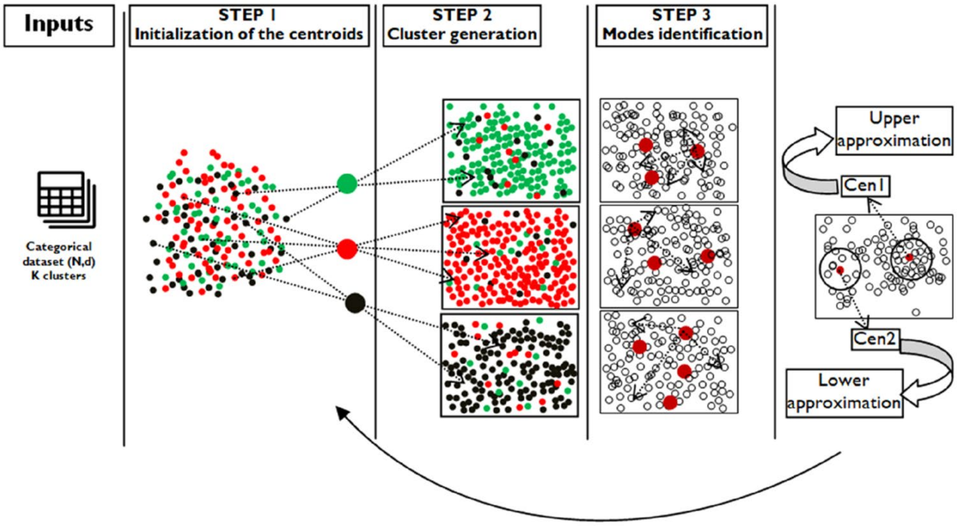


Fig. 3 The DRk-M clustering process

In the first step of the DRk-M, K initial observations are randomly selected as cluster modes. This initial random selection is the same approach also used in the k-modes. However, many previous methods proposed initialization methods to select the most appropriate initial modes [6, 8, 35–37]. It can be also possible to integrate in the upcoming researches initialization methods to the DRk-M. In the second step, the simple matching dissimilarity measure is used to assign the observations to their closest clusters. The focus of this step is to minimize the cost function defined in Eq. 10. In the third step, all possible candidate modes are computed for each obtained cluster considering the modality frequency for each attribute and put either in the lower or upper approximation.

3.6 Evaluating the complexity of the DRk-M

The DRk-M is scalable when compared to the standard k-modes since it does not affect the clustering paradigm but only introduces a new approximation step towards identifying the most adapted centroid in the cluster. In order to assess the scalability of the DR*k*-M it is required to computationally analyze all the different steps involved in the clustering process. As for the standard k-modes, the N observations of the DR*k*-M will be assigned into K clusters in t iterations and thus the complexity will be O($NKdt$). Then, in each iteration t, the computational complexity required to compute the modes is O($NKtd\, n_{c_{kj}}$) where $n_{c_{kj}}$ is the number of objects having the category c_{kj}. Finally, a time complexity of O($|C_k|pKtd$) where $|C_k|$ is the cardinality of the considered cluster is required to assign the identified rough modes either to the upper or lower approximation of the rough sets. As a conclusion, the overall complexity of the DRk-M will be $O(NKdt) + O(NKtd\, n_{c_{kj}}) + O(|C_k|pKtd) = O(NK|C_k|ptd\, n_{c_{kj}})$. Considering the approximation that K, t, d, $|C_k|$, p, $n_{c_{kj}}$ are $<<< N$, it is possible to conclude that the overall time complexity of the algorithm is O(N).

4 Experimentations

In this section, it is proposed to evaluate the clustering performance and scalability of the DR*k*-M. The algorithm will be compared to many states of the art algorithms including the Huang's *k*-modes [4] (1998), the Ng's *k*-modes (2007) [50], the Cao's dissimilarity (2012) [11], the improved Huang's *k*-modes [7] (2014), the Weighted *k*-modes [7] (2014), the Improved Weighted *k*-modes [7] (2014), Improved Ng's *k*-modes [7] (2014), the Bai's fuzzy k-modes [10] (2013), the Khan's initialization method [36] (2013) and the Fuzzy k-modes [10] (2013). Various experimental datasets will be used with several testing configurations either in terms of the number of observations N, clusters K or dimensions d. The efficiency of the DR*k*-M will be validated using several well known evaluation metrics. The algorithms were coded using the Java coding language and the experiments were executed with an Intel Core i7-3.5Ghz machine with 16 GB memory capacity.

4.1 Evaluation metrics

4.1.1 The accuracy

The accuracy is used to qualify the correctly classified cases. To compute the accuracy, each cluster is assigned to the most frequent pattern in the cluster according to the modalities in the attributes. The accuracy of this assignment is then measured by counting the number of correctly assigned observations and dividing it by the total number of observations N. The accuracy is computed according to Eq. 13:

$$purity(\mathcal{C},\mathbb{C})(AC)=\frac{1}{N}\sum_{k}\max_{j}\left|\omega_k\bigcap c_j\right| \tag{13}$$

$\mathcal{C}=\left\{\omega_1,\omega_2,\ldots,\omega_3\right\}$ is the set of clusters and $\mathbb{C}=\left\{c_1,c_2,\ldots,c_j\right\}$ is the set of classes identified from the patterns. The accuracy is always a positive value that ranges from 0 to 1 where a higher value of the accuracy depicts a better clustering.

4.1.2 The entropy

The entropy is used to measure the disorder in a distribution of objects. The smallest value for the entropy is 0. An increasing value of this metric indicates a bad clustering. This metric denoted H is defined in Eq. 14:

$$H(\mathcal{C})=-\sum_{k}P\left(\omega_k\right)logP\left(\omega_k\right)=-\sum_{k}\frac{\omega_k}{N}log\left(\frac{\omega_k}{N}\right) \tag{14}$$

$P\left(\omega_k\right)$ and $P\left(c_j\right)$ are the probabilities of an observation being in cluster ω_k and class c_j respectively.

