Enhancing the Efficiency and Scalability of Big Data Using HK-Hybrid Clustering Algorithm

Dr. Meenu Dave[1], Hemant Kumar Gianey[2]
Computer Science Department
Jagan Nath University, Jaipur, Rajsthan, India.

Abstract- Clustering is a technique which is very important and popular technique used for Big Data minning. Clustering is basically a part of unsupervised leaning. We can also say that clustering is a process of organising objects into groups whose members are similar in some way. It does not only organize, but also identifies structure in the given set of unlabeled data. It is a technique to group number of systems in such a way to work together like a single system. K-means is simple and an efficient method used in data clustering technique. Hierachical technique is also important and useful in data clustering. In this paper we present an efficient HK-hybrid data clustering algorithm whereby we combine the properties of both k-means and hierarchical clustering together.

Keywords-Big Data, K-means clustering, Hierachical clustering.

I. INTRODUCTION

Data clustering techniques are an important and useful aspect which is used in various fields such as data mining [1], pattern recognition and pattern classification [2], data compression [3], machine learning [4], image analysis [5], and bioinformatics [6]. Clustering is basically a part of unsupervised leaning. It deals with finding structure in a collection of unlabeled data [7].We can also say that clustering is a process of organising objects into groups whose members are similar in some way. It is a technique to group number of systems in such a way to work together like a single system.

Clustering problems basically have four types of components [8]:

- Data set physical representation;
- Similarity between data points;
- The criterion function to optimize clustering solutions;
- The procedure of optimization.

II. CLUSTERING- TYPES OF DATA

A. Text or Document Data:

Text data are mostly words used in documents that are used to form phrases, sentences, paragraphs headings, names, and other forms of communication. Text data [9] can contain letters, numbers, and special characters such as *!, &, etc*. Some of the software/tools available for text clustering are Lingo3G, Vivisimo and Lemur. For using the text data, the following steps are essential.

- Organising a Document:

The documents [9] are hierarchically organized into coherent categories because this can be very useful for systematic browsing of the document collection.

- Classifying a Document:

In the [9] application of supervised learning, to improve the accuracy of the classification of a document, methods of co-training and word clusters are used. This is also very useful to improve the quality of results.

B. Numerical Data:

A Numerical Data is the type of data which involves digits only, either in integer or real format. Numerical data is used in statistics and quantitative research methodology. The result of numerical data is generally represented in terms of equations and sometimes it is also represented in graphs, tables and charts.

Areas that make extensive use of numerical data are in the field of sensing and monitoring - such as in mineral exploration, environmental sensing over large areas or multiple sensors, financial data- such as financial service institutions that integrate many financial sources, or in electronic commerce and web 2.0 applications where the focus is on user data, etc. [10].

C. Image Data:

We can say that image clustering is a high-level description of image content. The basic goal of image clustering is to find out a mapping of the archive images into clusters. We classify the images as they provide same information about the image. This classification provides a concise summarization and visualization of the image content. Image clustering is also useful for image database management system and for the creation of a user-friendly interface to the database. Images of the CT scan of brain, is an-example of image clustering. For image clustering, the following issues need to be properly addressed:

1. Identification of ways to represent the different features of the given image.
2. Methodology to organize the identified features
3. Classification of the image, i.e. assigning the image to a particular cluster.

D. Categorical Data:

Categorical data is the type of data which includes examples like sex, educational level, age group, race etc. In this example the term educational level and age group is involving the highest grade completed and the exact values for age. This can easily describe the type of categorical Data.

Categorical data is generally analysed by the use of data tables.

A two-way table presents categorical data by counting the number of observations that fall into each group for two variables, one divided into rows and the other divided into columns.

The structure of categorical data is different from the continuous data. Categorical data does not support the distance functions which are supported by continuous data. Categorical data also does not directly support or applly the algorithms which are directly applied on continuous data. For example cluster of intra-attribute values.

E. Binary Data:

Data analysis domain has an special place for binary data. Binary data clustering is used for the application market basket data clustering and document clustering. A binary vector [11] is used where market basket data is required. The elements of the vector indicate the purchase status as true or false. A binary vector is used to represent each document for document clustering, where each element indicates whether a given word/term was present or not. Binary data clustering is commonly used in market research, astronomy, sociology, medicine, World Wide Web, geography, and archaeology.

