RESEARCH ARTICLE OPEN ACCESS

Enhanced Kprototypes EKP and FCM for Improving Clustering of the Large Mixed Data

Faten Alkrdy [1], Jabr Hanna [2]
PhD Student [1], Professor [2]
Computer Engineering
Tishreen University
Syria

ABSTRACT
Data mining is one of the most significant and developing trends in computer science world. It includes several concepts such as classification and clustering. In the current research we focus on the clustering methods. In this paper, we introduces a new enhanced Kprototypes algorithm (EKP) with three enhancements. The first is the centroids initialization methods, the second is the convergence condition which is connected to the distance between the new and old centroids, while the third is vectorization process which minimize the time complexity of KP significantly. We also modified the FCM algorithm to make the distance computation step more effective. We applied our test on two dataset (Adult and CoverType) which are large mixed dataset and one of them included missed values which treated by two different ways. In the experiments, we defined the best clustering algorithm, compare our results with traditional algorithms and the international studies. The results showed that EKP achieved 4.769s only to converge and 68±4% clustering accuracy. The results showed that EKP was better than EKP at CoverType dataset. The best initialization method was the space attribute-splitting with minimum time and maximum accuracy. We compared our research with the previous ones and found that our research achieved the best time and accuracy clustering.

Keywords :— Data Mining, Kprototypes, FCM, Kmeans, Clustering, Adult Dataset, CoverType Dataset.

I. INTRODUCTION

Data mining is one of the computer science trends and it takes the importance of researchers due to its benefits and importance in many applications like analysing big dataset to investigate acknowledge about weather prediction, military application, statistics of study centers, educational concepts etc.

With the presence of large amounts of data stored in databases and data stores, the need to develop powerful tools for analyzing data and extracting information and knowledge has increased. Hence, data mining has emerged as a technique aimed at extracting knowledge from vast amounts of data. But although these techniques have led to good knowledge and useful databases, they are time-consuming in the exploration process, making their effectiveness limited.

Our goal in the current research is to minimize the clustering time and improve the performance in order to make the data mining process more effective and robust.

II. RELATED WORK

There are many studies in the field of clustering of big data. Kolen et al. [1], modified FCM to make it depending on linear complexity approach. They minimized the time complexity from $O(NC^{2p})$ into $O(NCP)$. The algorithm took 90.67s for each iteration instead of 745s. However, their algorithm added a new time complexity and didn't minimize time effectively as they said.

Sun et al. [2], used the FCM and a series of integer indexes to generate the initial centroids. His algorithm minimize time but didn't add any performance accuracy enhancement. The time minimized by 48% off the original algorithm. The used dataset in both Kolen and Sun studies were very small (no more than 500 data points).

Liu [3], in 2008, applied FCM and partitioning of the data points space to enhance the initial centroids generation. The accuracy of clustering improved in case of little datasets only.

Chen et al [4] at 2011 used the Landmark-based Spectral Clustering using random sampling to select landmarks (LSC-R). They applied their experiments on CoverType dataset. They got only 24.75% clustering accuracy and 134.71s as average converge time.

Celebi [5] at 2013, modified FCM in order to generate initial centroids based on the idea of that each data point is a centre of other neighbours. They could define K number of clusters for dataset. Their method lack the application on real datasets.

Steteco et al [6], in 2015, used an improved FCM to minimize the iteration of FCM. They reduced the iterations of FCM by Factor 2.1 off the original FCM. They used very small dataset.

In 2014, Cheung [7] used the Kprototypes algorithm. He achieved a medium clustering time and high performance on adult dataset. He got 61.45% ± 1.4% accuracy and 15s average time.

Kim [8] at 2017 used the speed Kprototypes to minimize the calculations of distances. He minimized time significantly but on the other hand the accuracy degraded. The one iteration took 5.8 seconds.

Jang [9], at 2018, used the Efficient Grid-Based K-Prototypes for clustering a randomized 10000 data points. He achieved a time improvement against Kim's method where the iteration took only 3.5s and less.

All the previous studies lacks many important research points such as some of them depended only on one type of datasets. Other studies processed the time or the accuracy. A very little studies focused on them both. Very little studies treated the problem of detecting the best ways of centroids initialization methods. Some other studies [10,11,12,1314] used the classification concepts which are another different type of data mining.

In the current study we try to deal with all that points together to make a robust clustering system which is effective in both time and accuracy.

III. MATERIALS AND METHODS

In this study, we used and improved two different data mining algorithms which are Kprototypes (KP) and Fuzzy C-Means (FCM).

A. The New Enhanced Kprototypes (EKP)

We modified three stages of the original KP [15]. First, we replaced the traditional random centroids initialization with a new approaches which could initialize centroids with more fit values and improve the performance.

The second modification is about how the algorithm converge and stop iteration. In the original KP, there was a number of iteration to perform clustering. In our method, we related the convergence of KP with the difference between the old and new centroids.

The third enhancement we made is the speedup of the algorithm by applying the vectorization process on the KP stages. Figure 1 illustrates the modified KP block diagram.

