Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 6, 2387-2395

An Improved K-Means Clustering Algorithm For Pattern Discovery In Data Mining

Dr.K.KARPAGAM

Assistant Professor of Computer Science H.H. The Rajah's College(Autonomous) Pudukkottai. (Affiliated to Bharathidasan University, Tiruchirappalli) kkarpaga05@gmail.com

Abstract

In health monitoring systems enormous amount of information has been extricated by media sensors, with the assist of medical diagnosis which produces text, audio, video images (media contents). The traditional method provides immense process lots of complexity so health service provider finds difficulty in analyzing the data. In database the enormous data has been grouped in terms of clustering. Without getting help from class labels the data can be separated into multiple set of data and can receive data inputs in k means clustering. This research works and implements on the disease data of the patient using k-means clustering. The functionality technique has been contributed by the data extraction approaches to reform the heterogeneous information's into useful quality information for making decisions. This article educates application and uses of data mining in assist care medical field. Importantly, we learn large dataset clustering on k-means clustering algorithm and produce an improvement to k-means clustering, which needs k or peripheral amount of data which is passed to the dataset. We suggest an algorithm called as G-means, which avail a greedy method to generate preparatory centroids and from that it takes k or peripheral progress among the given data in datasets to make modification in centre points. Our exploratory outcome which has been used in developing way on the similar data set, displays to us that G-means surpasses k-means in phase of F-scores and entropy. The execution time and coefficient of variance has been executed and it produces best outcomes for G-means clustering.

Keywords: The Ideal, "A-divisor", "A-Potent, Rees quotient", "near subtraction semigroup".

1. INTRODUCTION

In current days in order to provide better healthcare services, healthcare information are obtained from different providers providing health care services such as sensory environment. This data holds overall information like patients detail, treatment and medical tests. The acquired data is very complex and indefinite. So, it is very hard to quickly analyze and sort data in order regarding the health of patient to make important decisions. Data extraction is the operation where the raw information is taken and processing is done for data to make it as a valuable resource. In the application of health care it is a main process for effectively identifying unknown and valuable data's from enormous heterogeneous information example heart monitoring system. In the applications of healthcare unknown diseases can be recognized with the purpose for the disease and appropriate needed medical treatment procedure with the assist of data mining techniques. Drug recommendation systems improving policies of health care and efficient health care policies of individuals can be given to medical researchers for developing healthcare appliances [1]. A powerful tool is

needed to analyze and extract valuable needed information's from the complex data set. The result of this data mining technique is to give maximum advantages to health care organizations for combining patients having same types of diseases or issues in health so the need and effective treatment can be provided by the organizations. Association, classification, clustering are the various types of data mining techniques provided by the service providers of health care for giving report according to the health conditions of the victims. In health care data analysis one of the completely established and most important techniques done by data mining algorithm is classification. The destination point for all information points has been predicted by the data classification approach. Example based on the diseases of the patient can be classified as low risk or high risk patient. Ensemble approach, SVM (Support Vector Machine), decision tree, K-NN (K-nearest neighbour) is the different classification functionality used by classification methods. Based on the assumption of class categories classification is considered as supervised learning approach. One of the unsupervised learning techniques is clustering. No predicted classes are present under cluster method classification. Considering on the similarity measures enormous data has been segregated into various small subgroups or clusters in clustering [2]. For knowing likeness the among the data point this approach has been used. In clustering neither no information nor less information has been used for analyzing the data it is one of the important methods in clustering. The set of data having n points of data's have been portioned in k clusters or group in the one of the mainly used method of data clustering which is called as the k-means clustering algorithm [3]. To aggregation of corpus discernment provision and vastly used appliances the k means is considered has a handy algorithm. With the help on memory information the definite k-means algorithms have been worked, but it could effectively taken out for out of memory datasets. In each cycle one analysis is done for all data set it is considered as the k-means calculation principle issues. To acquire best result it should focus on many cycles. Basically for the significantly enormous local disk data sets it is extremely high in cost for utilization. To maximize the performance of algorithm amount of passes done by k-means have been reduced by the researchers. But comparing too conceivably with destined or classifier limits on nature of outcome this methodology only gives deduced outcomes. Combining it with the local minimum is the key component of k-means, for estimated version it does not load accurate values on it [2]. In this technique a fascinated enquiry arises that whether we can achieve methods of calculation with required less passes on enormous data set and can we handle similar intersect result as k-means algorithm? In this article we produce certain algorithm which is called as G-means. With the guide of greedy approach calculations of initial centroids can be easily obtained using G-means algorithms.

