Journal of Education for Pure Science- University of Thi-Qar
Vol.10, No.1 (March., 2020)
Website: jceps.utq.edu.iq Email: jceps@eps.utq.edu.iq

DOI: http://doi.org/10.32792/utq.jceps.10.01.04

A Proposed Algorithm for Clustering Data Based On Relations Among Points

Mustafa A. Naser[1], Firas S. Miften[2]

[1] *Ministry of Education, Thi-Qar Education Directorate*
[2] *University of Thi-Qar, College of Education for Pure Science, Computer Science Department*

Received 3/6/ 2019 Accepted 17/7/2019 Published 20/1/2020

Abstract:

This paper deals with automatic clustering algorithm, which partitions a many of objects to get smaller sets (clusters). Where objects of the one set are related to each other rather than to those in the other sets. The number of clusters in automatic clustering don't given priori, they determinates automatically, the number resulted of clusters exact closed to the numbers of the real dataset structure.

This paper represents a stimulus for the current study to introduce an algorithm that automatically finds the number of clusters based on distance among vertices. The study is based on the hypothesis that the proposed algorithm is able to efficiently find the clustering partitions for the whole dataset.

Keywords: Clustering, K-means Algorithm, Hierarchical Clustering Algorithm, Proposed Clustering Algorithm, CURE Clustering Algorithm, Partitioning Clustering Algorithm.

1. Introducti:

Data Mining is known as the process of analyzing data to extract interesting patterns and knowledge. Data mining is used for analysis purpose to analyze different type of data by using available data mining tools [3].

This information is currently used for wide range of applications like customer retention, education system, production control, healthcare, market basket analysis, manufacturing engineering, scientific discovery and decision making etc [1]. Data mining is studied for different databases like object-relational databases, relational database, data ware houses and multimedia databases etc.

Clustering: is partitioning data to a number of clusters (groups, subsets, or categories). Description most researchers the cluster by the internal homogeneity and external separation such that the content of cluster must be similar to each other's and the opposite is true. There are taxonomies different of algorithms that works on clustering data such as (partition and hierarchical clustering, Neural Network-Based clustering and Kernel Based clustering) these algorithms may be automatic or non-automatic [22,23,24].

Algorithms of clustering usually has been designed with some types of biases and assumptions. By this sense, don't can be say best and accurate of clustering algorithms context, although possible some comparisons, most these comparisons rely on around specific applications, by specific conditions, the results may become quite changes if occur different in the conditions [2,24].

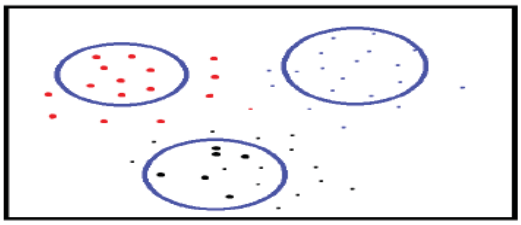

Figure (1): Clustering [2]

Clustering algorithms can be divided into:

- **Automatic technique**: In this type of group, the number of groups is not given in advance, and is automatically determined by the group algorithms used. The resultant number of groups is often closed accurately to determine the number of groups involved in the actual structure of the data set [17].

Non-Automatic techniques: In this type of group, a number of groups must be given in advance by the programmer. K-mean algorithm is a sample of this type. The accuracy of the results obtained depends on the expected number of groups selected by the user when implementing this algorithm on a real data set [17].

1.1 Types of Clustering

There are various types of clustering in data mining.

1.1.1 Partitioning Clustering

The general criterion of division is a combination of high similarity of samples within clusters with high variance between distinct groups. Most partitioning methods depend on distance. These aggregation methods work well to find spherical clusters in small to medium databases [5].

One of partition clustering algorithms is ***k*-MEANS CLUSTWRING**

Let $X = \{\mathbf{x1}, \mathbf{x2}, \ldots, \mathbf{xN}\}$ N p-dimensional data points to be clustered into K clusters. Let clusters set as $U \sim \{X1, X2, \ldots, Xk\}$. K-mean searches for a partition that has the minimum total square error between the sample of the clusters and the points in the clusters. The mean value **mi** of cluster *Xi is* given by:

$$\mathbf{m_i} = \frac{\sum_{\mathbf{x_j} \in X_i} \mathbf{x_j}}{|X_i|}$$

The Euclidean norm is the usual choice as follow equation:

$$\text{Dist (p,q)} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \qquad \ldots\ldots (2)$$