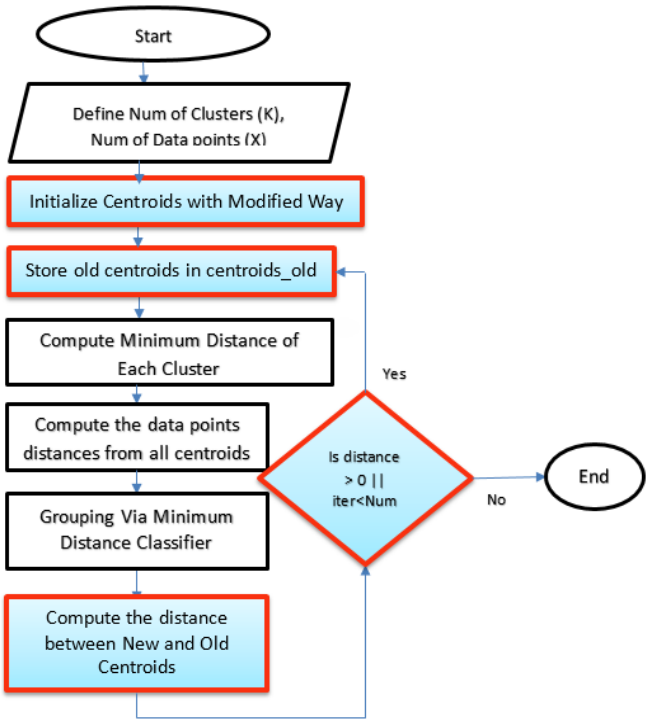
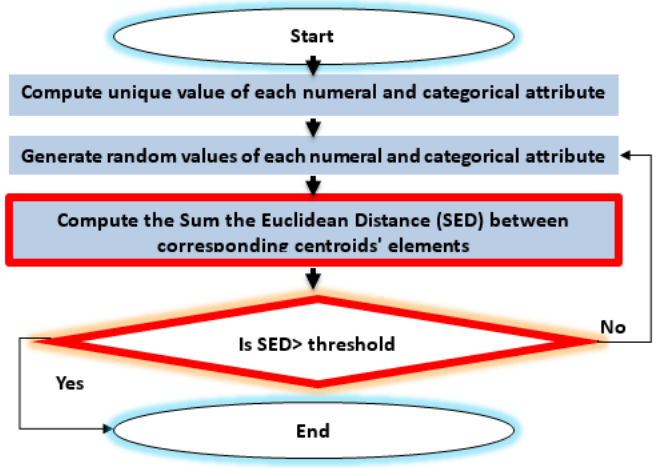


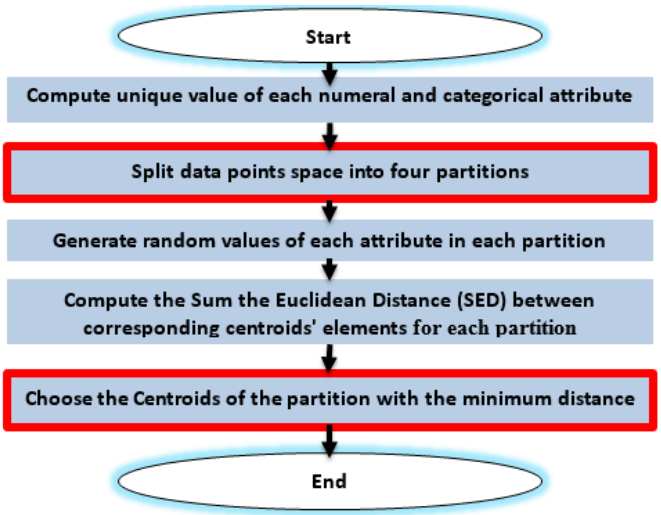
Figure 1. The Suggested Modified KP algorithm

The modifications are with color red on figure 1.

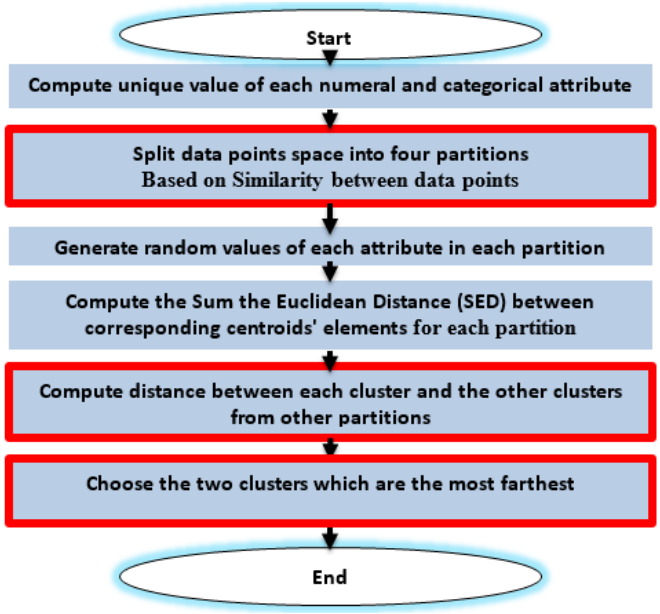
We propose three different initialization approaches. Figure2 shows them.