2. Clustering Algorithms

To illustrate best known issues in clustering k-means is the one among unsupervised algorithm used [3]. This methodology follows a simple approach of an available data set via a certain amount of group (leaving k groups) that have been established. The overall idea is to characterize one k centroid for all groups. Characterize area of diverse effect process can be done by keeping centroid set in artifice manner. According to this the best decision is to locate each and every one far away from each another. The next step is to take every pint among given data set and accomplice it with the nearest centroid till it reaches a state where each point has been correlated with a group. Abrupt aggregation is carried out automatically after successful completion of the first step, because of the last step we need to rearrange k new centroid as barycenters of each group. Another binding must be done between the same set of centroids and nearest new centroids after successful production of k new centroids. A cycle will be manufactured. Reorganization shall be done for k centroids which will modify their synchronize areas and at the end of day centroids will not change positions [2,6] because of the effect of cycle. The objective description is displayed below and the main aim of algorithm is to decrease squared error in functions.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 .$$

The represented centroids c_j are $\|x_i^J - c_j\|^2$ for the n dataset points is x_i^j of distance indicator. The four steps of k mean clustering algorithm has given in below order.

(1)The item coordinates which are clustered are represented in space and k elements have been kept randomly.

(2) Assign each item to a space to a cluster that is most common to it.

(3) Compute the k centroid elements and modify their places after the successful assignment of all items in space.

(4) Replicate process (2) and (3) completely till the centoids reach places where there is no modification needed with respect to interval among all group elements. But definitely the technique will stop. But most exceptional model has not been found in k-means calculations. Correlating it to the lowest global objective function which has been started. The algorithm fundamentally depends on the starting arbitrarily elected centre groups. Different set of centroids has been selected at every time to minimize the impact of algorithm which should run at different times, so that initial indications are very tough to compare [7]. K-means algorithm and its result have been depicted in algorithm 1.

Fig 1 describes execution of algorithm one step at every time. It exhibits 3 groups in space where k is equal to 3 (k=3). And they are differentiated with colors to represented data elements (blue, brown and green). They have been kept separately among three parts for getting additional clarifications. Random centroids have been selected in intial fig 1(a) where k amount of initial centroids have been portrayed based on coloring their whole group with similar color. Plus sign is used for characterizing the centroids.

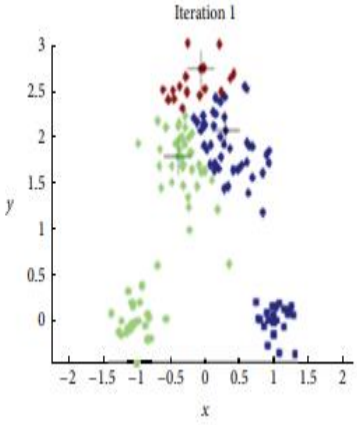
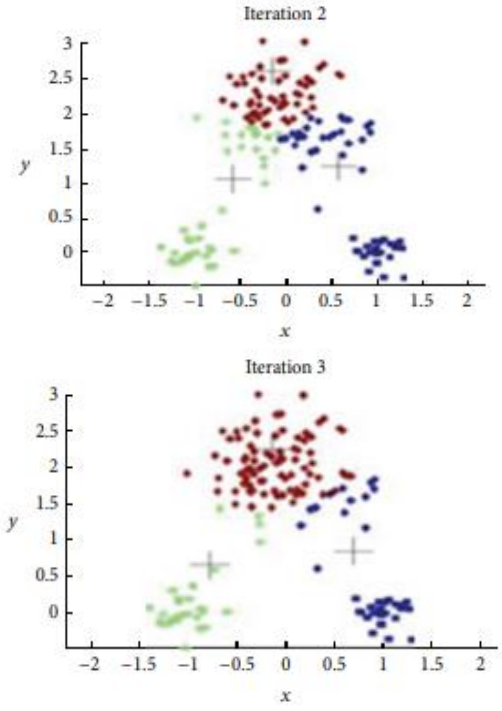


FIG 1(a)

Four iterations have been done in Fig 1(b). The placement of every elemnt to its nearest centroids is done by calculation the interval between starting cluster centroids and each element in the space. The evaluation and fluctuations of cluster centroids are determined at every iteration over this stage.