4.1.3 Normalized mutual information

The mutual information (MI) of two random variables is a measure of the mutual dependence between them. The Normalized Mutual Information (NMI) metric ranges from 0 to 1 and as its value is high, better clustering is obtained. The NMI is defined according to Eq. 15:

$$NMI(\mathcal{C},\mathbb{C})=\frac{2\times I(\mathcal{C},\mathbb{C})}{H(\mathcal{C})+H(\mathbb{C})} \tag{15}$$

I is the mutual information defined in Eq. 16:

$$I(\mathcal{C},\mathbb{C})=\sum_{k}\sum_{j}P\left(\omega_k\bigcap c_j\right)\times log\left(\frac{P\left(\omega_k\bigcap c_j\right)}{P\left(\omega_k\right)P\left(c_j\right)}\right)=\sum_{k}\sum_{j}\frac{\left|\omega_k\bigcap c_j\right|}{N}\times log\left(\frac{N\left|\omega_k\bigcap c_j\right|}{\left|c_j\right|\left|\omega_k\right|}\right) \tag{16}$$

$P\left(\omega_k\bigcap c_j\right)$ corresponds to the probability of an observation being in the intersection of ω_k and c_j.

Three other evaluation metrics will also be used in this study which are precision, recall and the F1-score. These metrics can be directly computed from the confusion matrix. In order to better understand how these metrics are computed, let's consider the example given in Table 4:

Table 4 Confusion matrix for two classes

		Predicted	
		Cancer = YES	Cancer = NO
Actual	Cancer = YES	TP = 25	FN = 5
	Cancer = NO	FP = 5	TN = 65

The considered example concerns the classification resulting from a dataset composed of two classes: *Cancer = Yes* and *Cancer = NO*. The goal is to predict whether a patient has Cancer or not for a total of 100 patients. The confusion matrix represents the predicted (returned by the model) and the actual (real) results for each of the two classes. It is possible based on this matrix to identify the observations that were correctly classified and those that were not.

The confusion matrix obtained can be interpreted as follows:

- 25 patients that have cancer were correctly classified by the system as True Positives (TP), i.e. they represent patients that have truly cancer and were predicted with cancer by the system.
- 65 patients that do not have cancer were also correctly classified by the system as True Negatives (TN).
- Only 10 (5 + 5) patients were wrongly classified by the system where either the patients have cancer and were spotted as not having cancer or vice-versa. These two groups correspond respectively to False Negatives (FN) and False Positives (FP)

4.1.4 The precision

The precision is a measure of the correctly classified positive cases from all the predicted positive cases. Thus, it is useful when the costs of False Positives is high. The precision can be directly computed from the confusion matrix as follows:

$$\text{PR}=\frac{\text{True Positive (TP)}}{\text{True Positive (TP)+False Positive (FP)}}$$

4.1.5 The recall

The recall is a measure of the correctly identified positive cases from all the actual positive cases. It is important when the cost of False Negatives is high. This metric is defined as follows:

$$\text{RE}=\frac{\text{True Positive (TP)}}{\text{True Positive(TP)+False Negative (FN)}}$$

One other way to compute the accuracy using the confusion metric is to apply the formula:

$$\mathrm{AC} = \frac{\text{True Positive (TP)+True Negative (TN)}}{\text{True Positive (TP) + False Positive (FP) + False Negative (FN)+True Negative (TN)}}$$

4.1.6 The F1-score

The F1-score is the harmonic mean of the precision and recall and gives a better measure of the incorrectly classified cases than the accuracy metric.

$$F1 - score = 2 \times \frac{PR \times RE}{PR + RE}$$

For example, in the considered cancer dataset, according to the values presented in the confusion matrix, the values of the precision, recall, accuracy and F1-score can be given as follows:

$$\mathrm{PR} = \frac{25}{25+5} = \frac{25}{30}=0.83$$

$$\mathrm{RE} = \frac{25}{25+5}=0.83$$

$$\mathrm{AC} = \frac{25+65}{25+5+5+65} = \frac{90}{100}=0.9$$

$$\text{F1-score} = 2 \times \frac{0.83 \times 0.83}{0.83+0.83}= 0.83$$

4.1.7 The Silhouette score

The Silhouette score is a statistical interpretation and validation of clustering results that provides a measure of how well a data point is classified when it is assigned to a cluster. Thus, this metric can potentially be used as a quality measure to validate the clustering results according to the number of clusters in our case. The silhouette ranges from – 1 to + 1, where a high value indicates that the object is well matched to its cluster and poorly matched to neighboring clusters.

Considering a data observation $obs_i \in C_j$ that was classified in cluster C_j, it is possible to measure the mean distance between obs_i and all the other data points in the same cluster as follows:

$$\mathrm{a}(\mathrm{obs_i}) = \frac{1}{|\mathrm{C_i}|\text{-}1} \sum_{\mathrm{j\epsilon C_i}, ij} \neq \mathrm{d}(\mathrm{obs_i}, \mathrm{obs_j})$$

where d is the distance used such as the Euclidean distance. This metric can also be interpreted to qualify how well obs_i is assigned to a cluster: the smaller the value of a, the better the assignment is.

It is also possible to define the mean dissimilarity of obs_i to other clusters C_k as the mean of the distance from $obs_i$to all the points in C_k (where $C_k \neq C_i$). For each data point obs_i, the mean dissimilarity is computed as follows:

$$\mathrm{b}(\mathrm{obs_i}) = \min_{k \neq i} \frac{1}{|\mathrm{C_k}|} \sum_{\mathrm{j\epsilon C_k}} \mathrm{d}(\mathrm{obs_i}, \mathrm{obs_j})$$

This distance should be the *smallest* mean distance of obs_i to all the points in any other cluster, of which obs_i is not a member. The cluster with this smallest mean dissimilarity is said to be the "neighboring cluster" of obs_ibecause it is the next best fit cluster for point obs_i. The silhouette value of one data point obs_i is then given as follows:

$$\mathrm{s}(\mathrm{obs_i}) = \begin{cases} \frac{\mathrm{b}(\mathrm{obs_i})\text{-}\mathrm{a}(\mathrm{obs_i})}{\max\{\mathrm{b}(\mathrm{obs_i}), \mathrm{a}(\mathrm{obs_i})\}}, \text{ if } |\mathrm{C_i}|>1 \\ 0 \text{ if } |\mathrm{C_i}| = 1 \end{cases}$$

For $s(obs_i)$ to be close to 1, it is required that $a(obs_i) < < b(obs_i)$. As $a(obs_i)$ is a measure of how dissimilar $obs_i$is to its own cluster, a small value means it is well matched. Furthermore, a large $b(obs_i)$ implies that obs_i is badly matched to its neighboring clusters. Thus, if $s(obs_i)$ is close to one, this means that the data is appropriately clustered. If $s(obs_i)$ is close to negative one, then by the same logic, it is evident that it would be better to classify obs_i in a neighboring cluster. An $s(obs_i)$ near zero means that the data is on the border of two natural clusters.

