International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:6, 2017

Clustering Categorical Data Using the K-Means Algorithm and the Attribute's Relative Frequency

Semeh Ben Salem, Sami Naouali, Moetez Sallami

***Abstract*—**Clustering is a well known data mining technique used in pattern recognition and information retrieval. The initial dataset to be clustered can either contain categorical or numeric data. Each type of data has its own specific clustering algorithm. In this context, two algorithms are proposed: the *k*-means for clustering numeric datasets and the *k*-modes for categorical datasets. The main encountered problem in data mining applications is clustering categorical dataset so relevant in the datasets. One main issue to achieve the clustering process on categorical values is to transform the categorical attributes into numeric measures and directly apply the *k*-means algorithm instead the *k*-modes. In this paper, it is proposed to experiment an approach based on the previous issue by transforming the categorical values into numeric ones using the relative frequency of each modality in the attributes. The proposed approach is compared with a previously method based on transforming the categorical datasets into binary values. The scalability and accuracy of the two methods are experimented. The obtained results show that our proposed method outperforms the binary method in all cases.

***Keywords*—**Clustering, *k*-means, categorical datasets, pattern recognition, unsupervised learning, knowledge discovery.

I. Introduction

THE considerable increase of information technology devices manufacturing and the advances in scientific data collection methods lead to the creation of growing data repositories. Besides, traditional exploratory methods have shown their inefficiency in dealing with such data quantities to discover new findings. Thus, recent developed knowledge-discovery systems should implement an innovative and appropriate machine learning algorithms to explore these huge structures and to identify initially hidden patterns [1], [2].

In data mining, clustering [3] is the most commonly encountered knowledge-discovery technique applied in information retrieval and pattern recognition. It refers to unsupervised learning aiming to partition a dataset composed of N individuals embedded in an$d$-dimensional space into K distinct clusters without any prior knowledge about the distribution of the resulting clusters. The resulting data points in the same cluster are more similar to each other than to data points in other clusters. Three sub-problems are addressed by this process: (i) the similarity measure (distance) used to compare the data points, (ii) the iterative process of the designed algorithm to discover the clusters in an unsupervised way to guarantee the efficiency and (iii) derive a significant description for each obtained cluster to extract the corresponding proprieties and knowledge.

k-means is a well known clustering algorithm proposed for numeric datasets (containing numeric values) which makes it not adapted for clustering categorical datasets. This fact is a great restriction and limited the performance of this algorithm since, in many data mining applications, most considered datasets may contain categorical values. To deal with categorical datasets, the *k*-means was extended to obtain the *k*-modes algorithm that will be detailed in the next section. However, one other interesting issue is to convert the categorical data into numeric values and directly apply the *k*-means algorithm which is also interesting to discover.

This paper is organized as follows: in the second section, we present previous approaches towards clustering categorical data with their limits and provides a detailed description of the k-means that will be adopted in this study. In the third section, our proposed approach is detailed. Experimental results and discussion are provided in the fourth section, and the last section is devoted to the conclusion and perspectives.

II. Literature Review in Clustering Categorical Datasets

A. Categorical Clustering Algorithms

Although several proposals were made in the context of clustering categorical datasets, the most popular developed algorithm is the *k*-mode [4] and its variants [5]-[7]. It is an extension of the *k*-means algorithm where the Euclidean distance is replaced by the simple matching dissimilarity function, more suitable for categorical values, and the means by the modes, to identify the most representative element in a cluster (centroid). Besides, the modes are based on a frequency based method used in each iteration to update the centroids. The *k*-prototype algorithm [4] permits clustering mixed datasets with categorical and numeric values. Numerous variants were also proposed: the fuzzy *k*-modes algorithm [8] and the fuzzy *k*-modes algorithm with fuzzy centroids [9]. However, the main limitation when using the simple dissimilarity matching distance is that it does not provide efficient results since the simple matching often results in clusters with weak intra-similarity [10].