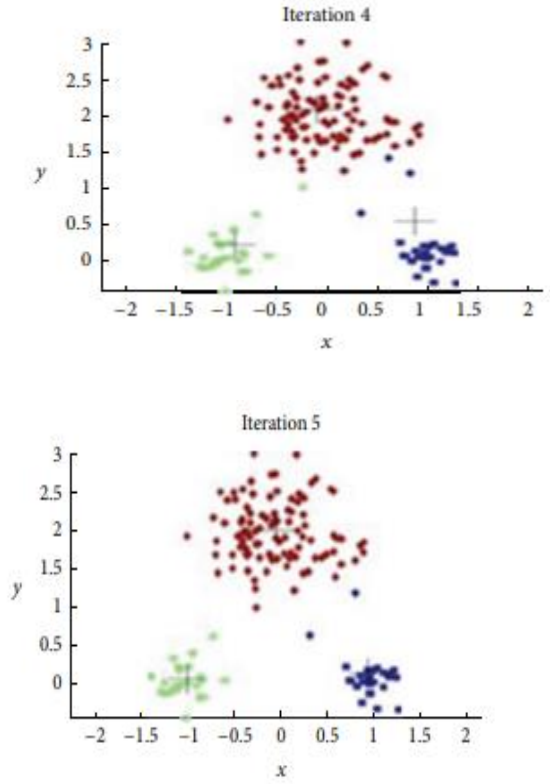


FIG 1 (b)

The algorithm coincides and come together after reaching final position in fig 1 (c). This is achieved when the algorithm compares centroid from the previous step to its current step and intimates that there are no modifications in centroids thus complete clusters have been achieved.

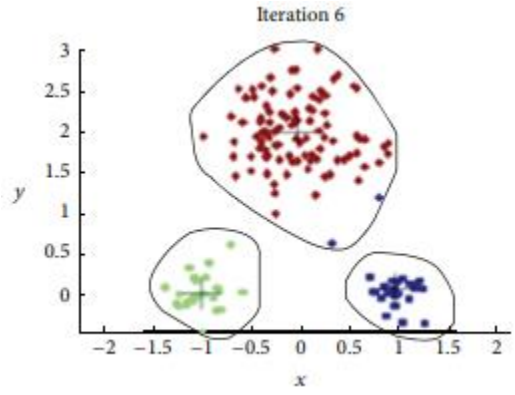


FIG 1 (C)

3.K-Means Algorithm:

Partitioning the k-means algorithm where every clusters centre is represented by mean value of the entity in clusters.

Input:

Amount of clusters K

Object N available in dataset D.

Output:

K clusters grouped inside a set.

Method:

(1) D initial cluster centres is randomly selected from K objects.

(2) Repeat

(3) Most similar objects are reallocated to the cluster, based on mean value of cluster.

(4) Cluster mean should be upgraded i.e. evaluate mean value of objects for every clusters.

(5) No change

4. The Proposed Approach

The quintessence beyond our G-mean algorithm we are going to use same distance and similarity so we have taken G from Greedy algorithm and means from K-means algorithm. For smoother approach we calculate the initial centrids using greedy approach, improvement should be done in centroids array initial random selection in the orginal algorithm. Our first step is deriving all available already presented elements that have greatest degree in space, so we can take a glance at the view of intial clusters from there. On the second step we delete all the centroids that are available in single cluster and we choose k clusters with greatest result of comparability purpose. To check whether the centroids are going too modified or not we iterate rest of data elements. So both the interval and correspondence functions will be accurately executed as similar to that of original k-means [27, 28]. The graph G(V,E) are the given input of G-means algorithm and constant K where V is set of vertices element present in the graph and E is set of edges connection between the

vertices, also mentioning G as an undirected graph, with higher amount of clusters which can be generated[29]. The analytical steps of G-means as follow.

(1) Identifying points with the higher degree (points which are close to the nearest points) is the starting step where the data is passed via the entire dataset this refers to the greedy part in algorithm.

(2) Select the K highest amount of vertices by differentiating the elements and taking in account to how they work in k number of clusters.

(3) Check this identity and distance function from all centroitsin clusters to every element reading via entire database without going above elements of the centroid array.