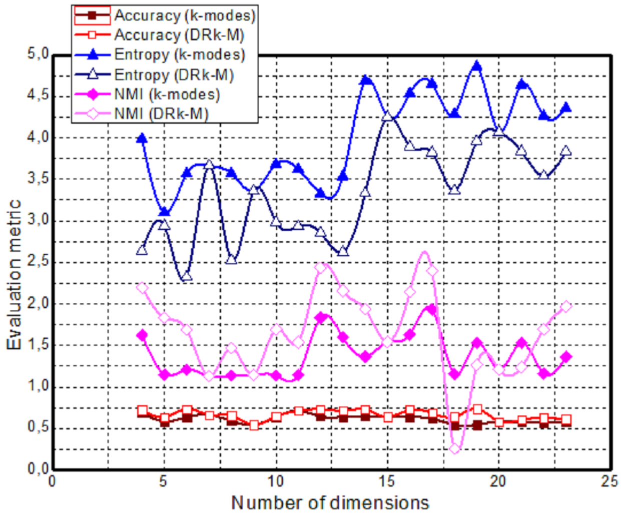


Fig. 4 Experiments for the Mushroom dataset with various dimensions ($N=8124$, $K=2$)

4.2 Experiments using the UCI datasets

In order to assess the efficiency of the DR*k*-M, the algorithm was experimented using five datasets extracted from the UCI Machine Learning Repository. These datasets were widely used in the literature to evaluate states of the art methods and are described as follows:

- *Mushroom data*: The data set includes descriptions of hypothetical samples corresponding to 22 species of gilled mushrooms in the Agaricus and Lepiota Family. It consists of 8124 objects and 23 categorical attributes. Each object belongs to one of the two classes, edible (4208 objects) and poisonous (3916 objects).
- *Breast cancer data*: The data set was obtained from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia. It consists of 699 data objects and 9 categorical attributes. It has two clusters: Benign (458 data objects) and Malignant (241 data objects).
- *Credit approval data*: The data set contains data from credit card organization, where customers are divided into two classes. It is a mixed data set with eight categorical and six numeric features. It contains 690 data objects belonging to two classes: negative (383 data objects) and positive (307 data objects). In the test, we only consider the categorical attributes on the data set.
- *Zoo data*: Zoo data set contains 101 elements described by 17 Boolean-valued attributes classified into seven classes.
- *Lung cancer data*: The data set was used by Hong and Young to illustrate the power of the optimal discriminant plane even in illposed settings. This data has 32 instances described by 56 categorical attributes. It contains three classes.

In the experiments, the DRk-M is first tested using the Mushroom dataset in terms of its dimensionality. Three evaluation metrics are used: the accuracy, the entropy and the NMI. The experiments are conducted by varying the number of dimensions d (4 → 24) and the obtained results are given in Fig. 4.

According to the results, the DR*k*-M provided better results in 86% of the total cases which makes it more accurate than the *k*-modes. For the accuracy, the values range from $a = 0.5437$ for $d = 9$ to $a = 0.7368$ for $d = 14$. In these cases, the DRk-M provided better results in 16 cases (80% of the total cases). The lines representing the accuracy are given in the bottom of Fig. 4. For the NMI represented with the lines in the middle of Fig. 4, the DR*k*-M provided values ranging from NMI = 1.1288 for $d = 7$ and NMI = 2.4387 for $d = 12$. In these cases, the DR*k*-M provided better results in terms of the NMI in 14 cases (70% of the total cases). The last metric used is the entropy, the values computed for the DR*k*-M range from $e = 2.3259$ for $d = 6$ to $e = 4.2498$ for $d = 15$. In terms of the entropy, the DR*k*-M provided better results than the *k*-modes in 16 cases (80% of the total cases). The conducted experiments are interesting since they permitted experimenting the effect of various dimensionalities on the performance of the DRk-M and compare it with the k-modes. It is important to mention that in this case, the number of clusters is $K = 2$ which is in concordance with the ground truth of the mushroom dataset since this dataset is in fact composed of two classes as mentioned in the dataset description. Different results would have been obtained if another value of K was selected.

The breast cancer dataset was also used to assess the efficiency of the DRk-M. In the experiments, the DRk-M and k-modes were compared for various numbers of clusters K (6 → 10). The corresponding results are provided in Table 5 where the accuracy, entropy and NMI are used as evaluation metrics.

For the breast cancer dataset, the DR*k*-M provided better clustering results in 12 cases.

In this second part of the experiments, more UCI datasets are used to compare the DR*k*-M to many state of the art methods as enhancements of the k-modes. Since most of these methods suffer from stability issues, 100 runs were carried out of the DR*k*-M with various initial modes. This technique was also used in many previous studies to ensure stable results. The comparison results of the DR*k*-M with each of the methods are given in Tables 6, 7, 8, 9, 10. Each value in these tables is the average of 100 times experiments.

Table 5 Experimental results computed for the Breast cancer dataset for K (6 → 10), $N = 644$, $d = 4$

K	6		7		8		4		10	
	k-modes	DRk-modes	k-modes	DRk-modes	k-modes	DRk-modes	k-modes	DRk-modes	k-modes	DRk-modes
Accuracy	0.4716	0.4440	0.324	**0.4631**	0.4512	**0.6237**	0.3603	**0.3777**	0.4435	**0.5035**
Entropy	3.5738	**3.3415**	4.3474	**3.2044**	4.3770	**3.342**	5.4385	**5.8406**	6.5446	**6.5130**
NMI	0.0461	0.0442	0.0045	**0.01**	0.0107	**0s.02**	0.0338	**0.04**	8×10^{-4}	2×10^{-4}

Values written in bold correspond to the metrics were the proposed algorithm performed better than state of the art methods

Table 6 The accuracy (AC) and F1-score computed for 100 runs for the Mushroom dataset

Methods	Huang's k-modes [4](1998)	Improved Huang's kmodes [7] (2014)	Weighted k-modes [7] (2014)	Improved Weighted k-modes [7] (2014)	Ng's k-modes [46,7] (2014)	Improved Ng's k-modes [7] (2014)	Bai's fuzzy NFKM [10] (2013)	Khan's initialization method [36] (2013)	Fuzzy k-modes [10] (2013)	DRk-M
AC	0.7176	0.8190	0.7106	0.8006	0.7969	0.8366	0.8298	**0.8815**	0.7001	**0,8591**
F1-score	0.7289	0.8250	0.7230	0.7827	0.7742	0.8411	0.8359	**0.8876**	0.6787	**0.8681**

Values written in bold correspond to the metrics were the proposed algorithm performed better than state of the art methods

