An Efficient Algorithm for Initializing Centroids in K-means Clustering

Dr. Ahmed Hussain Aliwy
Dept. of Computer Science
Faculty of CS and Mathematics, University of Kufa
Al-Najaf, Iraq
ahmed_7425@yahoo.com

Dr. Kadhim B. S. Aljanabi
Dept. of Computer Science
Faculty of CS and Mathematics, University of Kufa
Al-Najaf, Iraq
kadhim.aljanabi@uokufa.edu.iq

***Abstract*—**Clustering represents one of the most popular knowledge extraction algorithms in data mining techniques. Hierarchical and partitioning approaches are widely used in this field. Each has its own advantages, drawbacks and goals. K-means represents the most popular partitioning clustering technique, however it suffers from two major drawbacks; time complexity and its sensitivity to the initial centroid values. The work in this paper presents an approach for estimating the starting initial centroids throughout three process including density based, normalization and smoothing ideas. The proposed algorithm has a strong mathematical foundation.

The proposed approach was tested using a free standard data (20000 records). The results showed that the approach has better complexity and ensures the clustering convergence.

***Keywords*—Data Mining, Clustering, K-means, Centroids, Complexity**

I. INTRODUCTION

Data Mining (DM) is an interdisciplinary fields of statistics, computer science, Artificial intelligence, visualization, and many others. It is the computational process of discovering patterns in large data sets and its main goal is to extract knowledge from a data set and convert it into an understandable structure for further use. The main techniques of DM can be summarized into classification, association, clustering and prediction. Each has its own goals and algorithms [1,2,3,4].

Among all DM techniques, clustering represent one of the main and widely used concepts since it aims to find out hidden knowledge in huge data sets without any predefined attributes, and it is of great importance for the wide range of applications from which health care, business, marketing, bioinformatics, natural languages, recognition and many others [5,6,7,8].

A clustering algorithm attempts to find natural groups of components (or data) based on some similarity. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of objects in each cluster.

The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters. Generally, the distance between two points is taken as a common metric to assess the similarity among the components of a population.

The commonly used distance measure is the Euclidean metric which defines the distance between two points p and q where p= (p_1, p_2, ...) and q = (q_1, q_2, ...) is given by:

$$d = \sqrt{\sum_{i=1}^{k}(p_i - q_i)^2} \qquad (1)$$

The definitions of distance functions are usually very different for interval-scaled, Boolean, categorical, ordinal and ratio variables[1,2,3].

Distances are normally used to measure the similarity or dissimilarity between two data objects.

The mostpopular method is calledMinkowski distance where the distance is given by:

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \cdots + |x_{ip} - x_{jp}|^q)} \qquad (2)$$

Where (x_{i1}, x_{i2}, …,x_{ip}) and (x_{j1}, x_{j2}, …, x_{jp}) are two p-dimensional data objects, and q is a positive integer.If q = 1, d is called Manhattan distance and is given by:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}| \qquad (3)$$

If q = 2, d is called Euclidean distance as in equation (1).

The major clustering algorithms can be classified into hierarchical and partitioning[1,2]. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem using the partitioning technique. The mathematical foundation of K-means clustering algorithm can be stated as follows[9]:

In order to partition n data points into k disjoint subsets (clusters) $S_j$containingn_j data points, then the following Sum ofSquares Error criterion(SSE) is to be minimized:

$$SSE = \sum_{j=1}^{k} \sum_{n \in S_j} |x_n - \mu_j|^2 \quad (4)$$

Where x_(n)is a vector representing the n^{th} data point and μ_j is the centroid of the data points in S_j.

The procedure follows a simple and easy way to classify a given data set through initially defined number of clusters. The main idea is to define k centroids, one for each cluster. The better choice of the initial centroids is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid(many approaches are available such as Euclidian distance). The process is repeated until when no point is pending between different clusters.

Despite the fact that K-means is undoubtedly the most widely used partitioning clustering algorithm, unfortunately, this algorithm is highly sensitive to the initial placement of the cluster centroids.Numerous initialization methods have been proposed to address this problem.