F. Ordinal Data:

Ordinal Data is also a synonym used for the scale data. Order of magnitude or relative importance of data is presented by ordinal data. For example, a scale of small, medium and large amount of data or any range, but not their absolute values. We can also say that a data type which can be put into an order or we can give a rank or rating according to their priority or prominence than can also be known as ordinal data. You can count and order the ordinal data, but you cannot measure it. A popular fundamental marketing activity is grouping of similar customers and products. Market segmentation is done using ordinal data. Companies have to divide markets into groups of consumers to connect with all their customers with similar needs and wants which are known as segments. Each segment contains a unique set of positions to target similar type of firms. For example Ferrari in the high- end sports car in the market. All market segments are based on industrial practise, practical grounds and wisdom. These segments are formed to be based on data sets that are less dependent on subjectivity. For example ordinal data clustering allows accurate prediction of class labels of future.

Types of Clustering:

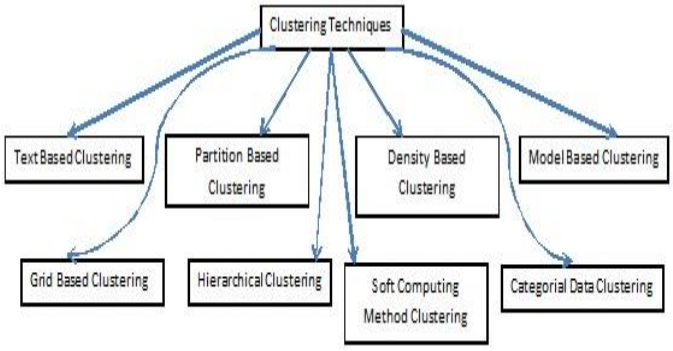


Figure 2.1 Types of clustering

Partition and Model based clustering is further divided in two parts:

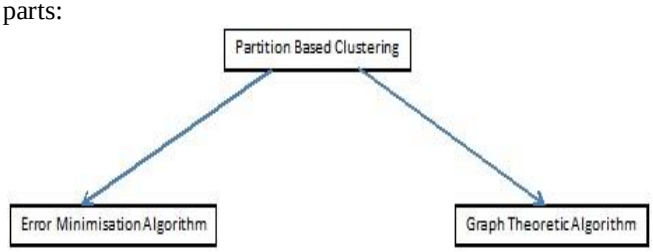


Figure 2.2 Partition based clustering

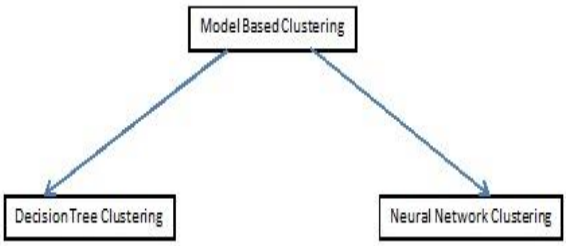


Figure 2.3 Model based clustering

Hierarchical and Soft computing clustering is also further divided in sub parts:

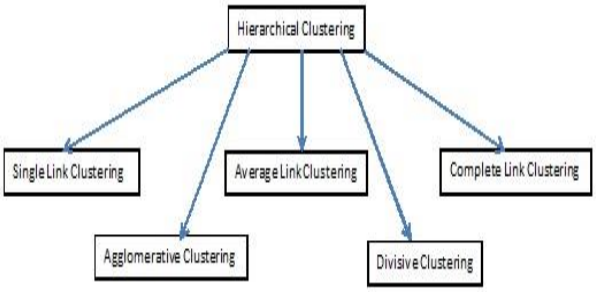


Figure 2.4 Hierarchical clustering

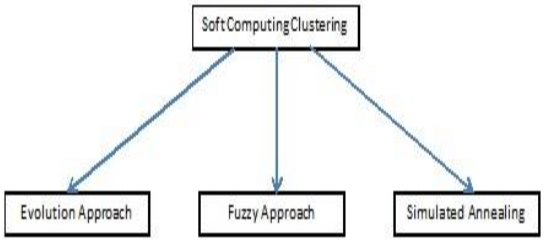


Figure 2.5 Soft computing clustering

III. GAP ANALYSIS

In today's era, the corporate and scientific environment produces massive amounts of data. To collect and analyze this data is a difficult task as data is increasing not in amount only but in complexity. Based on literature survey, there are various techniques which are used to analyze large datasets but these techniques are not efficient as some of them are related to the particular task and do not provide the global solutions, some of them are fast but they had to compromise with the quality of clusters and vice versa[12].