Table 7 The accuracy (AC) and F1-score computed for 100 runs for the lung cancer

Methods	Huang's k-modes [4] (1998)	Improved Huang's k-modes [7] (2014)	Weighted k-modes [7] (2014)	Improved Weighted k-modes [7] (2014)	Ng's k-modes [7,46] (2014)	Improved Ng's k-modes [7] (2014)	Bai's fuzzy NFKM [10] (2013)	Khan's initialization method [36] (2013)	Fuzzy k-modes [10] (2013)	DRk-M
AC	0.5322	0.5803	0.5344	0.5631	0.5516	0.6003	0.6012	0.5000	0.5306	**0,6111**
F1-score	0.5545	0.5967	0.5408	0.5557	0.5779	0.6265	0.6008	0.5735	0.5580	**0.6399**

Values written in bold correspond to the metrics were the proposed algorithm performed better than state of the art methods

Table 8 The accuracy (AC) and F1-score computed for 100 runs for the breast cancer dataset

Methods	Huang's *k*-modes [4] (1998)	Improved Huang's *k*-modes [7] (2014)	Weighted *k*-modes [7] (2014)	Improved Weighted *k*-modes [7] (2014)	Ng's *k*-modes [7,46] (2014)	Improved Ng's *k*-modes [7] (2014)	Bai's fuzzy NFKM [10] (2013)	Khan's initialization method [36] (2013)	Fuzzy k-modes [10] (2013)	DR*k*-M
AC	0.8482	0.9270	0.8530	0.8441	0.8645	0.8770	0.9446	0.9127	0.8343	**0,9329**
F1-score	0.8263	0.9191	0.8051	0.8771	0.8535	0.8499	0.9383	0.9042	0.8111	**0.9434**

Values written in bold correspond to the metrics were the proposed algorithm performed better than state of the art methods

Table 9 The accuracy (AC) and F1-score computed for 100 runs for the credit approval dataset

Methods	Huang's k-modes [4] (1998)	Improved Huang's k-modes [7] (2014)	Weighted k-modes [7] (2014)	Improved Weighted k-modes [7] (2014)	Ng's k-modes [7,46] (2014)	Improved Ng's k-modes [7] (2014)	Bai's fuzzy NFKM [10] (2013)	Fuzzy k-modes [10] (2013)	DRk-M
AC	0.7367	0.7647	0.7442	0.7578	0.7612	0.7942	0.7701	0.7441	**0.7822**
F1-score	0.7453	0.7628	0.7428	0.7550	0.7590	0.7680	0.7712	0.7630	**0.7907**

Values written in bold correspond to the metrics were the proposed algorithm performed better than state of the art methods

Table 10 The accuracy (AC) and F1-score computed for 100 runs for the soybean dataset

Methods	Huang's k-modes [4] (1998)	Improved Huang's k-modes [7] (2014)	Weighted k-modes [7] (2014)	Improved Weighted k-modes [7] (2014)	Ng's k-modes [7,46] (2014)	Improved Ng's k-modes [7] (2014)	Bai's fuzzy NFKM [10] (2013)	Khan's initialization method [36] (2013)	Fuzzy k-modes [10] (2013)	DRk-M
AC	0.8553	0.9234	0.8613	0.9068	0.9396	0.9979	0.9264	0.9574	0.8336	**1**
F1-score	0.8702	0.9288	0.8702	0.9100	0.9442	0.9978	0.9319	0.9643	0.8495	**1**

Values written in bold correspond to the metrics were the proposed algorithm performed better than state of the art methods

In the tables, comparison of the DR*k*-M with some fuzzy *k*-modes methods as reported in [10] are also provided. In this case, the fuzziness parameter was fixed to α = 1.1. In fact, according to an explanation provided in [10], several values of the fuzziness parameter were tested and it was found that α = 1.1 provided the least value of the cost function to be minimized, i.e. best results were provided using this value. Besides, in all the experiments, the number of clusters is set to be equal to the number of classes for each of the given data sets in order to respect ground truth conditions. In the experiments, two metrics were used: the accuracy and the F1-score.

According to Tables 6, 7, 8, 9, 10, the DRk-M provided better clustering results in terms of the accuracy and F1-score for all the datasets considered except for the Mushroom dataset when testing the algorithm with the Khan's initialization method [36]. In the state of the art methods, many variants of the *k*-modes were considered either by improving the simple matching dissimilarity measure, using fuzzy methods or implementing an initialization method to select the most accurate initial centroids. In all these cases, the DR*k*-M with the proposed Rough mode selection provided more accurate results. The results obtained confirm the dominance of using the DRk-M for categorical clustering and the advantage of implementing the RST in updating the modes during the segmentation process.

4.3 Experiments using the twitter datasets

In this section, two datasets collected from Twitter using the python coding language were considered. The twitter accounts targeted correspond to some profiles related to terrorist groups and the datasets are described as follows:

- **Dataset 1:** this dataset contains 1803 instances described by 13 categorical attributes ["*month of the tweet*", "*tweet_id*", "*source*", "*device*", "*in_reply_to_status_id*", "*in_reply_to_user_id*", "*in_reply_to_screen_name*", "*user_tweet_id*", "*user_tweet_name*", "*user_tweet_screen_name*", "*user_tweet_location*" and "*language*". The tweets collected correspond to specific key words related to cyber terrorism and are given as follows: *Islamic State*, *caliphate editions*, *state of the caliphate*, *daesh*, *Battalion Okba-Ibn-Nafaa*, *African media.*
- **Dataset 2:** this second dataset contains 284 instances described by 10 categorical attributes ["*tweet_date*", "*screen name*", "*tweet_id*", "*in_reply_to_status_id*", "*retweeted_status_id*", "*reply_to_user_ID*", "*user_verified*", "*retweeted*", "*user_tweet_location*", "*hashtags*". The tweets were collected from two specific page user name given as follows: "*Gzrawi*" (131 tweets) for tweets posted by terrorist groups affiliated with "daech" and "Daesh_Online_01" (153 tweets).