A.



B.



C.

Figure 2. The Iniial Centroid Generation: A) the randomized Euclidean distance way, B) the randomized Euclidean distance based on space splitting way, C) the randomized Euclidean distance based on space attribute-splitting way.

The vectorization process is performed on KP algorithm to minimize the time complexity of KP which is O(N*M*K*I) [13] where N is the total number of data points, M is the number of attributes, k is the number of clusters and I is the number of iterations. The time complexity is minimized by Z factor which is defined by Amdal's Law [16].

$O_{new}^{KP}=O(N*M*K*I)/Z;\ z=1/((1-F)+F/S)$ (1)

Figure 3 shows a simple example of the vectorization process.

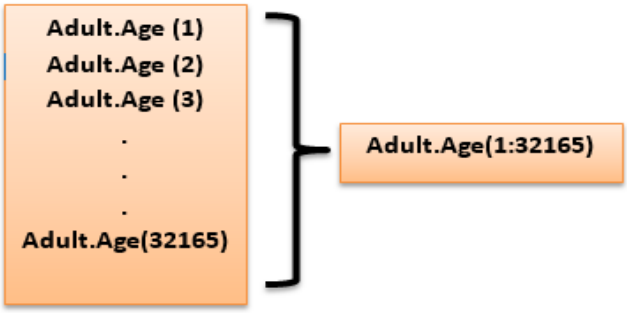


Figure 3. Example of vectorization process on KP algorithm.

B. The Modified FCM algorithm

We modified the original FCM algorithm [17] through two modifications. The first was the preparing of data which included the suntituion of missed values and the normaliztion of data points. The second modification was the way to compute the new distances from all centroids by the equation (2) which achieved more cluster separation.

$mf_{new}=1./(1+expo.\hat{}(U))$ (2)

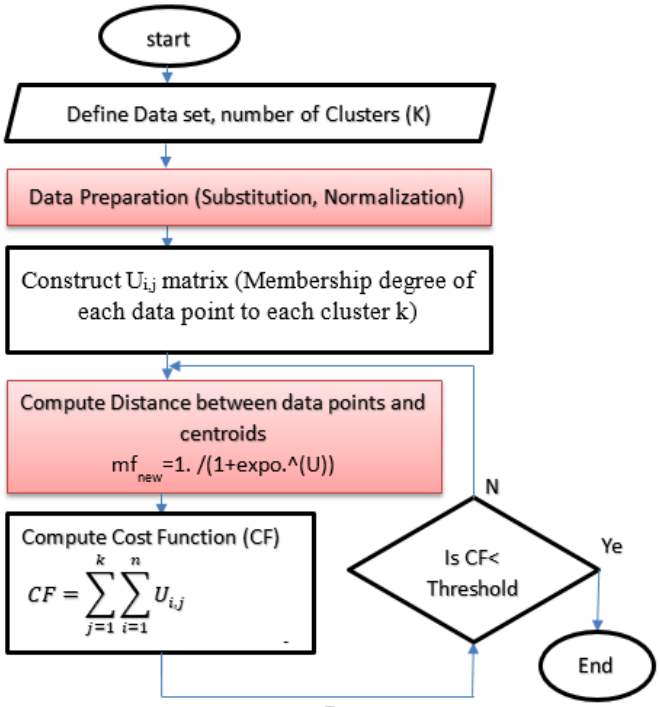


Figure 4. The Modified FCM Algorithm

C. The Used datasets:

In the current research, we used two different datasets. The first is the Adult dataset [18] which includes 8 numerical attributes and 6 categorical attributes. It also includes a missed values. Its Prediction task is to determine whether a person makes over 50K a year.

The second dataset is the CoverType dataset [19] which consists of 581012 records and 12 attributes. It is used to predict the cover type of Roosevelt National Forest of northern Colorado.

Table 1 includes detailed information about both datasets.

Table 1. Statistics of the used mixed data sets.

Dataset	Instances	Attributes	Class
Adult	30162	8 (cat) + 6 (num)	2
CoverType	581012	2 (cat) + 10 (num)	

D. Performance Metrics:

We used some evaluation metrics for the Kprototypes and fuzzy clustering techniques.

Equation 3 shows how can we compute the clustering accuracy while equation 2 explains the Rand_Index metric [20].

$$Acc = \frac{\sum_{i=1}^{N} \delta(c_i, map(l_i))}{N} \quad (3)$$

Where N is the number of data points, C_i is the cluster number which the data point belong to, map(l_i) is the cluster number in which the data points classified, and δ is a number with two conditions (1 in case of matching between the classified and original cluster number, 0 otherwise).