(4) One of the options either the distance or identity in functions is given to each element.

   (a) The algorithm will never hold this again if it is either placed in similar clusters or centroid arrays.

   (b) To be kept for future lookup for a potential centroid at a later run because the distance function is signalling that it should be placed in another cluster.

   (c) We are taking higher degree element so to replace current centroid and thus both similarities can be achieved. This is a most aberrant case where two of the element is a centroid. Also k integer is lesser than the amount of clusters.

(5) Boundary point of cluster will be achieved at the 4(b) part of algorithm where this point gets ended but the algorithm will keep iterating until taking in account continuously.

(6) Successful coverage array of centroid is declared only if the run has not modified anything in centroid array.

(7) Mapping is done for centroid colors with element colors for displaying entire group to display each clusters loops once via centroid array.

4.1 PSEUDOCODE

The algorithms pseudo code is displayed in pseudocode1

Input: Constant *k* and graph G (V, E)

Output: cluster centers K

Begin

NULL --> CentroidsArray[k][Vertices];

NULL --> ColorArray [Vertices];

Read complete dataset;

Create DegreeArray array;

Variable should be created for number of runs called Runs;

For each k Do

GetMax(DegreeArray);

Vertex=Add centroidsArray[k][0];

Position degree of vertex to -1;

Color array of vertex is set to j;

Increments j and Runs;

K needs to be decremented;

End while Runs != ColorArray has zeros do OR k

Read the complete data set;

Findout neighbours of centroid elements;

Color the same as centroids color them;

An item is centroid or neighbour is created;

Vertex= Add CentroidsArray[Centroid][Vertex];

End

For each k do

CentroidArray[k][0]= resultsArray[k];

End

Output ResultArray;

End

4.2. Example of G-means clustering

In this part we exhibit with two examples the working of the proposed algorithm. In the below two examples the difference will be resulting clusters of k constant number and number of centroids but same number of elements is used.

Fig 2(a) represents the connected graphs starting state. Element in shape is represented by every rounded points. K highest degree regardless of total element is picked up by the algorithm in the second part.

`

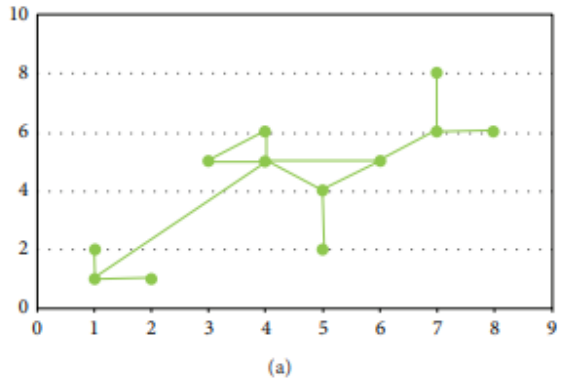


FIG 2 (a) Starting cluster space

Figure 2(b) demonstrates the selection using red rounded mark. To give precise number of element in centroids array the algorithm has been compared with centroid selection constant K. we take two examples when k equal to three it is taken as first and k is equal to four it is taken as second one.

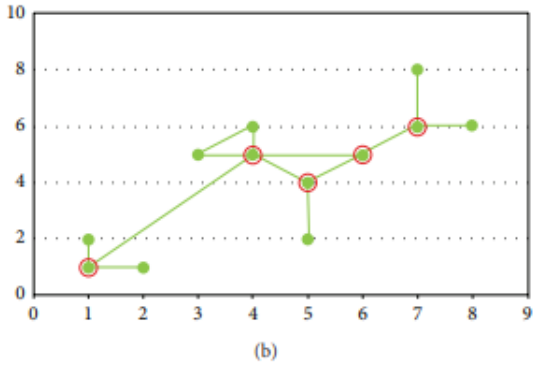


Fig 2 (b) Starting centroid selection

K is equal to three is denoted in the centroid pickup Figure 2(c). The greatest number of element inside every group and distance function decides the process. The purple circled element is known as new centroids. During the execution of the algorithm each element is colored to the corresponding centroids because we assume that centroids not going to change according to example.

