World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:11, No:6, 2017

A Computational Cost-Effective Clustering Algorithm in Multidimensional Space Using the Manhattan Metric: Application to the Global Terrorism Database

Semeh Ben Salem, Sami Naouali, Moetez Sallami

***Abstract*—**The increasing amount of collected data has limited the performance of the current analyzing algorithms. Thus, developing new cost-effective algorithms in terms of complexity, scalability, and accuracy raised significant interests. In this paper, a modified effective *k*-means based algorithm is developed and experimented. The new algorithm aims to reduce the computational load without significantly affecting the quality of the clusterings. The algorithm uses the City Block distance and a new stop criterion to guarantee the convergence. Conducted experiments on a real data set show its high performance when compared with the original *k*-means version.

***Keywords*—**Pattern recognition, partitional clustering, K-means clustering, Manhattan distance, terrorism data analysis.

I. Introduction

IN knowledge-discovery based systems [1], [2], the extensive collection of data from heterogeneous sources permits creating big data structures. Analyzing these forms of knowledge to extract initially hidden and undetectable patterns has raised significant interest in several fields including artificial intelligence, sentiment analysis and counter-terrorism. However, the analyzing process suffers from a major constraint which is the increasing computational cost required especially when dealing with huge datasets. Meanwhile, this analysis task should not compromise the effectiveness of the approaches and the reliability of the final results. Clustering was presented as an interesting issue in this context. It permits identifying prominent patterns without any previous intelligence concerning the shape or the requirements of the process in an unsupervised way. This "blind" method of discovering the profiles makes the provided algorithms complex and complicated and requires more scalable techniques [3]. Besides, each clustering algorithm has its strengths and weaknesses, due to the complexity of information. Three main topics are addressed by the clustering process: the *Similarity measure*, the *Clustering process* and *Cluster validation*. The *k*-means is a well-known clustering algorithm proposed for quantitative datasets. Although it is characterized by its simplicity and fast convergence, it is unable to deal rapidly with massive datasets leading to an expensive computational cost.

In this paper, an enhanced *k*-means approach was experimented. The objective is to reduce the computational time required without affecting the effectiveness of the algorithm. In the new algorithm, the Euclidean distance is replaced by the Manhattan metric to evaluate the similarity between the observations of the dataset, and the stop criterion based on comparing the centroids between two consecutive iterations is modified to consider the cardinality of each constituted cluster instead. Conducted experimental results show that the new approach has significantly superior performance than the direct *k*-means in major cases.

The second section of this paper presents previous works related to clustering using *k*-means. The third section details the proposed approach. In Section IV, we introduce the experimental datasets, in Section V the results are detailed, and the last part is devoted to the conclusion.

II. Clustering Approaches in Data Mining

The major task executed by a clustering process aims to divide a data set $\mathcal{D}$ into K non-empty disjoint subsets $\mathcal{C}_i$ where $\mathcal{D} = \bigcup_{i=1}^{K} \mathcal{C}_i = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K\}$. Initially, K elements of $\mathcal{D}$ are selected as the K initial centroids of the clusters. The distance is then measured between each centroid Cen_j and each remaining observation of $\mathcal{D}$ that will be then assigned to the closest cluster. The purpose is to minimize a cost function defined by (1):

$$\mathcal{F} = \sum_{i=1}^{N} \sum_{j=1}^{K} \left\| obj_i - Cen_j \right\| \quad (1)$$

Cen_j is the centroid of the j^{th} cluster, obj_i is the i^{th} object selected from $\mathcal{D}$, and $\|.\|$ is the distance metric. The main problems encountered in a clustering process are its dependency to the choice of the initial number of clusters K and the distance metric.

In Partitional Clustering Algorithms, the initial dataset is divided into K distinct groups [4], [5] as described above, while in Hierarchical Clustering Algorithms a hierarchical set of clusters is created and the clusters are merged either in an agglomerative or divisive way. For agglomerative algorithms, each object obj_i of $\mathcal{D}$ forms an independent cluster and thus K=N. The obtained clusters are then merged iteratively until

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).

reaching a stop criterion. CACTUS [6] and BIRCH [7] fall into this category. For divisive algorithms, all the objects of $\mathcal{D}$ form initially one cluster (K=1). As the clustering process is executed, the initial super cluster is divided into K sub-clusters according to some similarity criterion until reaching a stop condition.