Second, we used the Rand Index metric. Given a set of n elements S={o{1},... ,o{n}} and two partitions of S to compare, X={X{1},... ,X{r}}, a partition of S into r subsets, and Y={Y{1},... ,Y{s}}} a partition of S into s subsets, define the following [21]:

$$R=(a+b)/(a+b+c+d) \quad (4)$$

Where a is the number of pairs of elements in S that are in the same subset in X and in the same subset in Y, b is the number of pairs of elements in S that are in different subsets in X and in different subsets in Y, c is the number of pairs of elements in S that are in the same subset in X and in different subsets in Y and d is the number of pairs of elements in S that are in different subsets in X and in the same subset in Y.

Third, we used the following fuzzy metrics [22,23,24,25]:

Partition coefficient: which compute the intersection between clusters.

$$PC = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n} u_{i,j}^2 \quad (5)$$

where $U_{i,j}$ is the membership matrix.

Coefficient entropy: compute the fuzziness of the clustering.

$$CE = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n} u_{i,j}\log_a(u_{i,j}) \quad (6)$$

Inner Scattering: compute the degree of correlation of data points inside the cluster.

$$Scatt(n_c) = \frac{1}{n_c}\sum_{i=1}^{n_c}\|\sigma(v_i)\|/\|\sigma(X)\| \quad (7)$$

Where n_c is the number of clusters.

The cluster variance is defined as follows:

$$s_{v_i}^p = \sum_{k=1}^{n_i}(x_k^p - v_i^p)^2 / n_i \quad (8)$$

Where X_k^p is the data point of P column and K raw in dataset. V_i^p is the centre of i^{th} cluster while n_i is the number of clusters.

The inner variance is gives as follows:

$$s_{v_i}^p = \frac{1}{n}\sum_{k=1}^{n_i}(x_k^p - \bar{x}^p)^2 \quad (9)$$

Where $\bar{X}^P$ is the mean value of the all data.

Outer Scattering: computes the distance between clusters. It is given by equation 10.

$$Dis(n_c) = \frac{D_{max}}{D_{min}}\sum_{k=1}^{n_c}\left(\sum_{z=1}^{n_c}\|v_k - v_z\|\right)^{-1} \quad (10)$$

Where D_{min}, D_{max} are the minimum and the maximum distances between centroids of clusters z and k.

E. Test Scenarios:

For the test step, we suggested many scenarios which include comparatives between traditional and modified algorithms besides the comparatives between our algorithms and the other ones at the same datasets.

The test scenarios includes the following:

i. Compare EKP with traditional one.
ii. Compare the suggested initialization approaches and define the best one.
iii. Compare the missed values approaches.
iv. Evaluate FCM on adult dataset based on the fuzzy performance metrics.
v. Evaluate EKP on CoverType dataset under different iterations.
vi. Evaluate FCM on CoverType dataset using fuzzy performance metrics.
vii. Compare EKP and FCM.
viii. Compare our algorithms with the recent published research on both adult and CoverType datasets.

IV. RESULTS

The results were carried out on a laptop with Intel core i5 processor and 4 GB of RAM.

A. Adult dataset results:

Here, we applied the first four test scenarios. Table 2 includes the result of applying KP on adult dataset before and after the suggested modification. We used the number 10 as a maximum iteration for KP.

Table 2. KP Time Comparative before and after The Suggested Enhancement.

Status/Time	Prepare Data Time	Missed Values Processing Time (s)	Normalization Time (s)	Clustering Time (s)	
				With Stop Condition	Without
Enhanced KP (EKP) (With Vectorization)	1.129	1.025	0.106	4.769	8.998
Original KP	1.13 s	223.410 s	1050.167 s	>20 Minutes	

The enhanced KP algorithm achieved 68.33% clustering precision compared to 64.9% for Kmeans.

We needed only 4.769 seconds until KP converges.

The next scenarios were to define the best initialization algorithm for the initial centroids generation process.

Figure 5 shows the difference between the four suggested approaches on Adult dataset.

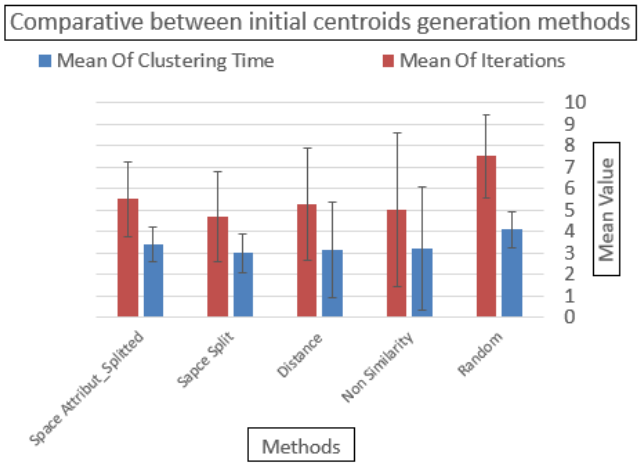
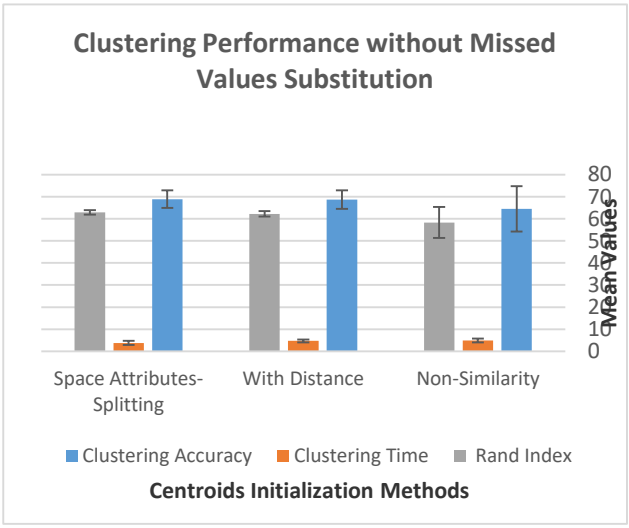