II. Related Works

M. EmreCelebi, Hassan A. Kingravi and Patricio A. Velain in theirpaper "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm" explained the different techniques and algorithms that are used in initializing the centroids placement, and highlighted where to use and where not to use each algorithm depending on the complexity, efficiency, reliability and data set size[9,10,11].

MacQueen [12] proposed two different techniques to initialize centroids. The first one takes the first K points in the data set as the centers and the second method chooses the centers randomly from the data points. Jancey's method assigns to each center a synthetic point randomly generated within the data space[13].

Forgy's method assigns each point to one of the K clusters uniformly at random. The centers are then given by the centroids of these initial clusters[14].

The Maxmin method chooses the first centroid arbitrarily and then the next centroid Ci where (i∈ {2,3 . . . ,k})ischosen to be the point having the greatestminimum distance to the previously chosen centroids[15].

The methods mentioned above have Linear Time-Complexity Initialization, however there are many other methods and algorithms that have complexity other than the linear like Loglinear Time-Complexity Initialization Methods[16,17,18], Quadratic-Complexity Initialization Methods[19,20,21,22], and others[23,24,25,26].

Our approach is work by grouping the close-points using zoom out for the points in data set and rounding these points to the nearest integer value. It is very simple method and has linear complexity with respect to n data points.

III. Problem Statement

The sensitivity of K-means clustering algorithm to the initial values of the centroids and its iterative nature in huge amount of data are the main drawbacks of this algorithm. In some cases, the randomly chosen initial centroids may not result in the expected clusters. Let's examine the data shown in fig.1. K-means clustering will not find the two clusters successfully without taking one initial centroid from group R and the other from group G. if both initial centroids are selected from the same group then K-means clustering may not give the correct clusters.

Solving the problem of the initial centroids placement in a linear complexity behavior is the goal of this paper.

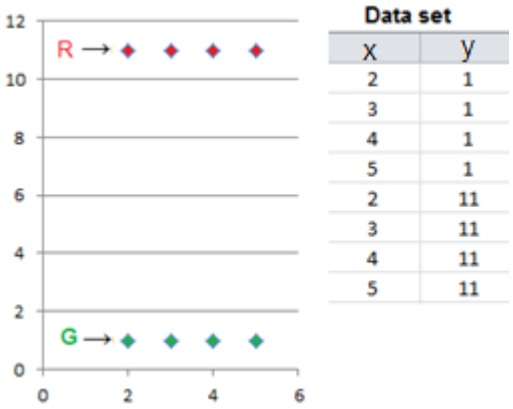


Fig. 1. Example for data set has problem with k-mean clustering with selecting intial centroid.

IV. Proposed Algorithm

The proposed algorithm is based mainly on the idea of zooming out all the data of each attribute(column). Many points can be reflected into one spot (one point) when rounded in some manner to nearest value as shown in Fig. 2.The received points can be expressed as initial centroids for the given clusters.

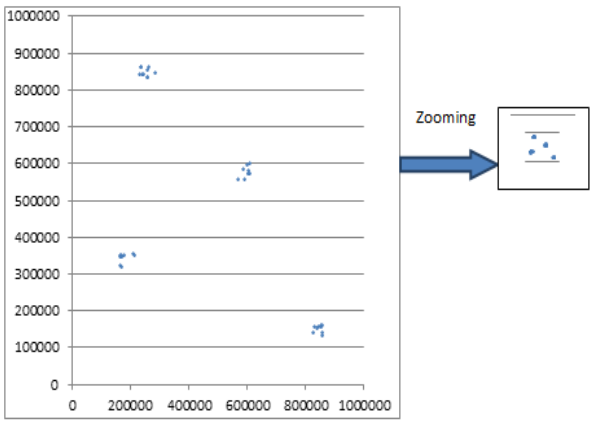


Fig. 2. Data and Cluster Zooming

A. The algorithm:

- Input: Given a data set X with L columns and k clusters
- Output: Initial centroids C_i (i=1,2,…k)
- For each column Xi of X