For clustering datasets K-means algorithm [13] is one of the most well-known unsupervised learning algorithm. The K-means clustering is the most widely used [14] due to its simplicity and efficiency in various fields. It is also considered as the top ten algorithms in data mining [15].

There has been a lot of work done to improve the efficiency of K-Means and Hierarchical algorithms to determine good quality clusters in less computation time but there are some shortcomings in both these techniques which are discussed as follows and there is a need to design new methodologies which can deal with the real time and online streaming data.

- Hierarchical algorithm can never undo what was done previously.
- The time complexity of at least $O(n^2 \log n)$ is required, where '*n*' is the number of data points.
- Based on the type of distance matrix chosen for merging different Hierarchical algorithms can suffer with one or more of the following:

i) Sensitivity to noise and outliers
ii) Breaking large clusters
iii) Arbitrary sized clusters and clusters with convex shapes are not easy to handle

- No objective function is directly minimized in Hierarchical algorithm
- In K-means algorithm it is difficult to predict K-Value.
- K-means does not work well with global clusters.
- With the help of dendograms it is difficult to identify the correct number of clusters.

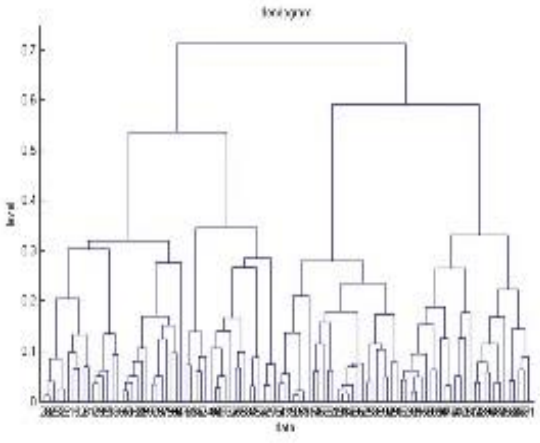


Fig. 3.1 : Showing dendogram formed from the data set of size *'N'* = 60

- Different initial partitions can result in different final clusters for k-means algorithm
- K-means algorithm does not work well with clusters (in the original data) of different size and different density
- The parallel k-means algorithm executes fast but it cannot handle non-arbitrary shape and also does not deal with the noisy data [16].
- The ELM feature is applied to K-Means to find accurate clusters in less execution time. The clusters produced are compromised in quality as large amount of resources are required to get an optimal solution [17].
- In another technique, k-means is modified to K-mode and K-prototype which generates a better result than k-means but these techniques are not able to hadle outliers [18].
- K-Means is also applied with modified coherent intelligence which provides efficiency and reliability but in this boundary points is the problem [19].

IV. PROBLEM STATEMENT

The current research frontiers are more focussed towards big data. Among various challenges in analyzing big data, the major issue is to design and develop the new techniques for clustering. Clustering techniques are used for analyzing big data in which cluster of similar objects are formed that is helpful for the business world, weather forecasting etc. Cloud computing can be used for big data analysis but there is a problem to analyze data on cloud environment as many traditional algorithms cannot be applied directly to cloud environment and also there is an issue of applying scalability on traditional algorithms, delay in the results produced and also the accuracy of the results produced. These issues can be addressed by application of K-Means and Hierarchical applied together. Therefore in this research work, a clustering algorithm is proposed and designed that helps in analyzing data in an efficient manner.

A. Proposed HK-Hybrid Algorithm

The design of proposed algorithms is as follows:
Input: D: Dataset having p data points.
K: Minimum number of clusters to be formed.
Output: K clusters having Cm center mean as cluster id.
Let $X = \{x_1, x_2, x_3, ..., x_n\}$ be the set of data points.