Figure 5. A Comparison between Initial Centroids Generation Suggested Approaches on Adult Dataset.

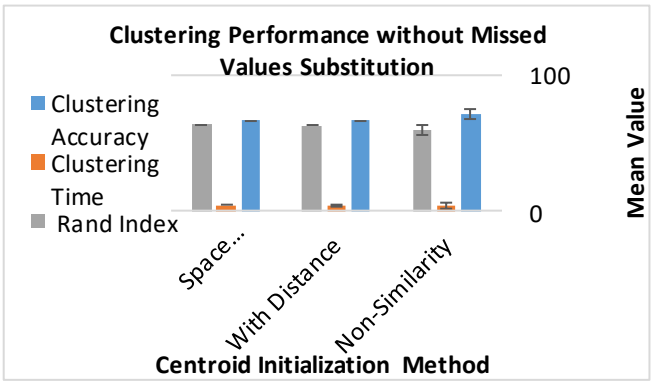
In other point of view, the time comparison between that four initialization approaches shows that the non-similarity approach is the most expensive approach with mean-time equals to 9.517 s, while the space splitting is the least one with mean-time equals to 0.78 s.

If we discuss these two results, we find that the splitting way which depends on the attributes similarity to portioning the space is the best way from time and accuracy point of view.

The last scenario of the KP algorithm on the adult data set is to define the nature of missed values of this dataset and the best way to treat them. We suggest to use two principles; the first lets them missing while the second substitute them with the ways introduced by our previous research on the same dataset [26]. The second way replace the numerical values with the mean value of all attribute values, and the categorical values with the most frequently value through them. Some of the missed values is inferred by other attributes values. The "Education" missing values are replaced by the "Education_num" attribute values and vice versa. The "Relationship" missed valued are substituted by means of "Material_status" attribute values and vice versa. Figure 6 shows the accuracy of enhanced KP algorithm under the two approaches of missed value processing.



a.



b.

Figure 6. Accuracy of Enhanced KP Algorithm under The Two Approaches of Missed Value Processing (a): Without any way, (B) with the suggested substitution method.

The results shows that the type of Adult dataset is "Missing completely at random (MCAR)" and the "Missing at random (MAR)" because some missed values can be inferred from the others while other missed values should be substituted with other way (mean, mode,…).

The results also illustrates that the accuracy didn't change significantly.

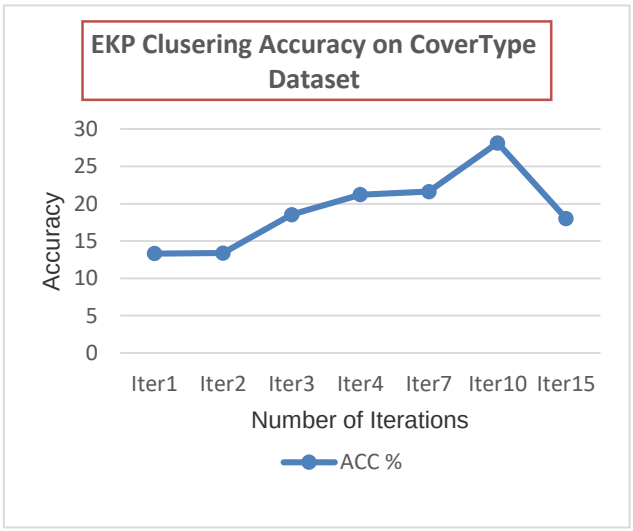
For the modified FCM clustering algorithm, we find that the clustering spends 2.449 seconds to converge after 50 iterations. Table 3 includes the performance evaluation of modified FCM on adult dataset. While table 4 shows a comparison between KP and FCM on adult dataset.

Table 3. The performance evaluation of modified FCM on adult dataset.