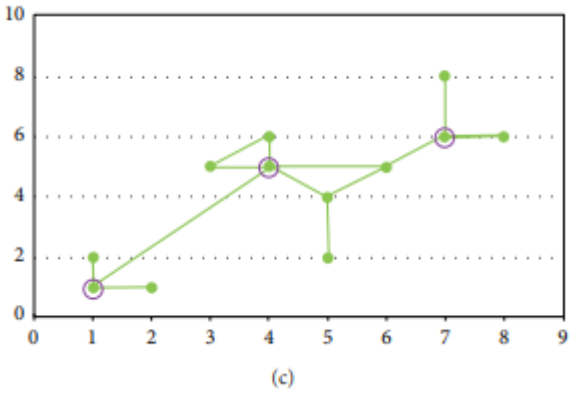


Fig 2 (c) k=3 centroid selection

Successfully completed algorithm is displayed in fig 2 (d). Centroid colored in black denotes three clusters. Each cluster is defined by unique shape. K=4 is next example we demonstrate while selecting resulting centroids distance and similarity function plays an important role. But the modification goes according to step two where comparison is done for centroid with constant k.

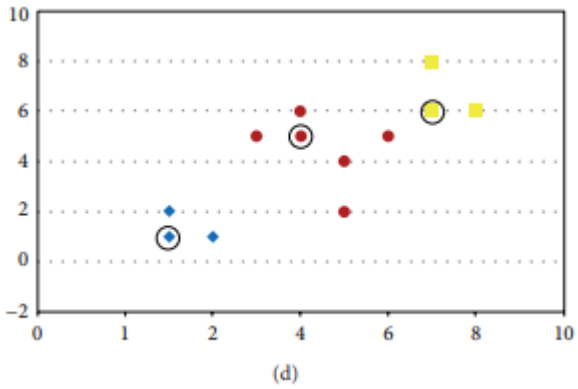


Fig 2 (d) the convergence of algorithms

While the modifications are same for both circled points in red the algorithm faces a difficulty which is displayed in fig 2(e). Each number having similar number of connected node is a common problem faced in greedy approach. But in this problem same function where both neighbouring elements of point (6, 5) are centroids to point (5, 4) which element is not centroid. The cluster will be based on single element if first element has been picked. iF second element has been picked it denotes it as centroids in centroids array.

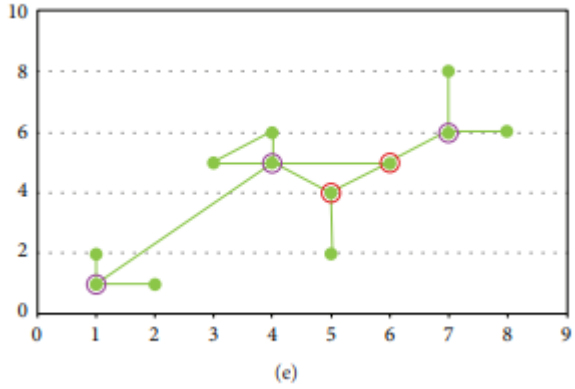


Fig 2 (e) selection problem in centroid

According to similarity function rules correct selection is displayed in fig 2(f)

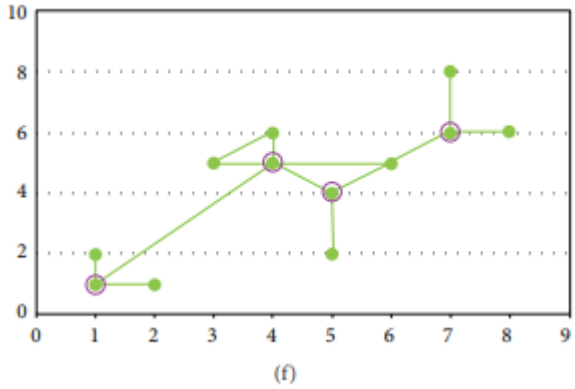


Fig 2 (f) Perfect centroid selection

Successful clusters with centroids are colored in black it replicates how algorithm works. Each cluster is defined by unique shape.

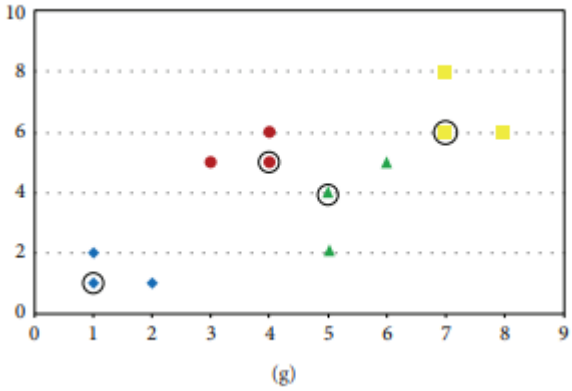


Fig 2 (g) algorithm converse.