1) Calculate the optimal value of k with ELB method
2) Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.
3) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.
4) Assign the cluster pair to same cluster
5) Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = mean (d[(k),(r)], d[(k),(s)]).
6) If all the data points are assigned to a cluster and total unique clusters is greater than optimal value of k then stop, else repeat from step 2).

B. Proposed HK-Hybrid Algorithm: Execution Stages

The detailed explanation of how the proposed algorithm works is given here.

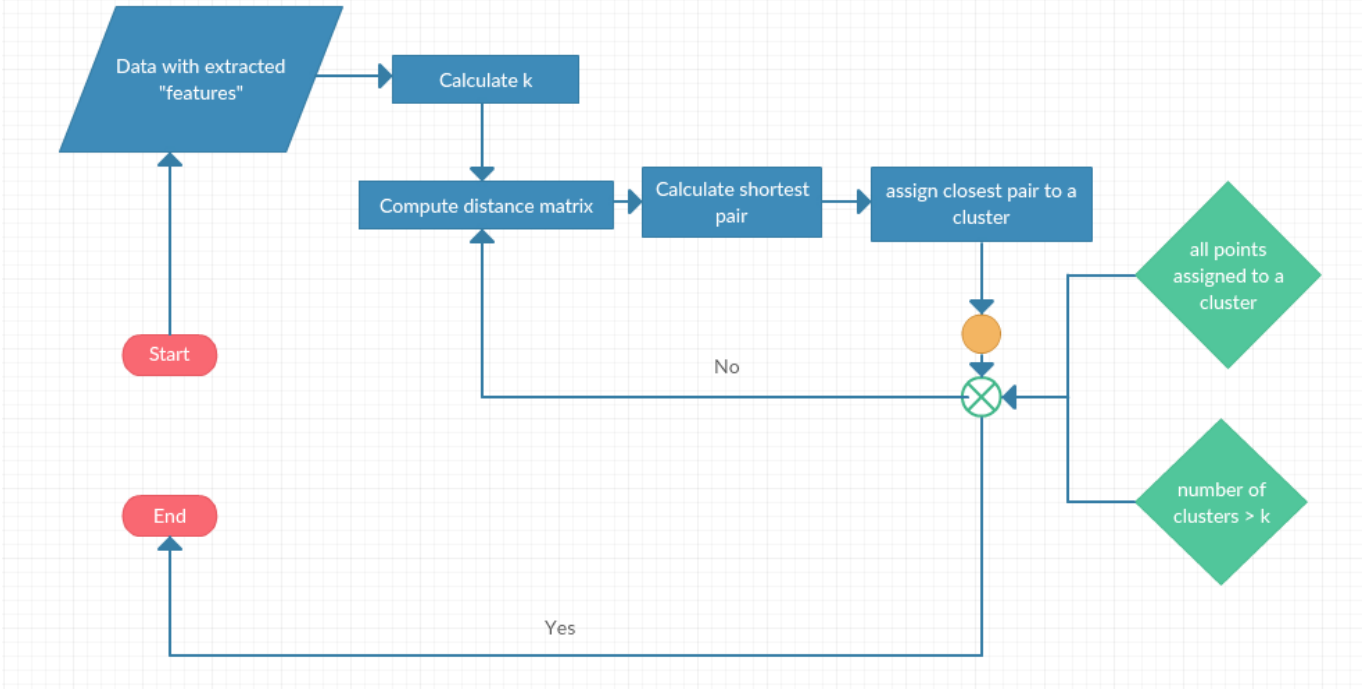


Fig. 3.2 : Proposed hybrid algorithm execution flowchart

Stage 1: Calculation of k

Compute the sum of squared error (SSE) for some values of k (for example 2, 4, 6, 8, upto 15.). The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid.