The similarity measure used in the *k*-means is the Euclidean metric derived from the Minkowski distance defined by (2) where (p=2).

$$D_p(X,Y) = \left(\sum_{k=1}^{d} |x_{i,k} - y_{j,k}|\right)^{\frac{1}{p}} \quad (2)$$

Let us consider a d dimensional Euclidean space $\mathcal{R}^d$ and $\mathcal{D} \subset \mathcal{R}^d$, a finite set of elements. Each element of $\mathcal{D}$ is a d-dimensional object defined by d attributes $X = (x_1, x_2, \dots, x_d)$. The Euclidean distance is defined as:

$$DIST_{Euclidean}(X,Y) = \sqrt{\sum_{k=1}^{d} |x_{i,k} - y_{j,k}|} \quad (3)$$

The *k*-means clustering process is defined as follows:
Input: a finite set $\mathcal{D} \subset \mathcal{R}^d$ with N elements; K clusters
Output: K clusters $\mathcal{C}_K \subset \mathcal{R}^d$
Goal: minimize the following cost function:

$$P(W,Q) = \sum_{l=1}^{K} \sum_{i=1}^{N} w_{i,l} d(X_i, Q_l) \quad (4)$$

$\sum_{l=1}^{K} w_{i,l} = 1$ and $w_{i,l} \in \{0,1\}$, $1 \le i \le N, 1 \le l \le K$. W is a NxK partition matrix and $Q = \{Q_1, Q_2, \dots, Q_K\}$ is a set of objects in the same object domain and d the squared Euclidean distance.

In [8], the authors present a detailed review of clustering techniques with more than 70 algorithms classified into 19 categories. A detailed definition of each algorithm was provided as well as its procedures. The associated similarity measures used as distance metrics were also introduced, and different evaluation indicators were cited. The essential and core idea of each commonly used clustering algorithm is then presented, and the advantages and disadvantages of each one are analyzed.

K-means algorithm is sensitive to the choice of the initial centroids [9]-[11], and different primary inputs may lead to multiple clusterings. Reference [12] proposed an incremental algorithm to determine the number of the initial clusters K. In [13], a Min-Max *k*-means clustering algorithm is presented to permit a more precise selection of the initial starting conditions. The algorithm starts with a randomly picked set of centroids and tries to minimize the maximum intra-cluster error. It has shown high efficiency even in intrusion detection systems [14]. In [15], the same research topic is also explored. The approach evaluates the distances between every pair of data points and finds out those which are similar. Finally, the initial centroids are constructed according to these found data-points. In [16], the authors proposed the global *k*-means to eliminate the singleton clusters generated during the clustering process, and they applied the Min-Max *k*-means clustering error method to global *k*-means to overcome the effect of bad initialization. The method is independent of any starting conditions and compares favorably to the *k*-means and the Min-Max *k*-means with multiple random restarts.

III. Modified *K*-Means Partitional Algorithm

The new algorithm uses the City Block distance and a new stop criterion. In *k*-means, the process ends when the centroids computed in the $(i+1)^{th}$ and the i^{th} iterations are equal, which is not adapted for the City Block distance since all the observations will be included in one cluster. The new stop criterion considers convergence when the cardinality of a cluster is equal to N. The retained effective final clustering result will correspond to the one obtained during the previous iteration. Besides, we propose to apply the approach for mixed datasets (qualitative and quantitative) which represents in its self an innovation since the original k-means is not adapted for this category. The qualitative data will be converted into numeric values using their relative frequency [17] to remove the numeric-only limitation of the *k*-means. The modified approach will be applied to the multidimensional context where each element of the dataset is described by N elements with d attributes.

In order to enhance the computational complexity of the original *k*-means, we replaced the Euclidean metric by the Manhattan distance defined as:

$$DIST_{Manhattan}(X,Y) = \sum_{k=1}^{d} |x_{i,k} - y_{i,k}| \quad (5)$$

The minimized cost function when using the City Block metric will be as:

$$cost\left(C_{i,1\le i\le K}\right) = \min_{\mathcal{C}_1,\dots,\mathcal{C}_K} \sum_{j=1}^{K} \sum_{i\in\mathcal{C}_j} \|obj_i - Cen_j\| = \sum_{i=1}^{n} \min_{j=1\dots K} |obj_i - Cen_j| \quad (6)$$

d_i is the i^{th} object in $\mathcal{D}$, C_j is the centroid of the j^{th} cluster.