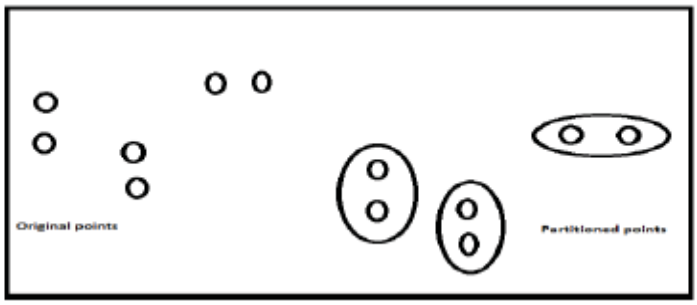

Figure (2): Partitioning Clustering [2]

1.1.2 Density Based Clustering:

In most partitioning methods depends on the distance between objects. In density method the group of objects is continue to grow while the density in the neighborhood overtake some threshold [6].

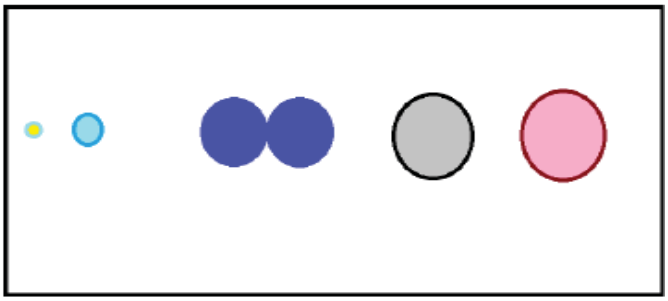

Figure (3): Density based Cluster [2]

1.1.3 Hierarchical Methods:

In hierarchical method is decompose of the given set of data objects is created. It can be classified as being either **agglomerative** or **divisive** based on how hierarchical decomposition is formed. Agglomerative approach is the bottom up approach starts with each object forming a separate group such as CURE algorithm. Hierarchical algorithms create hierarchical analysis of the data set for data objects. Hierarchical decomposition is represented by a tree structure called a dendrogram. Does not need groups as input. In this type of assembly, it is possible to view sections at different levels of detail using different types of K. For example, flat grouping.

It then merges the groups close together until all the groups are merged into one. The split method is the top-down approach, where all groups start in the same group, and in each iterative step, the cluster is divided into smaller groups until each object becomes a single block or cluster [8].

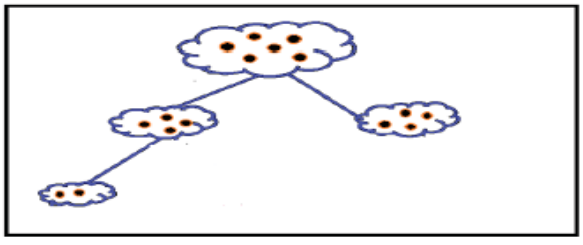

Figure (4): Hierarchal Clustering [2]

2. Data Analytics:

Data Analytics [15] is the science of examining raw data with the purpose of extract useful information to draw conclusions. Data Analytics can be used by many industries and organizations to get better business decision. Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known by the researcher. There are two types of data analytics:

a- Classification: Predictability models predict class labels; predictive models predict jobs of continuous value [2].

b- Prediction: Predictability model for predicting dollar expenditures for potential customers on computers given their income and occupation.

Analytics forecasts bring management, skills, information technology and modeling. Predictive analyzes are data science, multidisciplinary skills and a core set of non-profit organizations, business success and government. Marketing sales forecast or market share, a good retail site opportunity was found. It also identifies consumer segments, target marketing and risks associated with current products, and provides key predictive analytics for it [2].

Classification is a method of data extraction that comes within the technique of automatic learning to predict group membership of data representations. Classification technology is able to process a variety of data over gradient and its popularity increases.

3. A Proposed Algorithm:

The proposed algorithm works on clustering data based on distance(dissimilarity) matrix and adaptive threshold(Th). The distance matrix(D) calculates distance between data objects

where D_{ij} represents dissimilarity (distance) between object i and object j

Size of this matrix D is (n*n), where **n** is a number of objects in dataset that every object represented by vector of features.

The proposed algorithm takes dataset and computes the matrix (D) and calculate threshold for every iteration, as shown in algorithm below at step 2 then calculate size of D at step 4 and entry to(for) looping

For overall rows in matrix D and check first row if not have label if yes that assign label for this row as Noc, as shown in step 8 of algorithm and next step calculate threshold (Th) for every iteration based on formula specified in step 10 and check contents rows ith to others content by depends on adaptive threshold(Th) that calculated for this row and assign any value that less than or equal to (Th) to the same class or label for this row or object, as steps from (11 _ 17)

This algorithm considers one of the automatic technique because it finding the groups in dataset without determined by user, which good feature for its works.