Table 11 Clusters resulting from the segmentation process for K = 5

	C_1	C_2	C_3	C_4	C_5	
Tunisia	10	1	190	5	5	221
Egypt	5	160	9	10	5	189
Algeria	165	2	3	2	1	173
Lybia	2	4	2	48	3	59
Morocco	2	1	6	1	72	83
max	165	160	190	48	72	635/725 = 0.87

Table 12 Accuracy computed for the DRk-M and the *k*-modes for various *N* and *K*

Number of clusters (*K*)		5	6	7	8	9	10
Dataset 1 (N = 1803)	*k*-modes	0.6523	0.6310	0.7623	0.6845	0.6874	0.7239
	DR*k*-M	**0.7156**	**0.6861**	0.7623	**0.7260**	**0.6935**	**0.7819**
Dataset 2 (N = 284)	*k*-modes	0.7354	0.7325	0.6178	0.6912	0.7523	0.8234
	DR*k*-M	0.7354	**0.7819**	**0.6821**	**0.7361**	**0.8917**	0.8234

Values written in bold correspond to the metrics were the proposed algorithm performed better than state of the art methods

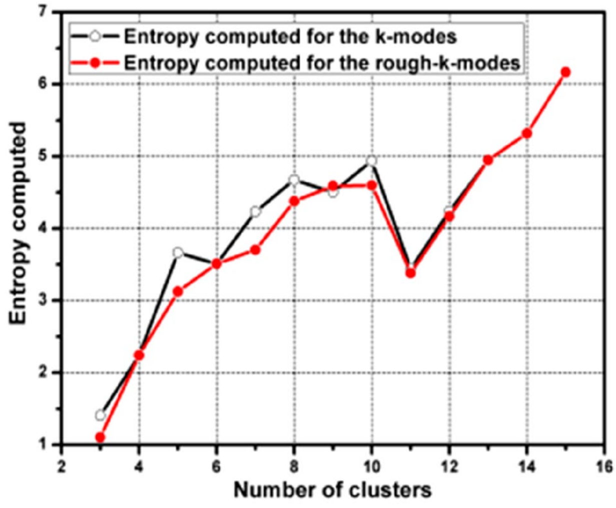


Fig. 5 Entropy computed for the for the two algorithms for $N = 10^3$ and K (3 → 15)

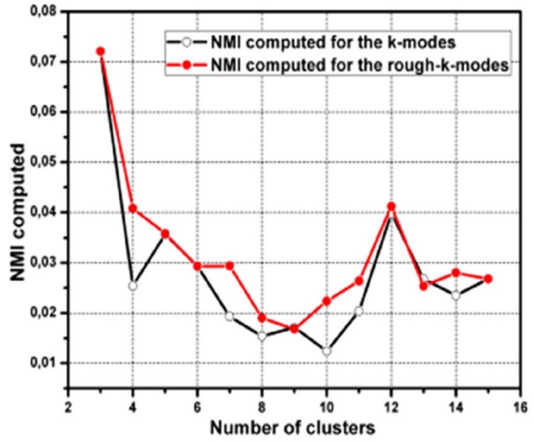


Fig. 6 NMI computed for the two algorithms for $N = 15 \times 10^3$ and K (3 → 15)

In this study, a categorical clustering algorithm is experimented. In Table 11, an example on how to evaluate the clustering results using the accuracy is given. The number of clusters K is initially defined. Then, the clustering process is launched. In this step, it is required to define the groups' class label which was set to the attribute location that corresponds to the place where the tweet was posted. A total number of 15 labels were then identified corresponding to various countries: Tunisia, Egypt, Algeria, Lybia, Morocco, Yemen, London, Iraq, KSA, Lebanon, Turkey, Kuwait, Syria, Yemen and NULL if no country is identified. For example, let's consider Table 11 where K = 5 groups is used:

In Table 11, the most frequent label in each cluster is identified in the last line = *max*. These values are then summed and used to compute the accuracy: $a = 0.87$.

Using the two datasets described above, the accuracy of the DRk-M and the *k*-modes is computed for various number of clusters *K* (5 → 10) and the obtained results are reported in Table 13.

According to the results given in Table 12, the DR*k*-M provided better results than the *k*-modes for the two datasets. For dataset 1, better results were obtained for all the experiments expect for K = 7 were the same accuracy was computed which can also be considered as an acceptable result. For dataset 2, better results were obtained in all cases expect for K = 5 and K = 10. Out of 24 experiments, the DRk-M

Table 13 Average accuracy, STD and average_Silhouette computed for the DRk-M and the *k*-modes for the two Twitter datasets for 50 runs of the algorithms

		Dataset 1 ($N = 10^3$)		Dataset 2 ($N = 15 \times 10^3$)	
		k-modes	DR*k*-M	*k*-modes	DR*k*-M
5	Average_accuracy	0.6412	0.7232	0.7557	0.7289
	STD	2.68%	2.27%	1.25%	0.79%
	Average_Silhouette	0.7854	0.7846	0.7432	0.7637
6	Average_accuracy	0.6251	0.6927	0.7597	0.7914
	STD	1.79%	1.12%	2.08%	1.83%
	Average_Silhouette	0.7284	0.7876	0.7013	0.7522
7	Average_accuracy	0.7543	0.7643	0.6210	0.6938
	STD	2.67%	2.39%	1.19%	0.72%
	Average_Silhouette	0.6832	0.6893	0.6144	0.6381
8	Average_accuracy	0.6718	0.7351	0.7183	0.7318
	STD	3.28%	2.83%	0.86%	0.59%
	Average_Silhouette	0.7291	0.7456	0.8267	0.8819
9	Average_accuracy	0.6706	0.7091	0.7418	0.9063
	STD	2.12%	1.88%	2.07%	1.80%
	Average_Silhouette	0.7156	0.7612	0.7155	0.7516
10	Average_accuracy	0.7164	0.7763	0.8161	0.8201
	STD	1.89%	0.93%	1.01%	0.91%
	Average_Silhouette	0.7396	0.8137	0.7236	0.7514

provided better results in 18 cases (75%). In the other cases, the same accuracy was computed.

In order to test the efficiency of the DRk-M for large datasets, the cardinality of dataset 2 (initially composed of 284 observations) was increased using several data copies. Thus, two datasets were generated with cardinalities $N = 10^3$ and $N = 15 \times 10^3$. In this part of the experiments, the entropy and the *NMI* were used as evaluation metrics to evaluate the effectiveness of the DRk-M and compare it with the k-modes. These experiments were conducted for various *K* (3 → 15) and the results reported in Figs. 5 and 6:

- For $N = 10^3$, in almost 46% of the cases, the DRk-M provided higher entropy than the *k*-modes.
- For $N = 15 \times 10^3$, the DR*k*-M provided better clustering results with higher NMI values than the k-modes in 52% of the cases.

In order to statistically validate the obtained results, additional experiments were conducted on the same datasets using the silhouette evaluation metric. The Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique indicates how well each object has been classified. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). In Table 13, the average accuracy computed for the DRk-M and k-modes for 50 runs of the two algorithms is reported. Two datasets were considered with various number of clusters K (5 → 10). Besides, the standard deviation was also computed for each set of 50 runs to identify the degree of confidence of the average accuracy calculated.