The relative frequency of the k^{th} category $C_{k,j}$ in attribute M_j is defined as:

$$f_r(M_i = C_{k;j}/\mathcal{D}) = \frac{n_{C_{k,j}}}{N} \quad (7)$$

where $n_{C_{k,j}}$ is the number of occurrences of category $C_{k,j}$. The following Table I represents the proposed algorithm:

IV. Experimental Results and Discussion

In the experiments, several datasets are considered with various cardinalities so to better test the scalability of the proposed approach. The datasets correspond to the terrorist attacks that occurred in four North African countries: ALGERIA (2704 attacks), EGYPT (1218 attacks), LYBIA (1098 attacks) and TUNISIA (75 attacks). These datasets are extracted from the Global Terrorism Database (GTD) [18]-[21]. Each dataset is described by 11 mixed attributes: five qualitative (*city, type, target, group, sub_type*) and six quantitative (*year, month, day, success, nb_kills, nb_wounds*).

In this section, a performance study is conducted to evaluate the approach and compare it with the *k*-means in terms of

effectiveness and computational costs. The algorithm was coded using JAVA on an Intel Core i3-2.1 GHz machine with a 4GB RAM running on Windows 7 operating system. All the figures presented in this paper are generated using MATLAB.

TABLE I
PROPOSED NEW CLUSTERING ALGORITHM BASED K-MEANS

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 Manhattan 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)
FOR each categorical 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 1: Randomly select K initial centroids (objects) from $\mathcal{D}$ for the clusters;
$cen_1^{(1)}, cen_2^{(1)}, \dots, cen_K^{(1)}$
WHILE $(|\mathcal{C}_{i,1\leq i\leq K}^{(t)}| \neq N)$ **DO**
STEP 2: **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

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

Re-compute new cluster centroid using the means;

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

A. Computational Costs Required (CCR).

In the following figures, the computational cost required by the new algorithm to discover the clusters is experimented according to various values of K. This parameter is necessary to identify the fastest algorithm which is very desirable in main Data Mining applications that deal with huge datasets. The computational time is evaluated according to two variables: the number of observations N of each dataset and the number of cluster K. For each conducted experiment, we considered the same initial centroids between the two approaches, so it is possible to eliminate the influence of that starting condition on the final results.

According to the previous results, the new experimented algorithm requires lower computational cost than the k-means for all the experiments executed. The difference between the two approaches is significant especially for high values of K. This highlights the importance of the new proposed algorithm in enhancing the capabilities of the original k-means algorithm. In the second part of the experiments, the number of runs before convergence for the two algorithms was also tested, and the corresponding results are provided in Table II:

TABLE II
NUMBER OF TERMS REQUIRED BEFORE CONVERGENCE

Algorithm	D1	D2	D3	D 4
***k*-means**	6-14	5-13	7-32	7-18
Proposed approach	2			

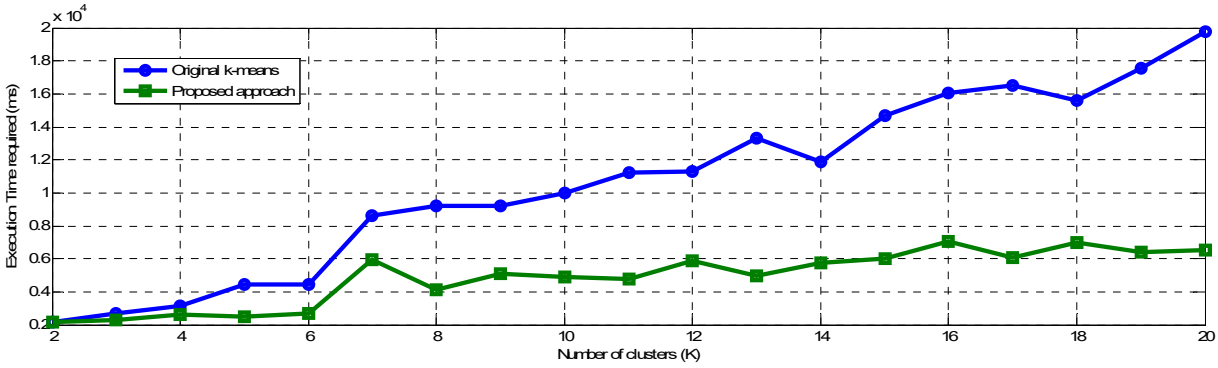

Fig. 1 Computational cost required to identify the K clusters of the dataset ALGERIA