In [11], the authors showed that the similarity between two categorical values can also be referred as their co-occurrence according to a common value or a set of values which represents the second techniques to clustering categorical data considering the co-occurrence of the attributes. The most popular algorithm that falls into this category is the ROCK [12]. It measures the similarity between the categorical

Semeh Ben Salem, Sami Naouali and Moetez Sallami are with the Virtual Reality and Information Technology (VRIT), Military Academy of Fandouk Jedid, Tunisia (e-mail: semeh.bensalem@yahoo.fr, snaouali@gmail.com, Sellami-Moetez@outlook.fr).

patterns using the concept of links, i.e. the similarity between any two categorical patterns depends on the number of their common neighbors. Thus, the aim of this algorithm is to merge the patterns into a group that have relatively large number of links.

The notion of relative frequency was used in [13] to define a new dissimilarity coefficient for the *k*-modes algorithm in which the frequency of the categorical values in current cluster has been considered to calculate the dissimilarity between a data point and a cluster mode since the simple matching distance metric is not a good measure as it results in poor intra-cluster similarity.

Although the *k*-modes based algorithms have shown their efficiency in clustering large categorical datasets, like the *k*-means types algorithms, they still have two major limitations: (1) impossibility to cover the global information effectively, i.e. the provided solutions are only local optimal and a global solution is not easy to find [14], (2) the accuracy of the obtained results is sensitive to the number and shape of the initial centroids. Besides, the modes are more difficult to move in iterative optimization processes because the attribute values of categorical data are not continuous. The mode represents the most frequent element in the considered modality which means that if two modalities have close frequencies, only one will be retained and the other one will be dismissed, which results in information loss.

In this paper, two approaches are discussed and experimented for clustering categorical datasets using k-means algorithm: in the first method, we use a binary data representation to convert the initial categorical dataset into numeric values. In the second method, the relative frequency of the modalities in the attributes is used to execute the transformation.

B. The k-Means Algorithm

The *k*-means algorithm is a widely used clustering technique where an initial training set $S_N=\{x^{(i)},\ i=1,..,N\}$ composed of *N* elements $x^{(i)} \in \mathcal{R}^d$ described by *d* attributes is divided into *K* clusters $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K$. The clustering process is based on measuring the distance between the initially randomly selected centroids $z^{(i)}$ and the observations. The algorithm is described as follows:

1. Initialize centroids $z^{(1)}, \dots, z^{(K)}$;

2. Repeat until there is no further changes in cost function

a. $\forall j=1,\dots,K$: $\mathcal{C}_j = \{i;\ x^{(i)} \text{is closest to} z^{(j)}\}$

b. $\forall j=1,\dots,K$: $z^{(j)} = \frac{1}{|\mathcal{C}_j|}\sum_{i\in\mathcal{C}_j} x^{(i)}$ (cluster mean)

The *k*-means aims to minimize a criterion known as with in cluster sum of squares. This function is defined as follows:

$$cost = \sum_{i=1}^{K}\sum_{x\in\mathcal{C}_i}\left\|x^{(i)} - z^{(j)}\right\|^2$$

The resulting clusters are described by the mean of the samples in the cluster called "centroids" which may not be points from the dataset, although they live in the same space.

The *k*-means clustering algorithm is well known for its efficiency in clustering large datasets, and few previous proposals aimed to use its original version in clustering categorical data. On the other hand, some attempts were proposed to cluster categorical datasets using hierarchical algorithms but would not present interesting issues due to their quadratic time complexity that hindered their usage.

The main motivation behind this approach is to take the benefits from the *k*-means algorithm in terms of complexity: it is well known for its low computational cost O(*KNTd*) that is linear to the number of clusters *K*, the number of observations *N*, the number of iterations *T* and the number of attributes *d*. In [15], the author proposed an approach to using the *k*-means algorithm to cluster categorical data. The approach is based on converting multiple category attributes into binary values using either 1 or 0 to represent if the category is absent or present and to consider the binary attributes as numeric data. However, once used in data mining applications, this approach needs to handle a huge number of parameters and an increasing number of attributes corresponding to huge number of modalities. This fact will increase both the computational and space costs of the *k*-means algorithm. Besides, according to the algorithm's process, the cluster means computed representing the centroids will be contained into 0 and 1, which does not indicate the real characteristics of the clusters.