Mathematically:

$SSE=\sum Ki=1\sum x\in cidist(x,ci)2$

If you plot k against the SSE, you will see that *the error decreases as k gets larger*; this is because when the number of clusters increases, they should be smaller, so distortion is also smaller. The idea of the elbow method is to choose the k at which the SSE decreases abruptly. This produces an "elbow effect" in the graph, as you can see in the following picture:

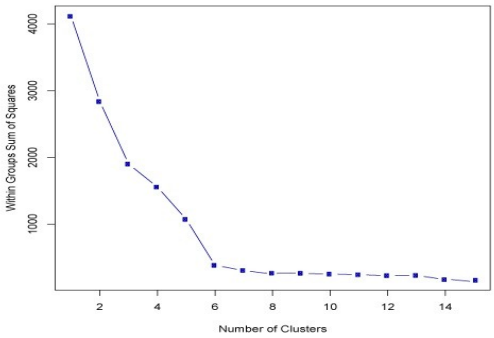


Fig. 3.3 : Elbow method for calculation of k

In this case, k=6 is the value that the Elbow method has selected.

Stage 2: Cluster Vector

For this stage a cluster vector is declared of the size of the total data points and each point is assigned a default cluster number.

Stage 3: Distance Matrix

Distance matrix is a square matrix (two-dimensional array) containing the distances, taken pairwise, between the elements of a set.

The minimum of this distance matrix gives the closest pair of data points from the data set. The centroid of these dataset is then calculated and the in the original dataset the corresponding rows for the calculated closest pair is replaced by the mean of these datapoints.

Stage 4:

After stage 3 we are left with a new different dataset than the original with mean value for the closest pair of data points calculated. If after stage 2 and 3 all data points have been assigned to a cluster and the total number of unique clusters is found greater than the value of k from Stage 1, we stop the execution otherwise we repeat Stage 2 and 3 until the desired result is achieved.

V. RESULT ANALYSIS

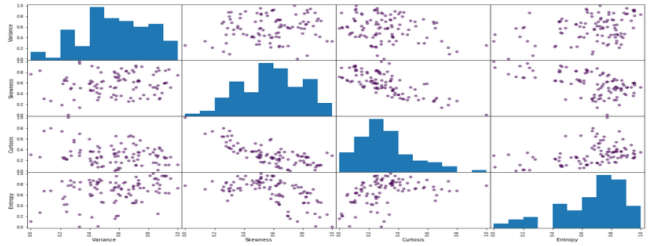
To implement the proposed approach five real time datasets are used to validate our proposed HK-hybrid algorithm. These datasets are collected from UCI machine learning repository. Tests for k-means in Python programming language, used the well known UCI Machine Learning Repository [20]. The UCI Machine Learning Repository [20] is among other things, a collection of databases, which is widely used by the research community of Machine Learning, especially for the empirical algorithms analysis of this discipline. All these datasets have different dimensionality. The proposed approach is implemented by python on platform under the environment of 2.3 GHz Intel(R) Core i3 VI generation processor, 12 GB RAM and Windows 10 operating system. Different clustering results for (K-means, Means shift, Agglomerative ,DBSCAN and HK-Hybrid) are shown in Plot 1 to 5. Different Parameters of K-means, DBSCAN, Meanshift, agglomerative and HK-hybrid clustering algorithms for banknote authentication data set are shown in table 1.

	Variance	Skewness	Curtosis	Entropy	Class
458	4.384800	-3.072900	3.042300	1.274100	0
1223	1.340300	4.132300	-4.701800	-2.598700	1
1306	-1.224400	1.748500	-1.480100	-1.418100	1
1337	0.234600	-4.515200	2.119500	1.444800	1
1171	-3.855200	3.521900	-0.384150	-3.860800	1
61	0.496650	5.527000	1.778500	-0.471560	0
649	-0.383880	-1.047100	8.051400	0.495670	0
1323	-0.025314	-0.173830	-0.113390	1.219800	1
722	4.845100	8.111600	-2.951200	-1.472400	0
1001	-0.036127	1.525000	-1.408900	-0.761210	1
600	1.155800	6.400300	1.550600	0.696100	0
122	-1.457200	9.121400	1.742500	-5.124100	0
163	2.400800	9.359300	-3.356500	-3.352600	0
599	3.929200	-2.915600	2.212900	0.308170	0
802	0.600500	1.932700	-3.288800	-0.324150	1
1048	-0.847100	3.132900	-3.011200	-2.938800	1
483	0.967880	7.190700	1.279800	-2.456500	0