  Find maximum element X_{imax}

  Find column factor $Fi = k/X_{imax}$

  Normalize and round Xi into $X_i' = ROUND(X_i * F_i)$,

- The new data set is X′ as vectors with integer values leas than k
- Calculate the frequency for each value in the vector X′ and group them together
- Sort the results according to their frequencies in a descending manner(Ci' (i=1,2,…$(k+1)^L$)
- Select the first k points as candidate centroids (Ci' (i=1,2,…k)).
- do Smoothing:

  If two candidate centroids from C_i' are very close to each other with distance $d \leq \sqrt{L}$. (then this means that they may be belong to the same cluster, then the one with lower frequency is deleted and $C_{(k+k')}$ is selected as new candidate centroid. This work is repeated to exclude the candidate centroids belonging to the same cluster. The result is a new vector C_i'' (i=1,2,…k). Where k' is the number of the deleted candidate centroids.

- Scan X to calculate the mean for each of the selected centroids C_i'' to get the initial Centroids C_i (i=1,2,…k).

B. Numerical example

Given the following data set X (40 samples) with two dimensions X_1 and X_2 as Table-I.

Applying the proposed algorithm for this data set (step by step)

- We have L=2 and Let k=4.
- X_{1max}= 855233 and X_{2max}= 863523
  - $F1 = k/X_{1max} = 4/855233$ and
  - $F2 = k/X_{2max} = 4/863523$
  - Normalize and round Xi into $X_i' = ROUND(X_i * F_i)$

    $X_1' = ROUND(X_1 * F_1)$

    $X_2' = ROUND(X_2 * F_2)$

    The result is shown in Table-II:

TABLE I. DATA SET X WITH L=2 AND X1 AND X2 AS DIMENSIONS

#	X1	X2	#	X1	X2
1	587171	587115	21	241071	844424
2	588100	557588	22	258416	835432
3	604678	574577	23	241687	844256
4	602013	574722	24	282603	846165
5	603145	574795	25	229131	842806
6	601376	579831	26	236274	861302
7	565148	557305	27	261819	863523
8	599808	596484	28	256512	837094
9	605250	573272	29	236495	861569
10	606738	601356	30	258528	856273
11	169274	348574	31	850993	157873
12	166799	318482	32	828179	155649
13	161780	324523	33	850965	156224
14	168805	351913	34	839974	154358
15	206519	355629	35	854338	135067
16	163091	350167	36	855233	141357
17	176144	353300	37	850538	160159
18	213951	352868	38	826499	142732
19	164046	346109	39	854922	159650
20	169569	346955	40	840375	155757

- All the values are integers less than or equal to 4(number of clusters k).
- Data in Table-II are grouped together which results in five groups (3,3), (1,2), (1,1), (1,4) and (4,1) with frequencies 10, 9,1,10 and 10 respectively as shown in Table-III.
- Sorting the data in Table-III according to their frequencies results in Table-IV.
- The first 4 points (3,3), (1,2), (1,4), (4,1) are selected as candidate centroids.
- No smoothing is required here because all the distances between any two candidate centroids are not$\leq \sqrt{L}$.
- The means for original points of each of the candidate centroids are shown in Table-V. for example the mean value for point (3,3) is given by:

$$\overline{X1} = \frac{5963427}{10} = 596342.7$$

$$\overline{X2} = \frac{5777045}{10} = 577704.5$$

TABLE II. RESULTS OF ROUNDING Xi TO Xi'

X_1	X_2	X_1'	X_2'	X_1	X_2	X_1'	X_2'
587171	587115	3	3	241071	844424	1	4
588100	557588	3	3	258416	835432	1	4
604678	574577	3	3	241687	844256	1	4
602013	574722	3	3	282603	846165	1	4
603145	574795	3	3	229131	842806	1	4
601376	579831	3	3	236274	861302	1	4
565148	557305	3	3	261819	863523	1	4
599808	596484	3	3	256512	837094	1	4
605250	573272	3	3	236495	861569	1	4
606738	601356	3	3	258528	856273	1	4
169274	348574	1	2	850993	157873	4	1
166799	318482	1	1	828179	155649	4	1
161780	324523	1	2	850965	156224	4	1
168805	351913	1	2	839974	154358	4	1
206519	355629	1	2	854338	135067	4	1
163091	350167	1	2	855233	141357	4	1
176144	353300	1	2	850538	160159	4	1
213951	352868	1	2	826499	142732	4	1
164046	346109	1	2	854922	159650	4	1
169569	346955	1	2	840375	155757	4	1