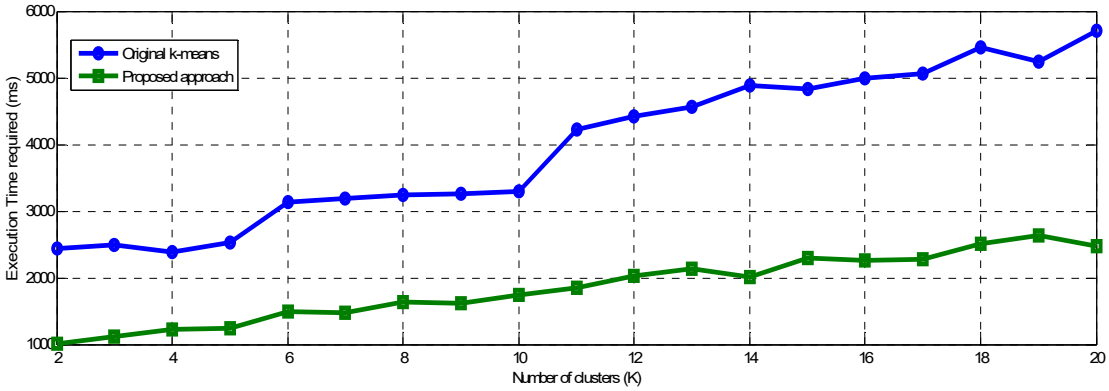

Fig. 2 Computational cost required to identify the K clusters of the dataset EGYPT

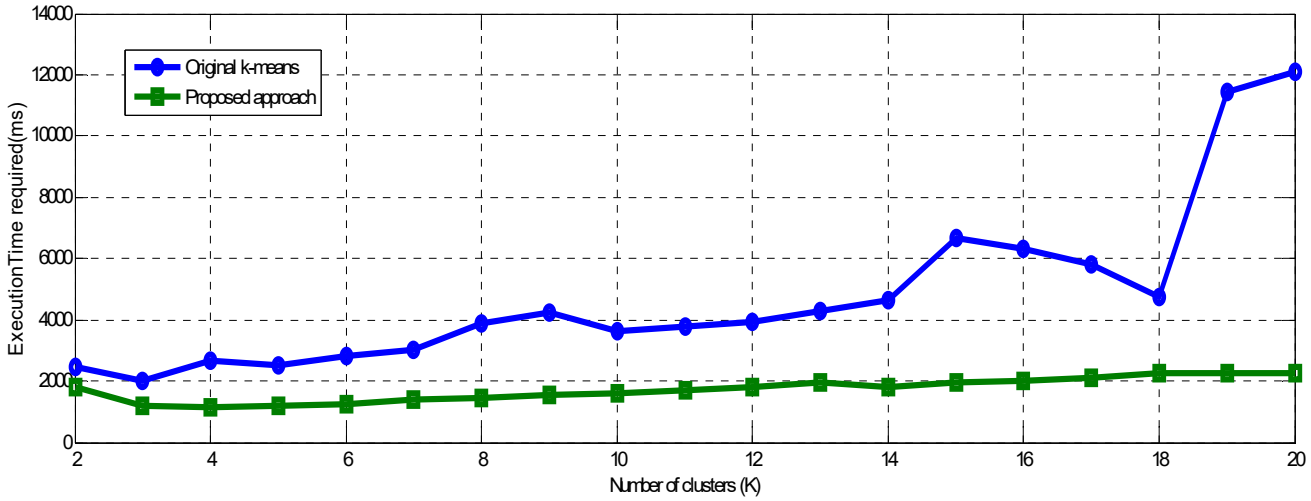

Fig. 3 Computational cost required to identify the *K* clusters of the dataset LYBIA