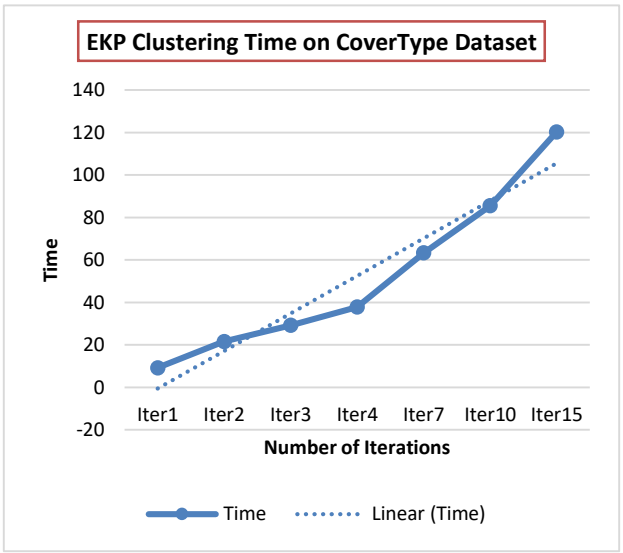
Attributes	Number of Clusters	Clustering Time	Performance Metrics

Age SurvyWeight CapitalGain CapitalLoss HoursPerWeek NativeCountry	2	2.449 s	Partition coefficient (similarity) PC: 0.1844 Coefficient Entropy (CE): -0.2091 (max=0.3) Inner Scattering: 0.7533 Outer Scattering: 3.6731 Clustering Precision: 55.09 %

The KP algorithm achieves accuracy 65.83% with clustering time 3.65 s, while FCM gets 55.09%, 3,74s at iteration 75 for accuracy and time respectively.



a.



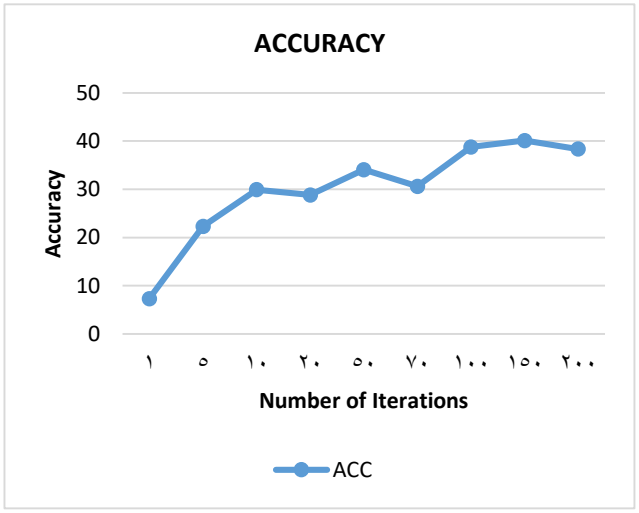
b.

Figure 7. EKP performance on CoverType dataset: (a) Clustering Accuracy, (B) Clustering Time.

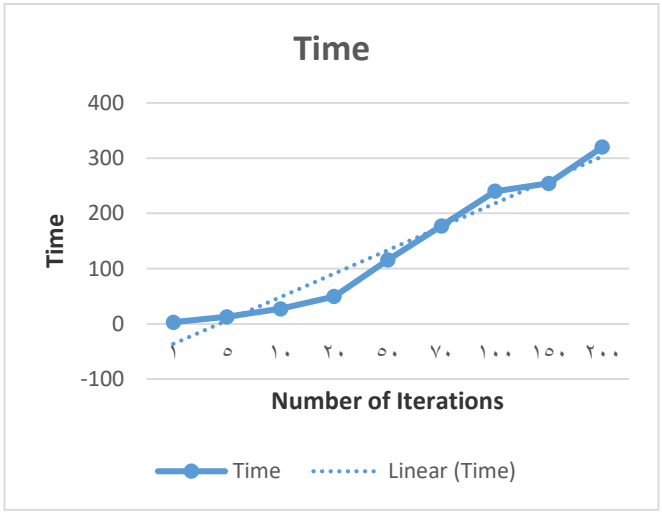
***B.* CoverType Dataset Scenarios:**

For the CoverType dataset, we introduces many scenarios of KP and FCM. Figure 7 (a,b) show the EKP time and accuracy results on CoverType dataset. On the other hand, figure 8(a,b) includes the modified FCM clustering results.

It can be noticed that the average accuracy time is linear and not exponential which approve that our enhancement minimize the accuracy time.



a.



b.

Figure 8. Modified FCM performance on CoverType dataset: (a) Clustering Accuracy, (B) Clustering Time.

Figure 9 shows the clustering result on CoverType dataset after 75 iterations of modified FCM.

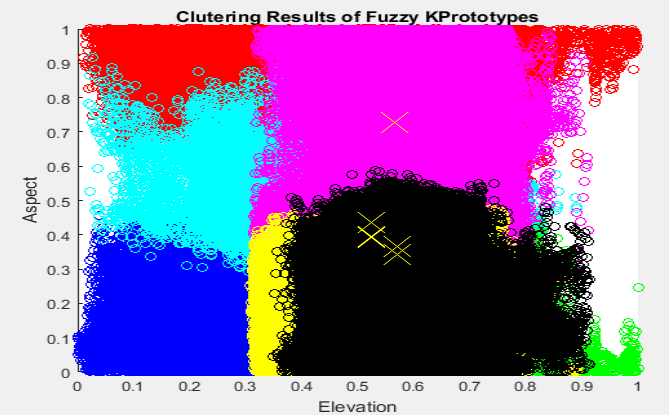


Figure 9. The clustering result on CoverType dataset after 75 iterations of modified FCM.