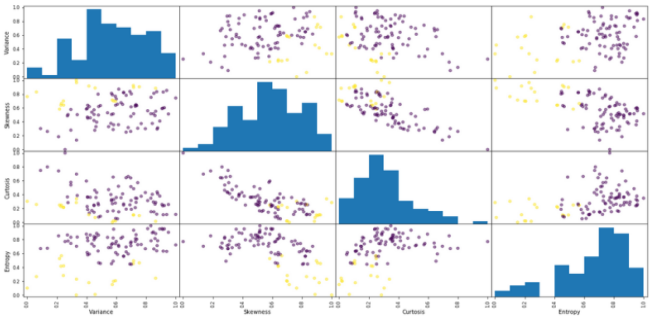
Fig. 3.4 : Dataset 1: banknoteAuthentication

Plot

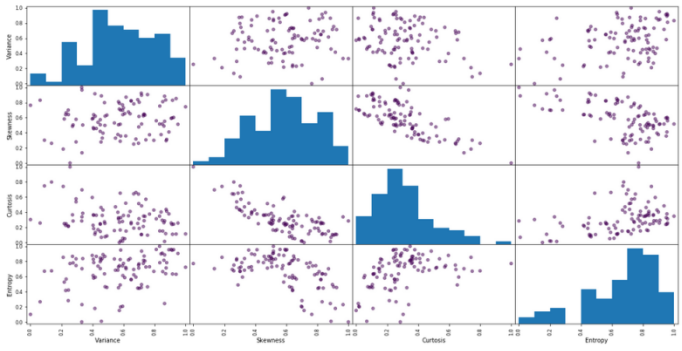
1. K-means



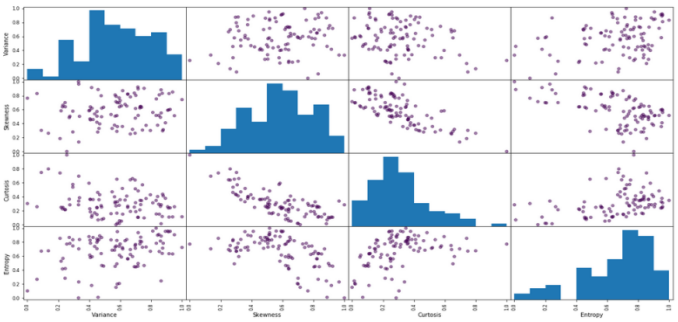
2. Mean Shift



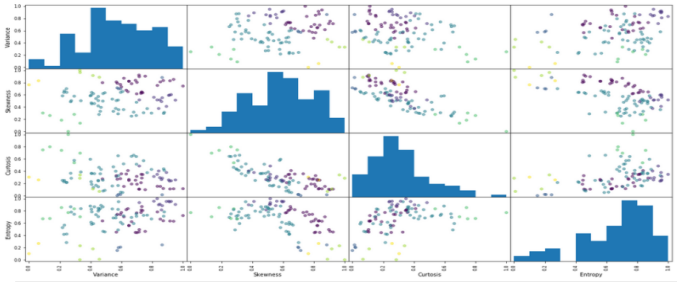
3. Agglomerative



4. DBSCAN



5. HK-Hybrid



Algorithm	Adjusted Rand Index	Time	Completeness	Homogeneity	V Score
Kmeans	0.220	0.035	0.348	1.000	0.517
MeanShift	0.087	0.043	0.256	0.236	0.246
Agglomerative	0.113	0.001	0.119	0.119	0.119
DBSCAN	0.021	0.002	0.230	0.108	0.147
HK-Hybrid	0.004	0.222	0.217	0.200	0.208

Table 1 : Parameters of K-means, DBSCAN, Meanshift, agglomerative and HK-hybrid clustering algorithms for banknote authentication data set.

VI. CONCLUSION

In this paper we present a new clustering algorithm known as HK-hybrid which is a combination of K-means and Hierachical clustering techniques. We tested our proposed method using the standard data-sets from UCI machine learning repository and compared our result with four different types of clustering algorithms. The experimental result indicate that our algorithm can produce a higher quality clusters with a smaller standard deviation on the selected data set compared to other clustering methods.