The below figure (5) shown steps flow a work.

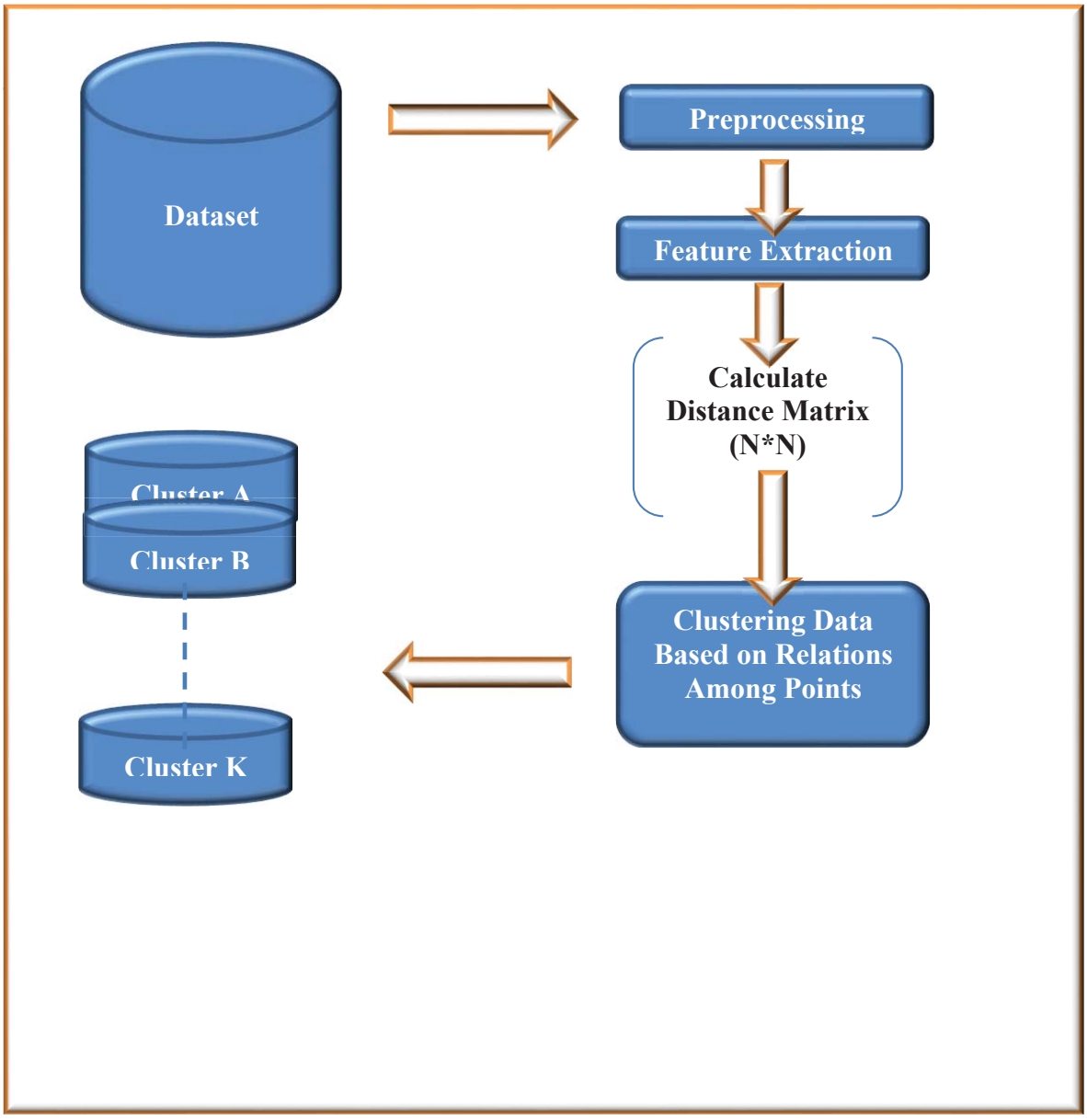

Figure (5): Block diagram for the proposed algorithm

A Proposed Algorithm:

Input:

X = {**x1**; **x2**; …… ; **xn**} // *where X is dataset with p*-dimensional data points.