A higher value of the silhouette indicates a more compact and separated cluster. Thus, according to the results provided in Table 13, the clusters generated using the DR*k*-M for all the experiments are more compact and isolated than those generated using the k-modes.

In Fig. 7, the value of the silhouette score is given in the y axis. Each red point corresponds to the silhouette score computed for a given clustering. The k-modes and DRk-Modes were executed 50 times using the Twitter dataset ($N = 10^3$ and $K = 5 \rightarrow 10$). The silhouette mean *SC* was also computed and is given as follows:

$$\mathrm{SC} = \max_{k} \tilde{s}(k)$$

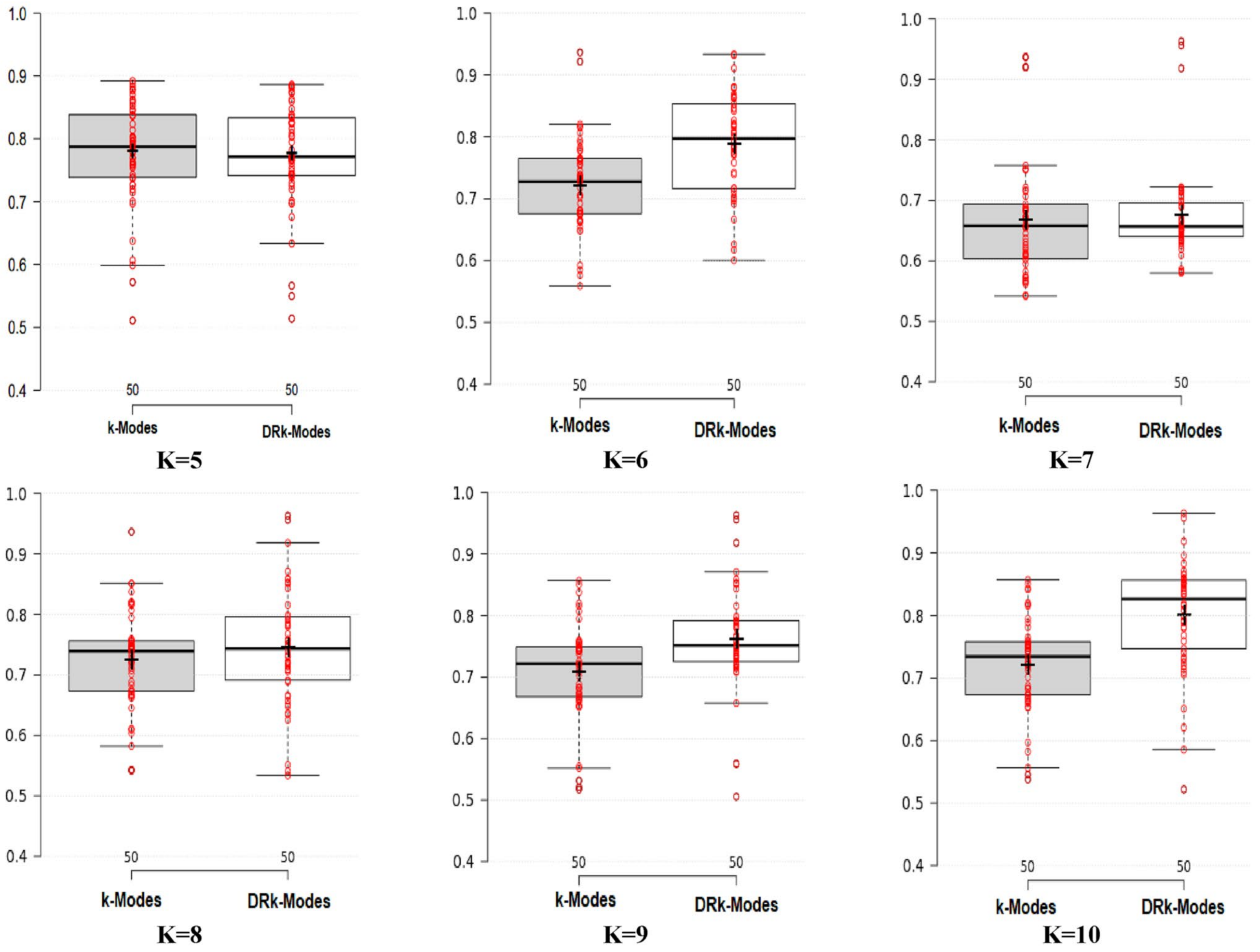
In the comparison between the resulting clusters, the focus is to identify the highest average Silhouette score resulting from the 50 runs conducted. From this purpose, a box plot representation was used. Box plots enable to study the distributional characteristics of a group of scores as well as the level of the scores. It is easy to identify the mean silhouette for each set of experiments which corresponds to the (+) sign. According to the results, the DRk-M provided a higher silhouette average than the k-modes and closer to one which indicates that its performance in producing more compact and isolated clusters is higher than the k-modes.

4.4 Experiments using the GTD datasets

In this section, the Global Terrorism Database was used to assess the scalability of the DR*k*-M in terms of the execution time and the accuracy. The tests were conducted for several dataset cardinalities *N* ($500 \rightarrow 25 \times 10^3$) and various number of clusters $K = 8$ and K = 10. The execution time was computed and reported in Fig. 8.

According to the results given in Fig. 8, the DRk-M provided higher computational time than the k-modes for large datasets. For small datasets, the two algorithms provided almost the same running performance. This issue is due to the time required for computing all the candidate modes in each cluster for the DRk-M which implies scanning the whole dataset multiple times for each attribute. Besides, it is possible to enhance the run time and obtain faster results by considering more powerful machines and resources either in terms of the memory or CPU.