VII. REFERENCES

[1]. M. Eirinaki and M. Vazirgiannis, 2003,"Web Mining for Web Personalization",ACM Transactions on Internet Technology (TOIT), vol. 3, no. 1 pp: 1-27. DOI: 10.1145/643477.643478

[2]. B.Bahmani Firouzi, T. Niknam, and M. Nayeripour, Dec 2008, "A New Evolutionary Algorithm for Cluster Analysis", Proc. of world Academy of Science, Engineering and Technology,vol.36. http://www.waset.org/journals/waset/v46/v46100.pdf

[3]. Gersho and R. Gray, 1992,"Vector Quantization and Signal Compression", Kulwer Acadimec, Boston. http://www-ee.stanford.edu/~gray/

[4]. M. Al- Zoubi, A. Hudaib, A. Huneiti and B. Hammo, 2008,"New Efficient Strategy to Accelerate k-Means Clustering Algorithm, American Journal of Applied Science",vol.5,no.9 pp: 1247-1250. DOI: 10.3844/ajassp.2008.1247. 1250
[5]. M. Celebi, 2009,"Effective Initialization of Kmeans for Color Quantization",Proc. of the IEEE International Conference on Image Processing, pp: 1649-1652. DOI: 10.1.1.151.5281
[6]. M. Borodovsky and J. McIninch, 1993,"Recognition of genes in DNA sequence with ambiguities", Biosystems, vol. 30, issues 1-3, pp: 161-171. DOI:10.1016/0303-2647(93)90068-N
[7]. Data clustering algorithm : https://sites.google.com/site/dataclusteringalgorithms/home.
[8]. Tio Li, Sheng Ma, IFD: Iterative Feature and Data Clustering, 2003,page 472-476.
[9]. Charu C. Aggarwal, Cheng Xiang Zhai, "A SURVEY OF TEXT CLUSTERING ALGORITHMS", Springer US,pp 77-128,Date: 07 January 2012
[10]. [10]. M. V. Jagannatha Reddy,B. Kavitha, "Clustering the Mixed Numerical and Categorical Dataset using
[11]. Similarity Weight and Filter Method", International Journal of Database Theory & Application,Mar 2012, Vol. 5 Issue 1, p121
[12]. [11]. Jennifer Dy, "A Unified View on Clustering BinaryData", Springer,Machine Learning
[13]. March 2006, Volume 62, Issue 3, pp 199–215
[14]. [12] Kosha Kothari1 Ompriya Kale2 "Enhancing the Efficiency and Scalability of Big Data Using Clustering Algorithms", (IJSRD/Vol. 3/Issue 03/2015/896).
[15]. L. Kaufman, and P. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis, (John Wiley & Sons). DOI: 10.1002/9780470316801
[16]. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, 1996. From Data Mining to Knowledge Discovery in Databases, Advances in Knowledge Discovery and Data Mining, AAAI/MIT press. DOI: 10.1.1.42.1071
[17]. XindongWu and et. Al., 2008. Top 10 Algorithms in Data Mining, Journal of Knowledge and Information Systems, vol. 14. Issues 1-37. DOI: 10.1007/s10115-007-0114-2 [
[18]. Q. He, X. Jin, C. Du, F. Zhuang and Z. Shi, "Clustering in extreme learning machine feature space," Neurocomputing, Vol.128, pp. 88-95, 2014.
[19]. A. Katal, M. Wazid and R.H. Goudar, "Big data: Issues, challenges, tools and good practices," Contemporary Computing (IC3), 2013 Sixth International Conference on, IEEE, 2013.
[20]. Izhar Ahmad, "K-Mean and K-Prototype Algorithms Performance Analysis", American Review of Mathematics and Statistics March 2014, Vol. 2, No. 1, pp. 95-109
[21]. Demchenko, Yuri, P. Grosso, C. Laat and P. Membrey, "Addressing big data issues in scientific data infrastructure," Collaboration Technologies and Systems (CTS), 2013 International Conference on, pp. 8-55. IEEE, 2013
[22]. UCI. Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.