TABLE III. DATA POINTS WITH THEIR FREQUENCIES

Group #	X'	Y'	Frequency
1	3	3	10
2	1	2	9
3	1	1	1
4	1	4	10
5	4	1	10

TABLE IV. THE SORTED GROUPS

Group #	X'	Y'	Frequency
1	3	3	10
2	1	4	10
3	4	1	10
4	1	2	9
5	1	1	1

TABLE V. THE MEAN VALUES FOR THE ORIGINAL POINTS OF THE CANDIDATE CENTROIDS

Group #	X1'	X2'	$\overline{X1}$	$\overline{X2}$
1	3	3	596342.7	577704.5
2	1	4	250253.6	849284.4
3	4	1	845201.6	151882.6
4	1	2	177019.9	347782.0
5	1	1	166799	318482

- The initial centroids are: (596342.7,577704.5), (250253.6, 849284.4), (845201.6, 151882.6),(177019.9, 347782.0),

V. TESTING RESULTS

In order to test the proposed algorithm effectiveness, four different data set groups each with 5000 2-attribute records (sample of50 records is shown in Table-VI) have been studied and tested. The results are shown in Table-VII with a comparison to the same clustering process with randomly chosen initial centroids shown in Table-VIII. The clusters representing each data set group are shown in fig. 2 – fig. 5.

TABLE VI. SAMPLE OF 50 RECORDS FROM THE COLLECTED DATA

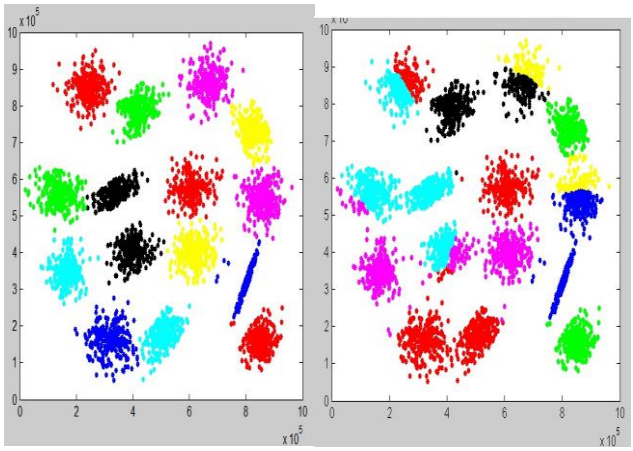
#	X1	Y1	#	X1	Y1
1	664159	550946	26	601182	582584
2	665845	557965	27	562704	570596
3	597173	575538	28	605107	563429
4	618600	551446	29	607214	575069
5	635690	608046	30	568824	570203
6	588100	557588	31	612485	518009
7	582015	546191	32	589244	573777
8	604678	574577	33	625579	551084
9	572029	518313	34	560237	500154
10	604737	574591	35	626224	569687
11	577728	587566	36	610666	551701
12	602013	574722	37	597428	569940
13	627968	574625	38	600582	599535
14	607269	536961	39	604168	555003
15	603145	574795	40	613871	550423
16	671919	571761	41	617310	551945
17	612184	570393	42	625728	579460
18	600032	575310	43	606300	566708
19	627912	593892	44	638559	558807
20	601967	604428	45	582176	630383
21	591851	569051	46	544056	577786
22	601444	572693	47	631297	578351
23	629718	558104	48	561574	621747
24	661430	603567	49	604973	574773
25	597551	556737	50	605284	556134

TABLE VII. INITIAL CENTROIDS ACCORDING TO THE PROPOSED ALGORITHM (K=15).