Where:

n - is a number of data points in dataset where every row represent object

Output:

Label. \\ contains the label for every object

Noc. \\ the number of clusters that found in dataset X

1) **Begin**
2) Compute (n, n) : D is distance matrix between data points in X by using the Euclidean distance.
3) Set Label = zeros(n)
4) Noc = 0; //counter for the number clusters, initial value zero

5) $for\ i = 1\ to\ n$
6) $if\ \ Label\ (i) ==\ 0\ then$
7) $Noc\ =\ Noc\ + 1;$
8) $Label\ (i) =\ Noc;$
9) $End\ if$
10) $Th = (min(D(i,:)) + max(D(i,:)))/2 * 0.16;$ // calculate adaptive threshold.
11) $for\ j = i + 1{:}m$
12) $if\ D(i,j) <= Th\ \ then$
13) $if\ Label(j) == 0$
14) $Label(j) = Label(i);$
15) $End\ if$
16) $End\ if$
17) $End\ for\ loop\ j$
18) $End\ for\ loop\ I$
19) $\boldsymbol{End\ algorithm}$

4.Results

In this paper proposed algorithm is applied on dataset named **Aggregation[15]** loaded from this link http://cs.joensuu.fi/sipu/datasets/ which consist of 788 data points by 2- dimensional and **7** classes and it shows good result for clustering this dataset.

Powerful this algorithm that could returns numbers of clusters in dataset(Automatic), in contrast with K-Means algorithm.

A proposed algorithm considers Automatic clustering technique, while K-Means algorithm Non-Automatic clustering technique [17]. In this section we presenting the experiment results By apply three algorithms the proposed, K-Means and CURE algorithm on same dataset then compared between them based on this factors (data size, accuracy, recall, execute time)

Where accuracy measure is given as the ratio of the number of samples correctly labeled to the total number of samples in the dataset, table (1) explains that.

Accuracy = # (Correctly labeled samples) / # (Total samples).

and recall measure computed as:

Recall = 1 – accuracy

Table (1) shows results comparison between three algorithms

Algorithm	Data size	NO.Clusters	Accuracy	Recall	Execute time
Proposed	312	3	0.9968	0.0032	0.020023
K-Means	312	3	0.3365	0.6635	0.023577
CURE	312	3	0.5801	0.4199	13.322092

Note: This table (1) we applied algorithms on data set called **Spiral [16] (**n=312, k=3, p=2).

Below figures shows a visualize results for the three algorithms that applied to another data set called **Aggregation** which mentioned previously:

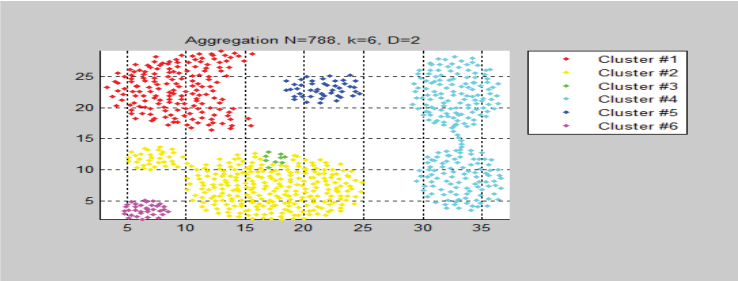

Figure (6) "Apply proposed algorithm on Aggregation"

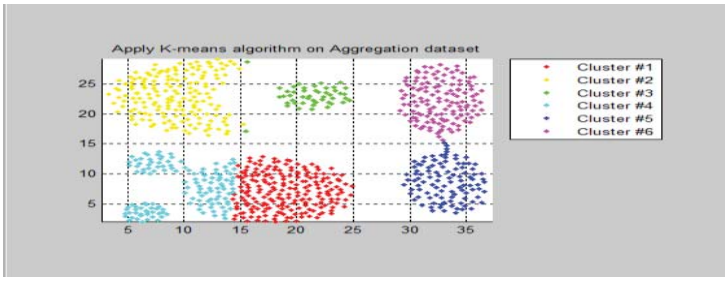

Figure (8) "Apply CURE algorithm on Aggregation"

5. Discussions and Conclusions

1.The proposed algorithm the clusters number is automatically determined through this algorithm, in contrast, for the K-means and CURE algorithms, the clusters number is a priori given by the programmer. As a result, the performance of others algorithm and K-means less than proposed algorithm because the second one is product clusters nearest from structure of dataset.