III. Proposed Approach for Clustering Categorical Datasets

In this paper, it is proposed to experiment a method to cluster qualitative data using the original version of the *k*-means algorithm. The considered dataset is assumed to be stored in a table, where each row (tuple) represents the observations described by the attributes arranged in columns. Encountered objects in many Data Mining applications are many times described by categorical information systems.

Definition 1. Formally, a categorical information system is represented by the quadruple *CIS*=(*U,A,V,f*), where:

- *U* is a non empty set of objects (universe).
- *A*a non empty set of attributes.
- *V* is a finite unordered set representing the union of all the attributes domain.
- *f* is a mapping information function.

Although the initial version of the *k*-means is not adapted for categorical data which represents its main limitation, in this paper we propose a new efficient approach to cluster categorical datasets based on *k*-means. To make it possible, our proposed solution consists of transforming the initial dataset into numeric values by considering the relative frequency of the modalities in each attribute.

Definition 2. The relative frequency is the number of occurrences of the k^{th} category $C_{k,j}$ in attribute $M_j$divided by the number of observations *N* in the dataset $\mathcal{D}$ and is defined as follows:

$$f_r(M_i = C_{k,j}/\mathcal{D}) = \frac{n_{C_{k,j}}}{N}$$

where $n_{C_{k,j}}$ is the number of occurrences of the category $C_{k,j}$.

This proposed approach will be compared to Ralambondrainy's method [15]. The corresponding clustering algorithm proposed in this context is described as follows:

Inputs:
$\mathcal{D} = \{ind_1, ind_2, \dots, ind_N\} \subseteq \mathcal{R}_N^d$ a set of N individuals;
$K\ (\ll N) \in \mathcal{N}$ desired clusters;
$\mathfrak{D} : \mathcal{R}_N^d \times \mathcal{R}_N^d \rightarrow \mathcal{R}$ the Euclidean distance;
Outputs:
a set of K clusters $\mathcal{C}_{i,1\leq i\leq K} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K\}$

Data Transformation (qualitative → quantitative)

STEP1: **FOR** each observation obs_i from $\mathcal{D}$**DO**
FOR each attribute M_j**DO**
Compute f_r of the k^{th} category $C_{k,j}$ in M_j

$$f_r(M_i = C_{k;j}/\mathcal{D}) = \frac{n_{C_{k,j}}}{N}$$

STEP 2: Randomly select K initial centroids (objects) from $\mathcal{D}$for the clusters;

$$cen_1^{(1)}, cen_2^{(1)}, \dots, cen_K^{(1)}$$

WHILE $(cen_{1,\dots,K}^{(t+1)} \neq cen_{1,\dots,K}^{(t)})$ **DO**
FOR each cluster $\mathcal{C}_i \in \mathcal{C}$ **DO**
FOR each individual $ind_{j,1\leq j\leq N} \in \mathcal{D}$**DO**
Compute $\mathfrak{D}(ind_j, cen_i)$
Assign each ind_j to the nearest cen_i

$$\mathcal{C}_i^{(t)} = \left\{ind_j : min \left\|\mathfrak{D}(ind_j, m_i^{(t)})\right\|\right\}$$

Re-compute new cluster centroid using the means;

$$cen_i^{(t+1)} = \frac{1}{\left|\mathcal{C}_i^{(t)}\right|} \sum_{ind_j \in \mathcal{C}_i^{(t)}} ind_j$$

The following example, gives an idea on how to implement the two methods.