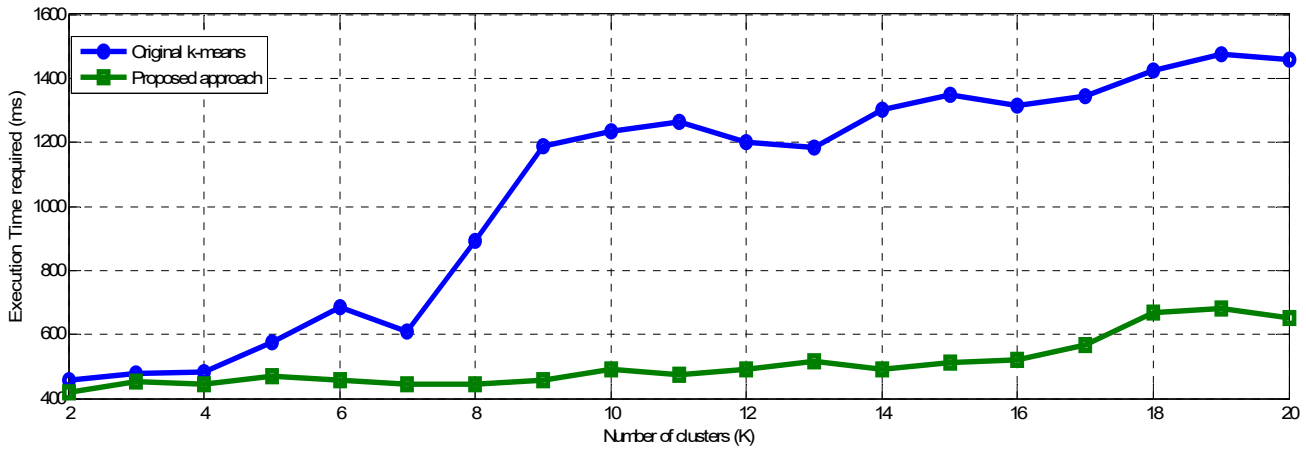

Fig. 4 Computational cost required to identify the *K* clusters of the dataset TUNISIA

According to the results, it is obvious that the original *k*-means requires higher iterations to converge than the new approach. This parameter is equal to 2, which makes it even more robust and scalable.

B. Accuracy Performance Evaluation.

Cluster validation evaluates the results using specific measures [22]. An external criterion called the accuracy or purity is used to quantify the similarity between the clusterings. A higher accuracy corresponds to better clustering. The accuracy is defined for a clustering $\mathcal{C}_i$ and j classes C_j as:

$$a\left(\mathcal{C}_i, C_j\right) = \frac{1}{N}\sum_{i=1}^{K} max_j \left|\mathcal{C}_i \cap C_j\right| \quad (8)$$

Figs. 9-12 present the efficiency according to these experiments.

Previous results show that the proposed approach provides better results than the *k*-means. Although, for some experiments, the values of the accuracy computed are almost identical, it is evident that, in the leading cases, the results favor the proposed approach when compared with the *k*-means.

The computed values of the accuracy presented in Figs. 5-8 can be divided into three groups reported in Table III. Evaluating the number and average of cases corresponding to each group is an interesting issue in order to estimate the effectiveness of our contribution.

TABLE III
NUMBER OF CASES COMPARING CLUSTERINGS

Comparing cases	$a_{new}=a_{k\text{-}means}$	$a_{new}>a_{k\text{-}means}$	$a_{new}<a_{k\text{-}means}$
NB cases	12	55	9
	15.79%	72.36%	11.84%

TABLE IV
AN AVERAGE OF NON-EMPTY CLUSTERS

Dataset	$\beta_{EUC} \leq \beta_{MAN}$	$\beta_{EUC} > \beta_{MAN}$	NB_experiments
ALGERIA	127 (79.37%)	33(20.62%)	160
EGYPT	99 (76.74%)	30 (23.25%)	129
LYBIA	95 (69.34%)	42 (30.66%)	137
TUNISIA	45 (70.31%)	19 (29.68%)	64

According to the results, 72.36% of the results correspond to good clustering ($a_{new}>a_{k\text{-}means}$). The case where $a_{new}=a_{k\text{-}means}$

can be considered as a good clustering since the new approach guarantees at least the same results than the *k*-means.

The number of non-empty clusters β also experimented. In fact, better clustering corresponds to a reduced number of groups since the elements would be more compact and their distribution denser over the clusters. We computed this parameter for various values of K (2→20) and reported the results in Table IV.