2.The value of threshold selected results of a proposed algorithm depending on it.in contrast, the CURE and K-means algorithm depend on initial value for clusters(K). large chosen value for threshold our proposed lead to one cluster contains different objects, on the other hand, if pick small value lead to data points that belong to one subject in two groups, or more. On the other hand, for the K-means and CURE algorithms, select K if small will lead to combine the data points of more subjects in one cluster, whereas splitting data points of specific subject into many groups

3.Table (1) shows results for K-means and CURE algorithms is less than a proposed algorithm.

4. The criterion of dataset size, a proposed approach is outperforming CURE and K-means algorithms when the data were very large.

6. References:

[1] K. Rajalakshmi, Dr.S.S. Dhenakaran, N. Roobin "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July 2015

[2] Shalini Goel. "Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining". International Journal of Computer Applications · January 2017.

[3] Shital A. Raut and S. R. Sathe, "A Modified Fastmap K-Means Clustering Algorithm for Large Scale Gene Expression Datasets", *International Journal of Bioscience, Biochemistry and Bioinformatics*, Vol. 1, No. 4, page 120-124, November 2011.

[4] Daljit Kaur and Kiran Jyot, "Enhancement in the Performance of K-means Algorithm", International Journal of Computer Science and Communication Engineering, Volume 2 Issue 1, 2013.

[5] Siddheswar Ray and Rose H. Turi, "Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation", School of Computer Science and Software Engineering Monash University, Wellington Road, Clayton, Victoria, 3168, Australia, 1999

[6] Azhar Rauf, Mahfooz, Shah Khusro and Huma Javed "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", *Middle- East Journal of Scientific Research* 12 (7): 959-963, 2012 ISSN 1990-92332012

[7] Madhu Yedla, T M Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center*", International Journal of Computer Science and Information Technologies*, Vol. 1 (2) 2010, page 121-125

[8] Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S "Reducing the Time Requirement of K-Means Algorithm" *PLoS ONE*, Volume 7, Issue 12, pp-56-62, 2012.

[9] Akhilesh Kumar Yadav, Divya Tomar, Sonali Agarwal, "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT) 2013

[10] Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey "Weather Forecasting using Incremental K-means Clustering", 2014

[11] Chew Li Sa; Bt Abang Ibrahim, D.H.; Dahliana Hossain, E.; bin Hossin, M., "Student performance analysis system (SPAS)," in *Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on* , vol., no., pp.1-6, 17-18 Nov. 2014

[12] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, 2010

[13] Qasem a. Al-Radaideh, Adel Abu Assaf 3eman Alnagi, "Predictiong Stock Prices Using Data Mining Techniques", The International Arab Conference on Information Technology (ACIT'2013)

[14] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Vol IWCE 2009, July 1 - 3, 2009, London, U.K

[15]. A. Gionis, H. Mannila, and P. Tsaparas, Clustering aggregation. ***ACM Transactions on Knowledge Discovery from Data (TKDD)***, 2007. 1(1): p. 1-30.

[16]. H. Chang and D.Y. Yeung, Robust path-based spectral clustering. ***Pattern Recognition***, 2008. 41(1): p. 191-203

[17]. Jassim T. Sarsoh, Kadhem M. Hashim, Firas S. Miften, "Comparisons between Automatic and Non-Automatic Clustering Algorithms", Journal of College of Education for Pure Sciences Vol. 4 No.1.

[18] Anton Riabov, Zhen Liu, Joel L. Wolf, Philip S. Yu and Li Zhang. "Clustering Algorithms for Content-Based Publication-Subscription Systems", in Proceedings of the 22 nd International Conference on Distributed Computing Systems (ICDCS'02) ©, IEEE Computer Society Washington, DC, USA, Page 133, 2002.

[19] Cen Li and Gautam Biswas. "Unsupervised Learning with Mixed Numeric and Nominal

Data". IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 4, Pages: 673 – 690, July/August 2002.

[20] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani and Sharad Mehrotra. "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases". Knowledge and Information Systems, Vol. 3, Pages 263–286, 2001

[21] Lepere R. and Trystram D. "A New Clustering Algorithm for Large Communication Delays" in Proceedings of16th IEEE-ACM Annual International Parallel and Distributed Processing Symposium (IPDPS'02), Fort Lauderdale, USA, 2002.

[22] Rui Xu and Donald C. Wunch,"**Survey of Neural Network**", Vol.16, No.3, 2005.

[23] Jain A., and Dubes R., "**Algorithms for Clustering Data**", Neural Computer Survey., Vol.37, No.12, 1989.

[24] Everitt B., Landau S., and Leece M., "**Cluster Analysis**", London, Arnold, 2001.