TABLE I
EXAMPLE OF THE INITIAL CONSIDERED CATEGORICAL DATASET

Obs_i	Sex (M/F)	Work	Criminal Records
Obs_1	M	S	Y
Obs_2	M	E	Y
Obs_3	F	J	N
Obs_4	M	E	Y

*S: Student, E: Employee, J: Jobless, **Y: Yes, N: No

Table I provides an example of a categorical dataset containing four observations described by three categorical attributes. The first attribute (*Sex*) has two modalities (*Male/Female*), the second attribute (*Work*) has three modalities (*Student, Employee, Jobless*), and the third attribute (*Criminal Records*) has two modalities (*Yes/No*). When considering the first transformation method to obtain binary dataset, the obtained result will be as follows.

TABLE II.A
BINARY TRANSFORMATION OF THE INITIAL DATASET

Obs_i	Sex		Work			Criminal Record	
	Male	Female	Student	Employee	Jobless	Yes	No
Obs_1	1	0	1	0	0	1	0
Obs_2	1	0	0	1	0	1	0
Obs_3	0	1	0	0	1	0	1
Obs_4	1	0	0	1	0	1	0

TABLE II.B
RELATIVE FREQUENCY OF THE MODALITY IN EACH ATTRIBUTE

Modality	Male	Female	Student	Employee	Jobless	Yes	No
$f_r(C_{k;j})$	0.75	0.25	0.25	0.5	0.25	0.75	0.25

TABLE II.C
NUMERIC TRANSFORMATION OF THE INITIAL DATASET CONSIDERING THE RELATIVE FREQUENCY

Obs_i	Sex (M/F)	Work	Criminal Records
Obs_1	0.75	0.25	0.75
Obs_2	0.75	0.5	0.75
Obs_3	0.25	0.25	0.25
Obs_4	0.75	0.5	0.75

The relative frequency of each modality in the example is provided in Table II.B.

The obtained result when considering the proposed approach will be as follows in Table II.C.

IV. EXPERIMENTAL ANALYSIS RESULTS

In this section, the experimental environment and the initial dataset are described. The efficiency is evaluated using the accuracy. Besides, the contribution on scalability is also tested considering different values of the number of clusters K in a first step and then with 50 runs of four different values of K with different initial centroids.

The complexity of the *k*-means algorithm depends on the number of iterations T, attributes (dimensions) d, observations N and clusters K. In the experiments, N and K are equal for the two methods, however, the experimental results show that T for the binary method is higher than for the frequency based method. Besides, the resulting datasets to be experimented have different number of dimensions: for the binary transformation, this parameter is higher than for the frequency based method. These facts show that our new proposed technique permits reducing the complexity of the *k*-means once executed. Some proposals were made to reduce the dimensionality [16], [17] and can be considered if it is proposed to experiment the issue of reducing the dimensions of the resulting binary transformation.

A. Experimental Environment and Evaluation Criterion

The algorithm was coded with JAVA language and experimented on an Intel Core i3-2.1 GHz machine with a 4 GB RAM running on Windows 7 operating system. To evaluate the efficiency of the *k*-means in clustering categorical datasets using the relative frequency of attributes transformation, the accuracy is considered as an evaluation criterion and as this metric increase, better clusters are obtained. The accuracy is defined as follows:

Definition 3. The accuracy AC of a clustering is an external evaluation criterion that permits comparing the effectiveness of two clusterings as follows:

$$AC = \frac{\sum_{i=1}^{K} a_i}{|U|}$$

K is the number of predefined classes, a_i is the number of

correctly assigned objects.

In the experiments, the dataset contains a list of 50 terrorist attacks that occurred in several countries worldwide extracted from the Global Terrorism Database (GTD) [18]. Each dataset is described by eleven qualitative attributes: year, month, day, country, region, city, type, target, target nationality, group, mean. Although the first three attributes are numeric, they were considered as qualitative measures since they have fixed ranges and would provide more significance. Two datasets are then generated: the first one contains binary values (58 *Ko*) and the other contains numeric values computed using the relative frequency of the modalities (26 *Ko*).