Cluster Centroid	Data Group 1		Data Group 2		Data Group 3		Data Group 4	
	x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
1	591556	575116	843222	648091	508163	617222	627818	721763
2	807088	326972	577566	252703	557352	311923	306749	707096
3	399685	782300	259799	732052	245043	328544	735554	437298
4	826483	724393	536460	446913	439368	390311	757563	579126
5	846435	145292	799866	249753	757074	810324	502866	520997
6	331706	568386	445714	604397	368212	578386	674674	263034
7	179205	337441	734962	478197	761651	449691	375712	379145
8	628618	393312	802371	798114	326461	764711	499748	203998
9	247998	842786	643824	717951	680810	242276	377178	515807
10	322612	179439	378722	393330	301640	460761	295740	275114
11	134301	569256	140919	250576	617110	763194	446768	644592
12	510298	183588	386668	184117	761303	637228	626856	595801
13	392194	399178	198477	468222	566976	443149	552156	390456
14	847289	533188	513836	849815	191237	202035	237942	476148
15	656682	851391	668766	143712	371714	255039	499503	770751

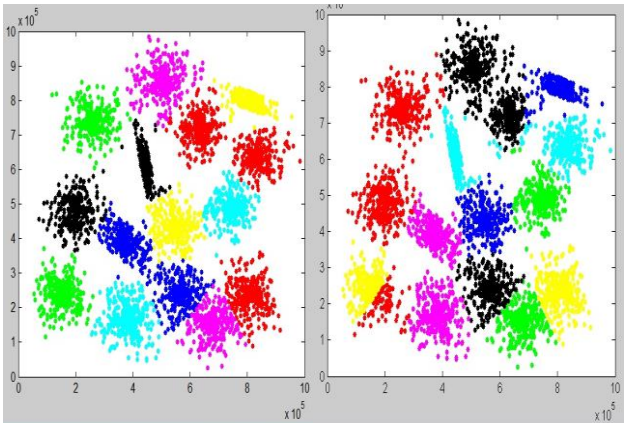
TABLE VIII. RESULTS OF APPLYING K-MEANS

Data Group	Proposed Algorithm Initialization		Random initialization	
	No of Iterations	*total sum of distances*	*No of Iterations*	*total sum of distances*
1	3	8.91762e+012	23	3.16298e+013
2	4	1.32792e+013	13	1.88525e+013
3	7	1.68906e+013	24	2.06249e+013
4	16	1.57031e+013	15	1.71045e+013



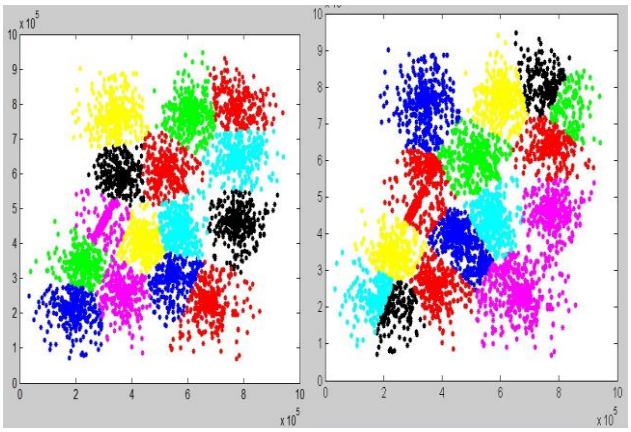
(a) (b)

Fig. 3. Cluster Representation of the DataGroups (1), a-Proposed Algorithm Clusters and b- Randomly chosen initialization



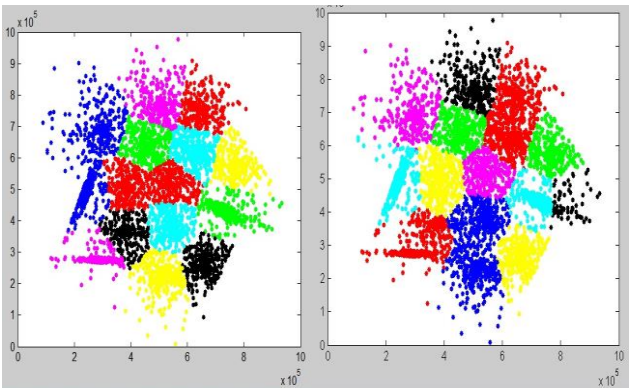
(a) (b)