In the other hand, to statistically validate the obtained results, the clustering outcomes were evaluated using the average accuracy for 50 runs of the two algorithms. Thus,



K=5

K=6

K=7

K=8

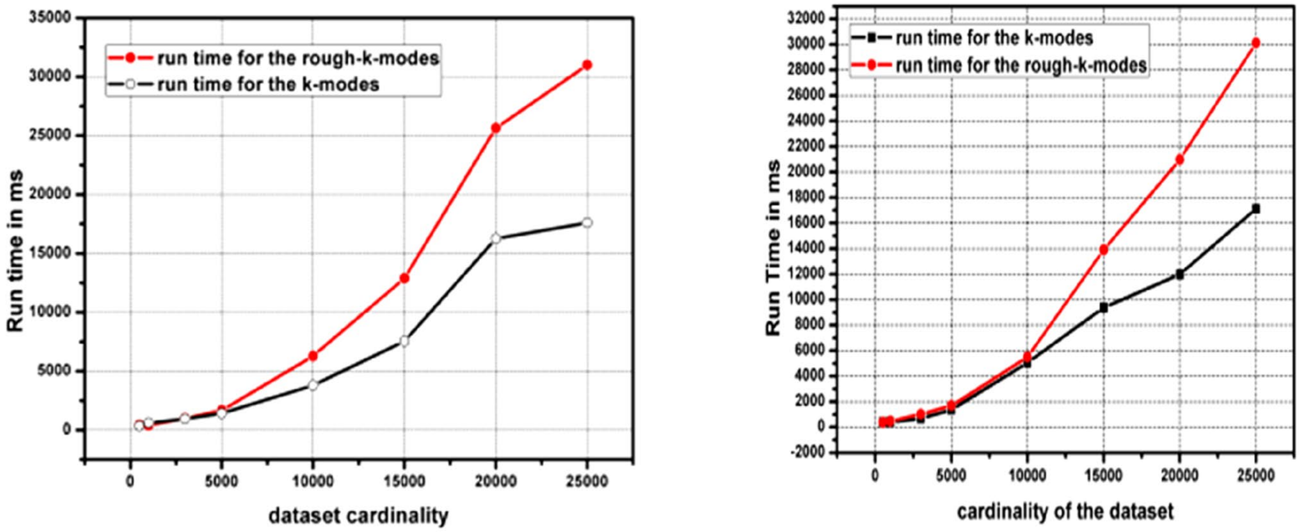
K=9

K=10

Fig. 7 Distribution of Silhouette scores for various clusterings according to the number of clusters for the DRk-M and k-modes ($N=10^3$, K: 5 → 10 and 50 runs)

two large datasets ($N=2\times10^4$ and $N=25\times10^3$) were considered for this purpose. The experiments were conducted for various number of clusters K = 3 → 8. Each algorithm was executed 50 times with various initial centroids in order to deal with stability issues. The average accuracy, STD and Silhouette were then considered for these 50 runs. The obtained results are reported in Table14.

Table 14 provides results related to the average accuracy computed for 50 runs of the DRk-M and the k-modes for several numbers of clusters. The accuracy reported how well the observations are arranged in their corresponding clusters. According to the values of the accuracy computed, the DRk-M provided better results than the k-modes in all cases which makes it more effective and efficient. Besides, in order to statistically characterize the values of the accuracy, the standard deviation (STD) is used to evaluate the overall spread of the 50 values calculate for each case. A lower value of the STD indicates more close values to the mean value (average) and thus depicts better results. According to Table 14, the DRk-M provides almost better results in all cases expect for K = 5 and K = 7 where the values computed for the k-modes were better. The silhouette was also used to characterize the compactness and density of the clusters generated by measuring the distance between the observations arranged in the clusters generated. A closer value to 1 of the silhouette indicates more compact and dense clusters. Once again and based on the results reported in Table 14, the DRk-M provided more accurate results than the standard k-modes.



a Execution time (K=8)

b Execution time (K=10)

Fig. 8 Execution time computed for *N* (500→25×10^3) and K = 8 and 10

Table 14 Accuracy, STD and average_Silhouette computed for the DRk-M and the k-modes for 50 runs using the GTD dataset

		Dataset 1 ($N=2\times10^4$)		Dataset 2 ($N=25\times10^3$)	
		k-modes	DR*k*-M	*k*-modes	DR*k*-M
3	Average_accuracy	0.6448	0.6837	0.6537	0.6967
	STD	1.13%	0.82%	1.72%	1.19%
	Average_Silhouette	0.6581	0.6819	0.7294	0.7628
4	Average_accuracy	0.6215	0.6928	0.6534	0.6976
	STD	2.34%	1.69%	2.91%	1.76%
	Average_Silhouette	0.6213	0.7089	0.6095	0.6780
5	Average_accuracy	0.6519	0.6686	0.6207	0.6911
	STD	3.67%	3.28%	1.68%	1.89%
	Average_Silhouette	0.6391	0.6814	0.6135	0.6318
6	Average_accuracy	0.7637	0.8019	0.6125	0.6308
	STD	3.68%	2.98%	2.59%	2.14%
	Average_Silhouette	0.7284	0.7446	0.8381	0.8734
7	Average_accuracy	0.7493	0.7739	0.6717	0.7193
	STD	2.57%	3.09%	1.89%	1.48%
	Average_Silhouette	0.7293	0.7675	0.8098	0.8539
8	Average_accuracy	0.7824	0.8239	0.8190	0.8382
	STD	1.32%	0.92%	1.68%	0.98%
	Average_Silhouette	0.7287	0.8097	0.6381	0.6937

5 Conclusion

Categorical clustering has gained great interest since the development of the k-modes algorithm. This algorithm has some major limitation related to the update of the modes in the last step of the process. In this paper, the RST was used as an uncertainty based model to identify the most accurate modes in a list of candidate ones when implementing categorical clustering. This consideration permits avoiding the random selection of the modes previously used in all states of the art k-modes based methods. The DRk-M is proposed based on computing the density of each candidate mode. This characterizes the number of observations that are closer to it as much as possible. Modes with high density would have higher probability to be considered as centroids for that cluster. In the experiments, multiple datasets with various configurations were used to assess the efficiency of the DRk-M. It was experimentally demonstrated that the DRk-M provided promising results. However, one main limitation of the DRk-M is the computational time required compared with the k-modes that is considerable due to the fact that more arithmetic computations are necessary to compute the list of all the candidate modes in the cluster. Since the DR*k*-M is more flexible and less exclusive than the *k*-modes in terms of the selection of the centroids in each cluster, the algorithm would provide more efficient results and thus the accuracy will be boosted.