B. Evaluation on Scalability

In this subsection, the scalability of the *k*-means applied for the two datasets is evaluated. This process is based on estimating two factors: the required execution time (run time) and the number of iterations necessary for the convergence. Besides, since the final results of the clustering depend on the initial centroids and to avoid the influence of their casual selection, we performed additional experiments to better experiment our proposed approach when fixing the number of clusters: for each experiment, the number of clusters *K* is fixed (*K*=3,4,5,6) and the initial centroids are modified to run the algorithm 50 times. Therefore, the average of 50 times runs is also provided to better illustrate the contribution on improving the scalability and effectiveness of our proposed technique.

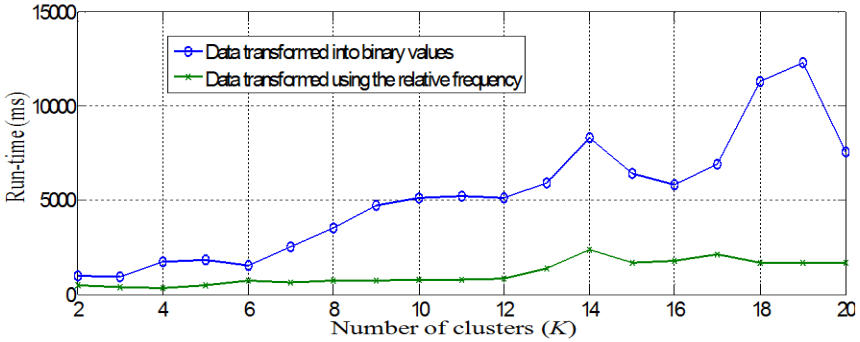

Fig. 1 Execution Time comparison using the two methods for the considered datasets over the number of clusters

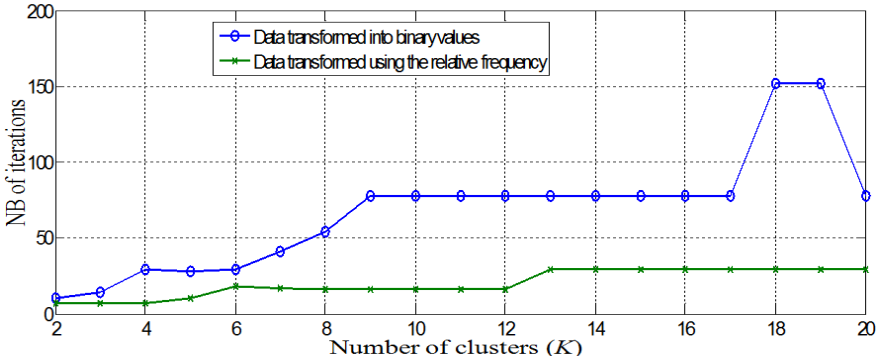

Fig. 2 NB of iterations comparison using the two methods for the considered datasets over the number of clusters

The two previous figures represent the scalability of the k-means clustering algorithm when considering two different initial datasets obtained according to our approach. The scalability is defined by two parameters: the run time and the number of iterations required by the algorithm to converge. According to these two factors, the relative frequency transformation is lower than the run time required by the binary method. This fact highlights the convenience of the proposed approach and the value of our contribution. The difference in the execution time is very significant and makes our proposed approach more adapted for data mining applications when dealing with huge datasets.

In the previous experiments, the scalability of the algorithm and its performance in clustering a transformed categorical dataset into numeric values were experimented according to different values of K. To better experiment it, it is proposed to test the scalability for four values of K with 50 runs and compute the average, the minimum and maximum values of the run time and number of iterations. The obtained results are summarized in Tables III.A-C.