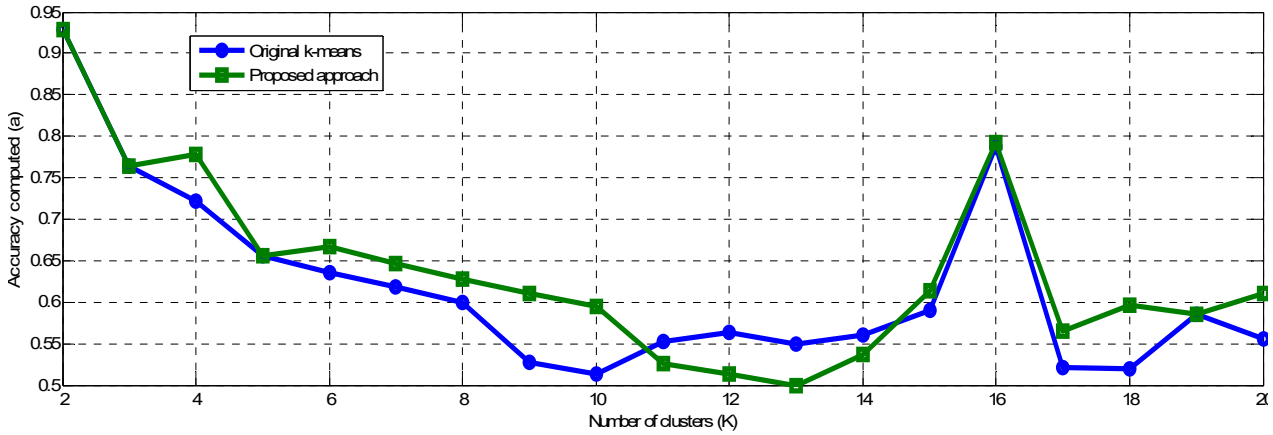

Fig. 5 Accuracy computed to identify the K clusters of the dataset ALGERIA

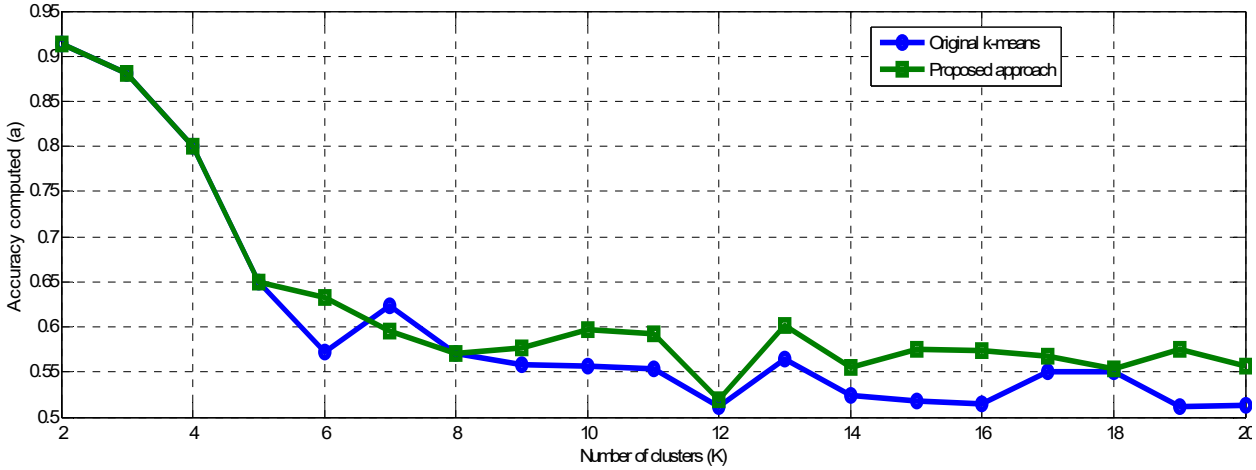

Fig. 6 Accuracy computed to identify the K clusters of the dataset EGYPT

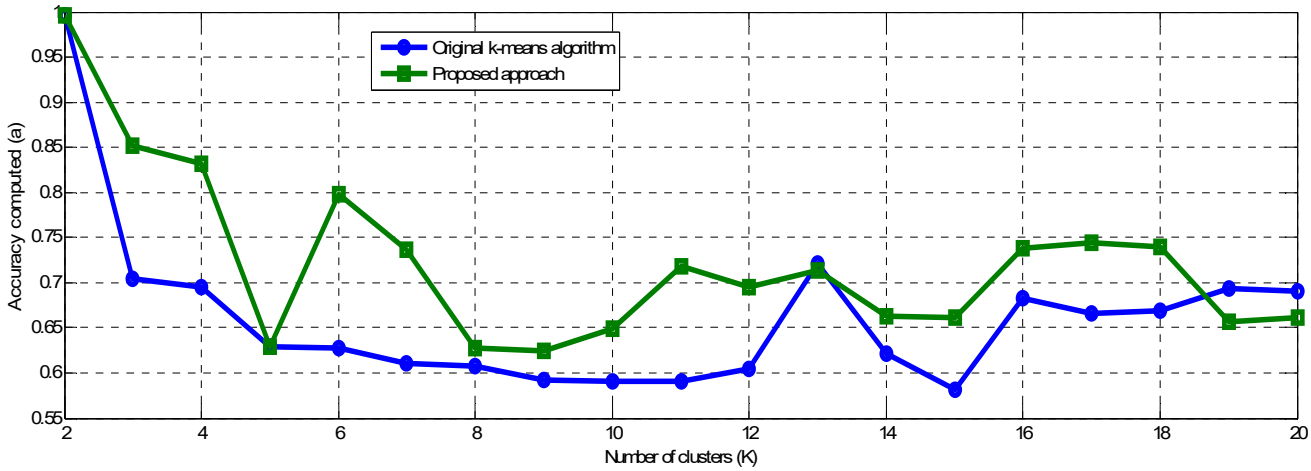

Fig. 7 Accuracy computed to identify the K clusters of the dataset LYBIA

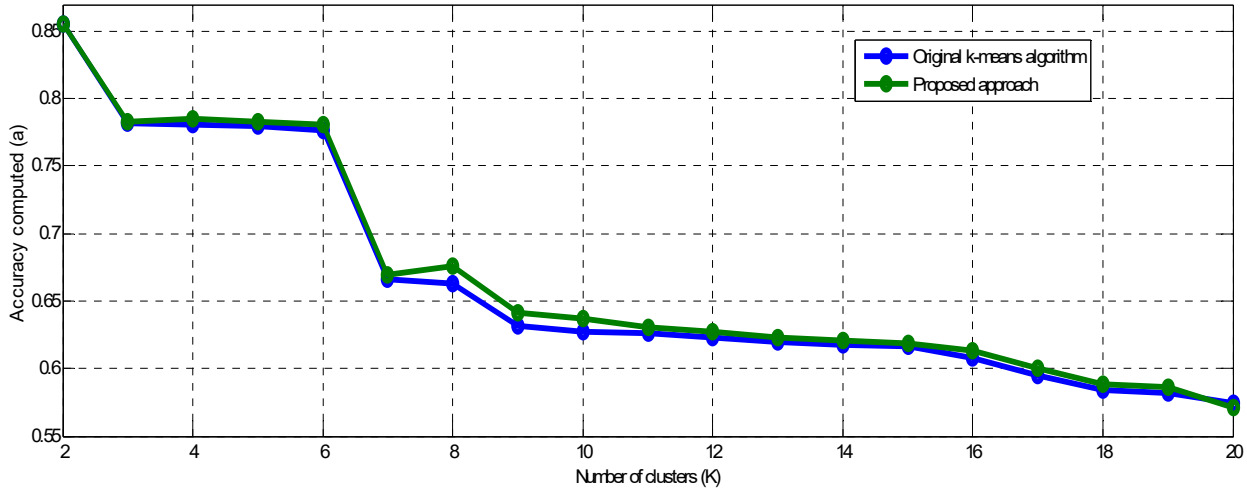

Fig. 8 Accuracy computed to identify the *K* clusters of the dataset TUNISIA

According to Table IV, the proposed approach provides more impressive results when compared with the *k*-means since it arranges the observations in a lower number of clusters, and consequently better clustering results could be provided.

V. Conclusion

The *k*-means is an unsupervised well known the partitional algorithm used in clustering for pattern recognition. Great research on the technical issues to enhance the computational complexity and accuracy of the algorithm was developed. In this paper, a modified version of the original *k*-means is discussed. The proposed approach uses the City Block and a new stop criterion to identify most accurate clustering. Experiments were conducted to evaluate the performance and the accuracy of the new algorithm with various numbers of initial clusters *K*. Obtained results show that it is more efficient than the *k*-means and provides more scalable and robust results.