References

1. Li H, Zhang L, Huang B, Zhou X (2020) Cost-sensitive dual-bidirectional linear discriminant analysis. Inf Sci 510:283–303
2. Bouguettaya A et al (2015) Efficient agglomerative hierarchical clustering. Expert Syst Appl 42(5):2785–2797

3. Liu A-A et al (2016) Hierarchical clustering multi-task learning for joint human action grouping and recognition. IEEE Trans Pattern Anal Mach Intell 34(1):102–114
4. Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min Knowl Dis 2(3):283–304
5. Liang J, Bai L, Dang C, Cao F (2012) The K-means-type algorithms versus imbalanced data distributions. IEEE Trans Fuzzy Syst 20(4):728–745
6. Cao FY, Liang JY, Jiang G (2009) An initialization method for the k-Means algorithm using neighborhood model. Comput Math Appl 58(3):474–483
7. Bai L, Liang J (2014) The k-modes type clustering plus between-cluster information for categorical data. Neurocomputing 133:111–121
8. Bai L, Liang J, Dang C (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
9. Bai L, Liang J, Dang C, Cao F (2011) A novel attribute weighting algorithm for clustering high-dimensional categorical data. Pattern Recogn 44(12):2843–2861
10. Bai L, Liang J, Dang C, Cao F (2013) A novel fuzzy clustering algorithm with between-cluster information for categorical data. Fuzzy Sets Syst 215:55–73
11. Cao F, Liang J, Li D, Bai L, Dang C (2012) A dissimilarity measure for the k-Modes clustering algorithm. Knowl Based Syst 26:120–127
12. Yanto ITR, Ismail MA, Herawan T (2016) A modified Fuzzy k-Partition based on indiscernibility relation for categorical data clustering. Eng Appl Artif Intell 53:41–52
13. Cao F, Liangn J, Li D, Zhao X (2013) A weighting k-modes algorithm for subspace clustering of categorical data. Neurocomputing 108:23–30
14. Salem SB, Naouali S, Chtourou Z (2018) A fast and effective partitional clustering algorithm for large categorical datasets using a k -means based approach. Comput Electr Eng 68:463–483
15. Semeh BS, Sami N, Moetez S (2017) Clustering Categorical Data Using the K-Means Algorithm and the Attribute's Relative Frequency. In: ICMLA: 14th International Conference on Machine Learning and Applications.
16. Semeh BS, Sami N, Moetez S (2017) A computational cost-effective clustering algorithm in multidimensional space using the manhattan metric: application to the global terrorism database. In: ICMLA 2017: 14th International Conference on Machine Learning and Applications.
17. Wu Bo, Wilamowski BM (2016) A fast density and grid based clustering method for data with arbitrary shapes and noise. IEEE Trans Ind Inf 13(4):1620–1628
18. Güngör E, Özmen A (2017) Distance and density based clustering algorithm using Gaussian kernel. Expert Syst Appl 64:10–20
19. Deng C et al (2018) GRIDEN: an effective grid-based and density-based spatial clustering algorithm to support parallel computing. Pattern Recogn Lett 104:81–88
20. McNicholas PD (2016) Model-based clustering. J Classif 33(3):331–373
21. Alamuri M, Bapi RS, Atul N (2014) A survey of distance/similarity measures for categorical data. In: International joint conference on neural networks (IJCNN). IEEE.
22. Liang JY, Zhao XW, Li DY, Cao FY, Dang CY (2012) Determining the number of clusters using information entropy for mixed data. Pattern Recogn 45(6):2251–2265
23. Bai L, Liang J, Guo Y (2018) An ensemble clusterer of multiple fuzzy k-means clusterings to recognize arbitrarily shaped clusters. IEEE Trans Fuzzy Syst 26(6):3524–3533
24. Kuo RJ, Nguyen TPQ (2019) Genetic intuitionistic weighted fuzzy k-modes algorithm for categorical data. Neurocomputing 330:116–126
25. Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 38(11):341–356
26. Li M et al (2014) Hierarchical clustering algorithm for categorical data using a probabilistic rough set model. Knowl-Based Syst 65:60–71
27. Ma W et al (2014) Image change detection based on an improved rough fuzzy c-means clustering algorithm. Int J Mach Learn Cybern 5(3):364–377
28. Maji P, Roy S (2015) Rough-fuzzy clustering and multiresolution image analysis for text-graphics segmentation. Appl Soft Comput 30:705–721
29. Dubey YK, Mushrif MM, Mitra K (2016) Segmentation of brain MR images using rough set based intuitionistic fuzzy clustering. Biocybern Biomed Eng 36(2):413–426
30. Podsiadło M, Rybiński H (2014) Rough sets in economy and finance Transactions on Rough Sets XVII. Springer, Berlin, Heidelberg, pp 104–173
31. Lausch A, Schmidt A, Tischendorf L (2015) Data mining and linked open data–New perspectives for data analysis in environmental research. Ecol Model 245:5–17
32. Hruschka H (2014) Comparing unsupervised probabilistic machine learning methods for market basket analysis. Rev Manag Sci: 1–31.
33. Delmelle EC (2016) Mapping the DNA of urban neighborhoods: clustering longitudinal sequences of neighborhood socioeconomic change. Ann Am Assoc Geogr 106(1):36–56
34. Lulli, A, et al. (2015) Scalable k-NN based text clustering. IEEE Int Conf Big Data (Big Data). IEEE
35. Dinh D-T, Huynh V-N (2020) k-PbC: an improved cluster center initialization for categorical data clustering. Appl Intell 50:2610–2632
36. Khan SS, Ahmad A (2013) Cluster center initialization algorithm for K-modes clustering. Expert Syst Appl 40(18):7444–7456
37. Jiang F, Liu G, Junwei Du, Sui Y (2016) Initialization of k-modes clustering using outlier detection techniques. Inf Sci 332:167–183
38. He Z, Shengchun D, Xiaofei X (2005) Improving k-Modes algorithm considering frequencies of attribute values in mode. In: International Conference on Computational Intelligence and Security, 157–162.
39. Park I-K, Choi G-S (2015) Rough set approach for clustering categorical data using information-theoretic dependency measure. Inf Syst 48:284–295
40. Herawan T, Deris MM, Abawajy JH (2010) A rough set approach for selecting clustering attribute. Knowl-Based Syst 23(3):220–231
41. Indrajit S, Sarkar JP, Maulik U (2015) Ensemble based rough fuzzy clustering for categorical data. Knowl Based Syst 77:114–127
42. Suri NNR, Ranga M, Narasimha M, Gopalasamy A (2016) Detecting outliers in categorical data through rough clustering. Nat Comput 15(3):385–394
43. Tripathy BK, Adhir Ghosh (2011) SDR: An algorithm for clustering categorical data using rough set theory. In: IEEE Recent Advances in Intelligent Computational Systems. IEEE
44. Gao CAN, Witold PEDRYCZ, Duoqian MIAO (2013) Rough subspace-based clustering ensemble for categorical data. Soft Comput 17(4):1643–1658
45. Jie HU, Tianrui LI, Chuan LUO, Hamido FUJITA, Yan YANG (2017) Incremental fuzzy cluster ensemble learning based on rough set theory. Knowl Based Syst 132:144–155