TABLE III.A
AVERAGE OF THE RUN TIME AND NUMBER OF ITERATIONS REQUIRED BY THE TWO APPROACHES FOR VARIOUS VALUES OF *K*

K	Binary dataset		Dataset with relative frequency	
	Run time	iteration	Run time	iteration
3	860.6	16.6	369.13	7.16
4	1228.12	25.4	433.28	10.22
5	1792	28.68	511.08	11.64
6	2022.98	37.78	551.9	13.66

TABLE III.B
MINIMUM VALUES OF THE RUN TIME AND NUMBER OF ITERATIONS COMPUTED FOR THE TWO APPROACHES

K	Binary dataset		Dataset with relative frequency	
	Run time	iteration	Run time	iteration
3	608	10	312	6
4	780	8	312	6
5	811	11	331	7
6	952	14	390	8

TABLE III.C
MINIMUM VALUES OF THE RUN TIME AND NUMBER OF ITERATIONS COMPUTED FOR THE TWO APPROACHES

K	Binary dataset		Dataset with relative frequency	
	Run time	iteration	Run time	iteration
3	1545	29	468	9
4	3261	77	1435	41
5	2886	54	1108	30
6	4664	79	1170	29

According to the previous results, it is obvious that the proposed approach, consisting on transforming the categorical data into numeric values using the relative frequency of each modality in the attributes, is more scalable than the Ralambondrainy's technique: the average execution time and number of iterations calculated for 50 runs of the algorithm on the relative frequency dataset is lower when compared with the binary dataset. This fact highlights the importance of our proposed approach and its adaptability for huge datasets.

C. Evaluation on Clustering Efficiency

In this subsection, the clustering efficiency is experimented using the accuracy presented in section IV. Good clustering corresponds to higher values of the accuracy that represents the average of well clustered elements in their corresponding classes. In the first step of the experiments, the accuracy is computed for different values of the number of clusters *K* (2→20). The obtained results are shown in the following figure.

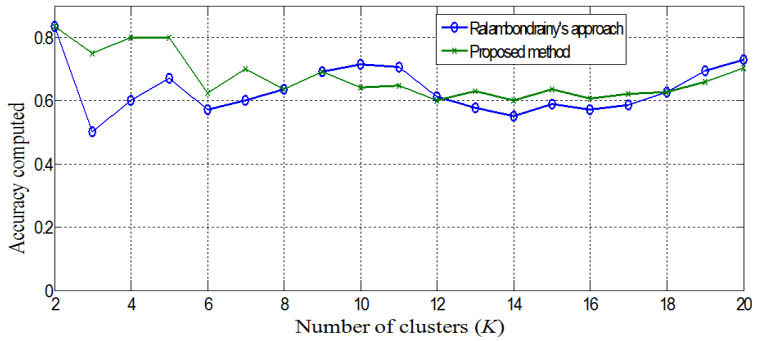

Fig. 3 Accuracy computed for various number of K

According to the previous results, it is obvious that the accuracy computed using the proposed approach is superior in major cases to the accuracy computed when considering the Ralambondrainy's approach [15]. The proposed approach is not only effective in terms of the run time and the number of iteration but also the efficiency is enhanced with the new proposal. The evaluation on clustering efficiency can be considered as more important than the scalability since it permits characterizing and identifying more imminent profiles, which is the aim scope of the clustering process.

As executed with the scalability experiments, it is proposed to consider accuracy computation over 50 runs of the algorithm for the two datasets. The same previous values of *K*(3,4,5,6) are considered. In Table IV, the average of the values of the accuracy computed in each case is provided.

TABLE IV
ACCURACY COMPUTED FOR 50 RUNS OF THE *K*-MEANS FOR THE TWO METHODS

	3	**4**	**5**	**6**
Ralambondrainy	0.5	0.57	0.51	0.65
Proposed approach	0.675	0.748	0.686	0.712

The provided results confirm again that the proposed approach is more effective in clustering categorical data if we consider the relative frequency of the modalities in the attribute in transforming the categorical data into numeric values. The obtained accuracies are higher for the proposed approach than the results provided for the Ralambondrainy's technique.