References

[1] De Bruin, J. S, Cocx, T. K, Kosters, W. A, Laros, "*Data Mining approaches to criminal career analysis*." In Proceedings of the 6th International Conference on Data Mining ICDM'06, pp 11-18, 2006.

[2] T. Abraham and O. de Vel, "*Investigating profiling with computer forensic log data and associations rules.*" Proceedings of the IEEE International Conference on Data Mining (ICDM'06), pp 11-18, 2006.

[3] Jiawei Han M. K, "*Data Mining concepts and techniques.*" Morgan Kaufmann Publishers, An Imprint of Elsevier, 2006.

[4] Huang Z, "*Extension to the k-means algorithm for clustering large datasets with categorical values*", Data Mining and Knowledge Discovery, (2):283-304, 1998.

[5] Amir Ahmad, Lipika Dey, "*A k-means clustering algorithm for mixed numeric and categorical data.*" Data and Knowledge Engineering 63, pp 503-527, 2007.

[6] V. Ganti, J. E Gekhre, R. Ramakrishnan, "*CACTUS clustering categorical data using summaries*", Proceedings of the 5th ACM SIGKDD International Conference On Knowledge Discovery and Data Mining, 1999, pp 73-83.

[7] T. Zhang, R. Ramakrishnan, M. Livny, "*BIRCH: an efficient data clustering method for very large databases.*" SIGMOD Conference, 1996, pp 130-114.

[8] Dong Kuan Xu, Yingjie Tian, "*A Comprehensive Survey of Clustering Algorithms*", Ann. Data. Sci. , Springer-Verlag Berlin Heidelberg 2015, DOI 10.1007/s40745-015-0040-1

[9] Celebi M E, Kingravi H A Vela P A, "*A comparative study of efficient initialization methods for the k-means clustering algorithm*". Expert Systems with Applications 40:200–210, 2013.

[10] Celebi M E, Kingravi H, "*Deterministic initialization of the K-means algorithm using hierarchical clustering*", International Journal of Pattern Recognition and Artificial Intelligence 26(7):1250018, 2012.

[11] Celebi M E, Kingravi H, *"Linear, deterministic, and order-invariant initialization methods for the K-means clustering algorithm."* Celebi M E (ed) Partitional clustering algorithms. Springer, Berlin, pp 79–98, 2014.

[12] Kalogeratos A, Likas A, "*Dip-means: an incremental clustering method for estimating the number of clusters.*" In: Advances in neural information processing systems (NIPS), pp 2402–2410, 2012.

[13] Tzortzis G, Likas A, "*The Min-Max k-Means clustering algorithm*". Pattern Recognition 47:2505–2516-2014.

[14] Eslamnezhad M, Varjani A Y, "*Intrusion detection based on Min-Max K-means clustering.*" In 7th International symposium on telecommunications (IST'2014), pp 804–808-2014.

[15] Yuan F, Meng Z. H, Zhang H, X and Dong C. R, "*A new algorithm to get the initial centroids.*" Proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pages 26-29, 2004.

[16] Xiaoyan Wang, Yanping Bai, "*The global Min-Max k-means algorithm*", Wang and Bai *SpringerPlus 5:1665*, DOI 10.1186/s40064-016-3329-4-2016.

[17] Zengyou He, Shengchun Deng "*Improving K-modes Algorithm considering frequencies of attributes values in mode.*" Conference paper in Lecture notes in computer science, December 2005.

[18] G. La Free, "*The Global Terrorism Database: Accomplishments and Challenges*", Perspectives on Terrorism, Vol. 4 (2010).

[19] X. Wang, E. Miller, K. Smarick, W. Ribarsky and R. Chang, "*Investigative Visual Analysis of Global terrorism.*", Proceeding of the 10th Joint Eurographics/ IEEE-VGTC conference on Visualization, Vol. 27 (2008): 919-926.

[20] M. Adnan, M. Rafi, "*Extracting patterns from Global Terrorism Database (GTD) sing co-clustering approach.*" Journal of independent studies and research computing, Volume 13, 2015.

[21] 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.

[22] Silke Wagner, Dorothea Wagner, "Comparing Clusterings-An Overview", January 12, 2007.