If we compare EKP and FCM clustering results on CoverType dataset, we can define that EKP has a better performance and this is due to the fact that KP is better than FCM with large data processing. Figure 10 shows that difference.

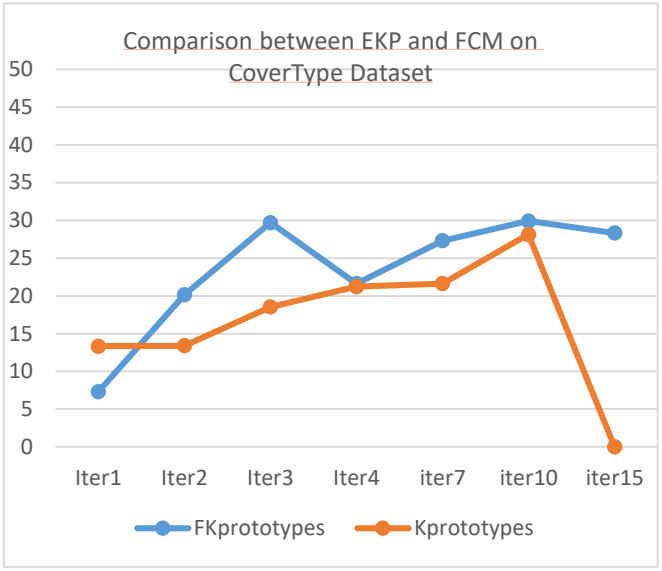


Figure 10. A Comparison between EKP and FCM on CoverType Dataset.

C. The Comparative Study with The International Researches.

The last step in our test scenarios is a comparison between our algorithms and the previous ones in the field of data mining and clustering. Table 4,5 include two comparatives between our algorithms and the previous studies on Adult and CoverType datasets.

Table 4. A Comparison between our Algorithms and the Previous Ones in Clustering on Adult Dataset.

Research	**Number of Records**	**Clustering Time**	**Accuracy (EKP)**	**Accuracy (Kmeans)**
Cheung	32162	15.279	61.45% ±	61.31%
[7]		5 s	1.4%	
Current Research	32465	1.129 s	68.33% ± 4%	64.9%

Table 5. A Comparison between our Algorithms and the Previous Ones in Clustering on CoverType Dataset.

Research	**Clustering Time (S)**	**Accuracy (%)**
Chen (Spectral Clustering) [4]	181006.17	44.24
Yan (KASP) [27]	360.07	22.42
Shinnou (CSC) [28]	402.14	21.65
Chen (Nystrom) [4]	258.25	22.31
Chen (LSC-R) [4]	134.71	24.75
Chen (LSC-K) [4]	615.84	25.5
Li (SeqSC(200) [29]	980	28.5
Li (SeqSC(400) [29]	620	27.8
Li (SeqSC(600) [28]	420	27.2
Li (SeqSC(800) [28]	230	27
Current Research (EKP)	85.55	28.14
Current Research (FCM)	254.3	40.1

V. CONCLUSIONS

In this paper, we introduced a new Kprototypes algorithm with three enhancements. The first is the centroids initialization methods, the second is the convergence condition which is connected to the distance between the new and old centroids, while the third is vectorization process which minimize the time complexity of KP significantly.

We also modified the FCM in order to make more separation between clusters. We applied tests on two different large datasets (Adult and CoverType). Experimental tests shows that the new EKP is better than the original KP from time and accuracy point of view. The new EKP is also better than FCM for the CoverType dataset clustering. The initialization process which depends on the splitting of data space and the Euclidean Distance between data points is the best initialization method. The adult dataset includes missed values which is treated by our algorithms using two ways. The tests shows that no significant accuracy or time enhancement is achieved using the substitution methods for both numerical and categorical values (i.e. the adult dataset missed values are from the types MAR and MCAR).

The comparative study shows that our algorithms exceeded the previous studies in case of accuracy and time enhancement. The future work can be done on more time enhancement using other parallel processing techniques and the fusion of our algorithms with the evolutionary ones in order to make more accuracy enhancement.

ACKNOWLEDGMENT

I will thank Dr. Mariam Saii for her scientific assistance and cooperation with our work.

REFERENCES

[1] Kolen, J.F. and Hutcheson, T., 2002. Reducing the time complexity of the fuzzy c-means algorithm. IEEE Transactions on Fuzzy Systems, 10(2), pp.263-267.

[2] Sun, H., Wang, S. and Jiang, Q., 2004. FCM-based model selection algorithms for determining the number of clusters. Pattern recognition, 37(10), pp.2027-2037.

[3] Liu, T., Zhou, Y., Hu, Z. and Wang, Z., 2008, October. A new clustering algorithm based on artificial immune system. In Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on (Vol. 2, pp. 347-351). IEEE.