Fig. 4. Cluster Representation of the DataGroups (2), a-Proposed Algorithm Clusters and b- Randomly chosen initialization



(a) (b)

Fig 5. Cluster Representation of the DataGroups (3), a-Proposed Algorithm Clusters and b- Randomly chosen initialization



(a) (b)

Fig 6 .Cluster Representation of the DataGroups (4), a-Proposed Algorithm Clusters and b- Randomly chosen initialization

VI. ALGORITHM PERFORMANCE

For a given data set with N records, K clusters and Lcolumns, the following time complexities are required. We can see that the number of groups of distinct points will be $\leq(k+1)^L$.

1. Finding maximum elementin each column for L columns is → (LN)
2. Normalize and round is → (LN)
3. Countfrequency →$(K+1)^L$N
4. Sort groups frequency→$((K+1)^L)^2$
5. Select top K groups →(K)
6. Find average for each group of all groups →(N)

Time complexity =$O(LN+(K+1)^L N + ((K+1)^L)^2)$

Since N represents the number of data points in the whole data set, and when K(number of clusters) and L (number of attributes)are constants then the time complexity =O(N)

7. Smoothing takes, simply, leas than k^2

This means that in all cases the time complexity for the proposed algorithm is linear with respect to number of data points n.

VII. DISSCUSION AND CONCLUSION

Initializing the centroids in K-means clustering represents a crucial factor in the whole clustering process since K-means technique is highly sensitive to both data set size and the initial values to start with. An efficient and effective algorithm has been introduced to solve the difficulties of initial centroids estimation. The proposed algorithm highly relies on the concept on density based, data normalization and smoothing. Tables-VII, Table-VIII and fig.2-fig.5 show the effectiveness of the proposed algorithm compared to the randomly chosencentroids and in turnwith other different initialization techniques.

The work in this paper presents a new algorithm for estimating the initial centroids in K-means clustering with a linear time complexity of almost O(n), where n represents the number of records in the data set. However, it can be effectively used in case of having variable number of clusters and attributes.

VIII. RECOMMENDATIONS

The proposed algorithm can be used effectively in for small and medium numbers of clusters and attributes since it results in reliable initial centroids with liner time complexity, however it can be used in cases with large number of clustersand attributes when reliability and convergence are crucial.

REFERENCES

[1] Jiawei Han and Micheline Kamber " Data Mining: Concepts and Techniques" 3rd Edition., Morgan Kaufmann, 2010.

[2] M. Steinbach, P.-N.Tan and V.Kumar, "Introduction to Data Mining",Addison-Wesley, 2006. ISBN: 0-321-32136-7

[3] M. H. Dunham, "Data Mining:Introductory and Advanced Topics",Prentice Hall, 2002.

[4] D. J. Hand, H. Mannila, and P. Smyth, "Principles of Data Mining", MIT Press,2001.

[5] A. K. Jain, M. N. Murty, P. J. Flynn, Data Clustering: A Review, ACM Computing Surveys 31 (3) (1999)264–323.

[6] L. Kaufman, P. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley-Interscience,1990.

[7] D. Aloise, A. Deshpande, P. Hansen, P. Popat, NP-Hardness of Euclidean Sum-of-Squares Clustering, MachineLearning 75 (2) (2009) 245–248.

[8] A. K. Jain, Data Clustering: 50 Years Beyond K-means, Pattern Recognition Letters 31 (8) (2010) 651–666.

[9] M. Emre Celebi,Hassan A. Kingravi and Patricio A. Velain in theirpaper "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm", Expert Systems with Applications, 40(1): 200–210, 2013

[10] J. M. Pena, J. A. Lozano, P. Larranaga, An Empirical Comparison of Four Initialization Methods for theK-Means Algorithm, Pattern Recognition Letters 20 (10) (1999) 1027–1040.