V. CONCLUSION

Clustering categorical data is a heavy and complex task, and specific clustering algorithm should be designed. In this paper, the relative frequency of each modality in their attributes is used to transform the categorical measures into numeric values. The *k*-means algorithm is the applied to the resulting dataset. Experimental results conducted show that our proposed approach permitted enhancing three parameters: (i) the scalability: the run time and number of iterations, (ii) the efficiency experimented using the accuracy and (iii) the complexity due to the reduction of the number of iterations and dimensions of the original dataset. These findings show the considerable contribution resulting from to the use of the relative frequency. This criterion is considered as the most appropriate statistical parameter to convert categorical into

numeric measures. However, more additional experiments should be conducted to evaluate the effectiveness of our proposed approach: in our future work, we propose to compare the experimented approach in this paper with other more advanced techniques proposed for clustering categorical datasets.

REFERENCES

[1] Jiawei Han, Jian Pei, Micheline Kamber, "Data Mining: Concepts and Techniques", Elsevier, 3rd edition, 2011, 744 p.
[2] Charu C. Aggarwal, "Data Mining: the textbook", Springer 2015, 734 pages.
[3] GuojunGan, Chaoqun Ma, Jianhong Wu, "Data Clustering: Theory, Algorithms, and Applications", ASA-SIAM Series on Statistics and Applied Probability, 2007.
[4] Zhexue Huang, "Extension to the k-means algorithm for clustering large data sets with categorical values." Data Mining and Knowledge Discovery 2, 283-304 (1998).
[5] Fuyuan Cao, Jiye Liang, Deyu Li, Liang Bai, Chuangyin Dang, "A dissimilarity measure for the k-modes clustering algorithm", Knowledge Based Systems 26 (2012), Elsevier, pp 120-127.
[6] Z. He, X. Xu, S. Deng, "Squeezer: an efficient algorithm for clustering categorical data" Journal of Computational Science and Technology 17 (5) (2002) 611-624.
[7] Z. He, X. Xu, S. Deng, "Scalable algorithms for clustering large datasets with mixed type attributes", International Journal of Intelligent Systems 20 (10) (2005) 1077-1089.
[8] Z. X, Huang, M. K Ng, "A fuzzy k-modes algorithm for clustering categorical data", IEEE transactions on Fuzzy systems 7(4) (1999) 446-452.
[9] D. W Kim, K. H Lee, D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids", Pattern recognition letters 25 (2004) 1263-1271.
[10] M. K Ng, M. J Li, Z. X Huang, Z. Y He "On the impact of dissimilarity measure in k-modes clustering algorithm." IEEE transactions on Pattern Analysis and Machine Intelligence 29 (3) (2007) 503-507.
[11] D. Gibson, J. Kleinberg, P. Raghavan, "Clustering categorical data: an approach based on dynamical systems", Proceedings of the 24th VLDB Conference, New York, 1998, pp 311-322.
[12] S. Guha, R. Rastogi, K. Shim, "ROCK: a robust clustering algorithm for categorical attributes"Proceedings of the IEEEInternationalConference on Data Engineering, Sydney, Australia 1999, pp 512-521.
[13] Ng M. K., Li M. J, Huang J. H, He Z, "On the impact of dissimilarity measure in k-modes clustering algorithm." IEEE transactions on Pattern Analysis and Machine Intelligence 29 (3); 503-507, 2007.
[14] A. Chaturvedi, Paul E. Green and J.D Caroll, "K-modes clustering.", Journal of classification, Vol.18, No 1, pp 35-55, 2001.
[15] Ralambondrainy, H, "A conceptual version of the k-means algorithm." Pattern recognition Letters 16, 1147-1157, 1995.
[16] Semeh Ben Salem, Sami Naouali, "Reducing the multidimensionality of OLAP cubes with Genetic Algorithms and Multiple Correspondence Analysis", international conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015), Tunisia.
[17] Semeh Ben Salem, Sami Naouali, "Towards Reducing the multidimensionality of OLAP cubes using the Evolutionary Algorithms and Factor Analysis Method", International Journal of Data Mining and Knowledge Management Process (IJDKM 2016).
[18] Semeh Ben Salem and Sami Naouali, "Pattern Recognition Approach in Multidimensional Databases: Application to the Global Terrorism Database" International Journal of Advanced Computer Science and Applications (IJACSA), 7(8), 2016.