[4] Chen X, Cai D, Large Scale Spectral Clustering with Landmark-Based Representation, Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011, pp: 313-318.

[5] Celebi, M.E., Kingravi, H.A. and Vela, P.A., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert systems with applications, 40(1), pp.200-210.

[6] Steteco A, Xiao-Jun Zeng, John Keane, 2015, "Fuzzy C-Means++: Fuzzy C-Means With Effective Seeding Initialization", Science Direct, Expert Systems With Applications, pp:7541-7548

[7] Cheung Y, Jia H, " Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number", Pattern Recognition, 2014, pp:2228–2238.

[8] Kim B, A Fast K-prototypes Algorithm Using Partial Distance Computation, Symmetry, 2017, pp:1-10.

[9] Jang H, Kim B, Kim J, Jung S, An Efficient Grid-Based K-Prototypes Algorithm for Sustainable Decision-Making on Spatial Objects, Sustainability, 2018, pp:1-20.

[10] Mayya A, Saii M. "Hybrid Recognition System under Feature Selection and Fusion". International Journal of Computer Science Trends and Technology (IJCST) V5(4): Page(78-84) Jul - Aug 2017. ISSN: 2347-8578. www.ijcstjournal.org.Published by Eighth Sense Research Group.

[11] Mayya A , Saii M. "Iris and Palmprint Decision Fusion to Enhance Human Recognition". International Journal of Computer Science Trends and Technology (IJCST) V5(5): Page(42-46) Sep - Oct 2017. ISSN: 2347-8578. www.ijcstjournal.org.Published by Eighth Sense Research Group.

[12] Mayya AM, Saii M, (2016) Human recognition based on ear shape images using PCA-Wavelets and different classification methods. Med Devices Diagn Eng: DOI: 10.15761/MDDE.1000103.

[13] Mayya M, Saii M, Hand and Finger Fracture Detection and Type Classification Using Image Processing, 2018, IJCST V6(4): Page(68-76) Aug-2018. ISSN: 2347-8578. www.ijcstjournal.org.Published by Eighth Sense Research Group.

[14] Raneem Knaj, Jafar Al- kheir, Creating a High-Quality Syrian Audio Database for Analysis of Speaker Personality, IJCST, 2017, V5(4), Page(113-116) Aug-2018. ISSN: 2347-8578. www.ijcstjournal.org.Published by Eighth Sense Research Group.

[15] HUANG Z, clustering large data sets with mixed numeric and categorical values, n The First Pacific-Asia Conference on Knowledge Discovery and Data Mining, 1997.

[16] https://study.com/academy/lesson/amdahl-s-law-definition-formula-examples.html, last accessed 10-10-2018.

[17] Fernando Lobo, Fuzzy c-means algorithm, Data mining Lectures, available at www.fernandolobo.info/dm/slides/fuzzy-c-means.pdf.

[18] Adult Dataset, available at: https://archive.ics.uci.edu/ml/datasets/adult, last accessed 15-7-2017.

[19] Adult Dataset, available at: https://archive.ics.uci.edu/ml/datasets/covertype, last accessed 2-10-2017.

[20] Alvarez M, Design and Analysis of Clustering Algorithms for Numerical, Categorical and Mixed Data, PhD thesis at Manufacturing Engineering Centre Cardiff University, ProQuest LLC publishing, 2013.

[21] Jinchao ji, Pang W, Zhou C, Han X, Wang Z, A Fuzzy Kprototype Clustering Algorithm for Mixed Numeric and Categorical data, Neuro-Computing, 2013, pp:129-135.

[22] Boriah S, Chandola V, Kumar V, Similarity Measures for Categorical Data: A Comparative Evaluation, Department of Computer Science and Engineering University of Minnesota, 2008, pp:243-254.

[23] Halkidi M, Batistakis Y, Vazirgiannis M, On Clustering Validation Techniques, Journal of Intelligent Information Systems, 31(3), 2001, pp:107-145.

[24] Halkidi M, Batistakis Y, Vazirgiannis M, Clustering Validity Checking Methods: Part II, ACM SIGMOD Record, 31(3), 2002.

[25] Aljawady R, Altaleb Gh, Clustering Validity, AL-Rafidain Journal of Computer Sciences and Mathematics , 5(2), 2008, pp:79-94.

[26] Alkrdy F, Hanna J, Time and Accuracy Enhancement Of Mixed Data Sets Using Data Mining Algorithms, Al-baath University Journal, 2017.

[27] Yan, D.; Huang, L.; and Jordan, M.I. 2009. Fast approximate spectral clustering. In Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'09).

[28] Shinnou, H., and Sasaki, M. 2008. Spectral clustering for a large data set by reducing the similarity matrix size. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08).

[29] Li Y, Huang J, Liu W, Scalable Sequential Spectral Clustering, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), 2016, 1809-1815.

[30] Suleiman A, Aldibaja I, Improving Web Cache Algorithm Using Semantic Similarity Measures, IJCST, 5(5), 12-17, 2017.