[11] J. He, M. Lan, C. L. Tan, S. Y. Sung, H. B. Low, Initialization of Cluster Refinement Algorithms: A Reviewand Comparative Study, in: Proc. of the 2004 IEEE Int. Joint Conf. on Neural Networks, 2004, pp. 297–302.

[12] [12] J. MacQueen, Some Methods for Classification and Analysis of Multivariate Observations, in: Proc. of the 5thBerkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[13] R. C. Jancey, Multidimensional Group Analysis, Australian Journal of Botany 14 (1) (1966) 127–130.

[14] E. Forgy, Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classification, Biometrics 21(1965) 768.

[15] T. Gonzalez, Clustering to Minimize the Maximum Intercluster Distance, Theoretical Computer Science 38 (2–3) (1985) 293–306.

[16] J. A. Hartigan, M. A. Wong, Algorithm AS 136: A K-Means Clustering Algorithm, Journal of the RoyalStatistical Society C 28 (1) (1979) 100–108.

[17] M. Al-Daoud, A New Algorithm for Cluster Initialization, in: Proc. of the 2nd World Enformatika Conf., 2005,pp. 74–76.

[18] S. J. Redmond, C. Heneghan, A Method for Initialising the K-means Clustering Algorithm Using kd-trees,Pattern Recognition Letters 28 (8) (2007) 965–973.

[19] F. Cao, J. Liang, G. Jiang, An Initialization Method for the K-Means Algorithm Using Neighborhood Model,Computers and Mathematics with Applications 58 (3) (2009) 474–483.

[20] G. N. Lance, W. T. Williams, A General Theory of Classificatory Sorting Strategies - II. Clustering Systems,The Computer Journal 10 (3) (1967) 271–277.

[21] M. M. Astrahan, Speech Analysis by Clustering, or the Hyperphoneme Method, Tech. Rep. AIM-124, StanfordUniversity (1970).

[22] G. W. Milligan, An Examination of the Effect of Six Types of Error Perturbation on Fifteen ClusteringAlgorithms, Psychometrika 45 (3) (1980) 325–342.

[23] G. P. Babu, M. N. Murty, A Near-Optimal Initial Seed Value Selection in K-Means Algorithm Using a GeneticAlgorithm, Pattern Recognition Letters 14 (10) (1993) 763–769.

[24] A. Likas, N. Vlassis, J. Verbeek, The Global K-Means Clustering Algorithm, Pattern Recognition 36 (2) (2003)451–461.

[25] D. Aloise, P. Hansen, L. Liberti, An Improved Column Generation Algorithm for Minimum Sum-of-SquaresClustering, Mathematical Programming (2010) 1–26.

[26] [26] G. Babu, M. Murty, Simulated Annealing for Selecting Optimal Initial Seeds in the K-Means Algorithm, IndianJournal of Pure and Applied Mathematics 25 (1–2) (1994) 85–94.

الملخص

تمثل العنقدة وحدة من أكثر خوارزميات استخلاص المعرفة في تقنيات التنقيب عن البيانات. النمط الهرمي ونمط التجزئة والتقسيم هي الأكثر شيوعا في هذا المجال. ولكل منهما ايجابياته وسلبياته وأهدافه. تمثل خوارزمية k-meanالأكثر شيوعا في تقنيات التجزئة والتقسيم. ولكنها تعاني من نقطتي ضعف هما تعقيد الوقت وحساسية الخوارزمية تجاه المراكز الابتدائية. يقدم العمل في هذا البحث منهجية مقترحة لتحديد المراكز الابتدائية للعناقيد (k) من خلال ثلاث عمليات هي تحديد الكثافة (density based) التطبيع (normalization)والتنعيم (smoothing). كما ان الخوارزمية المقترحة ذات أساس رياضياتي رصين.
تم اختبار المنهجية المقترحة من خلال استخدام البيانات المتوفرة على الانترنيت وبواقع (20000 قيد أو سجل). وتشير النتائج المستخلصة من البحث إلىأن المنهجية المقترحة ذات تعقيد وقت أفضل وتضمن تقارب الخوارزمية للوصول إلى الحل.

الكلمات المفتاحية:

التنقيب عن البيانات، العنقدة، k-means، المراكز، التعقيد