Experimental Result:

The k-means and G-means algorithm can be compared based on the entropy of the algorithm. The ICD (International Classification of Diseases) is an important tool for domestic health information exchange. It is used to group diseases related to medical and to extract information from hospital management. ICD code is the diagnosis value of data set. It is contained with categories of 10 diseases. The disease categories based on specification is shown in below table.

.

Table 1. ICD-coded disease classification

Catego-ry	Diagnostic value range	Disease type	Number	Percentage
0	[1,140]	Infectious and parasit-ic diseases	2699	2.7%
1	[140,240)	Tumor	3353	3.4%
2	250	diabetes	8568	8.6%
3	[390,460)or785	Diseases of the circu-latory system	29753	30.0%
4	[460,520)or786	Breathe	14109	14.1%
5	[520,580)or787	digestion	9297	9.3%
6	[580,630)or788	Urogenital	5026	5.1%
7	[710,740)	Musculoskeletal	4826	4.9%
8	[800,1000)	damage	6815	6.8%
9	V or E or else	other	15045	15.1%

From above table we can see that 2.7 % of infectious parasitic disease, tumour disease is 3.4 %, diabetes percentage is 8.6% circulatory disease is 30%, 14.1 % respiratory diseases. 9.3% digestive disease 5.1 % uro genital disease, 4.5 connective tissue disease, 6.8 % injuries and poisoning, 15.1 % for other diseases. K-means algorithm model can be tested and summarized. The training set of processing data is used by process for above operations. The separated set data has been given as input into G-means algorithm model. And result is shown in below fig 3. The G-means algorithm has accurate training results on data processing. Actual disease classification problem is obtained from the categories of clustered data. Range diagnostic value and percentage of number is given after dividing it into categories.

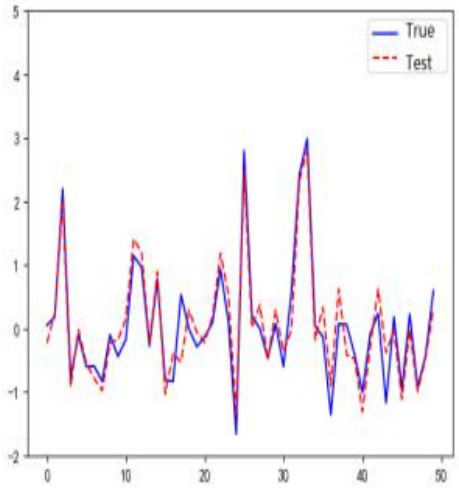


Fig 3. G-means clustering experimental result

Conclusion:

To compare clustering based on similarity matrices there are many research method done. But to obtain similar or minimum different results only k- means clustering algorithm can be used because there are no other algorithms in greedy approach. In this article we presented a new idea of grouping or combining enormous amount of information while minimizing the number of reads on complete data set but in k-means it is not possible. Distributed G-means algorithm is considered as future work where computation power or data can transfer over remote machines. Constant k number of cluster is used by most clustering algorithms this shows the basic way why we are using clusters. Continuously we try to we move our work forward on keeping a k variables rather than of constant this is very useful for us because there is no change in data in real world as k cannot be kept in same fixed position. Our next plan is to work on generalization clustering algorithm. This state's pre-processing should be obtained for both data and algorithm using certain rules so performance and result of clustering algorithm can be increased.

References

1. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. Foundations of Machine Learning; MIT Press: Cambridge, MA, USA, 2012.
2. Bishop, C.M. Neural Networks for Pattern Recognition; Oxford University Press: Oxford, UK, 1995.

3. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering: A review. ACM Comput. Surv. 1999, 31, 264–323. [CrossRef]
4. Ahmed, M.; Choudhury, V.; Uddin, S. Anomaly detection on big data in financial markets. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Sydney, Australia, 31 July–3 August 2017; pp. 998–1001.
5. Ahmed, M. An unsupervised approach of knowledge discovery from big data in social network. EAI Endorsed Trans. Scalable Inf. Syst. 2017, 4, 9. [CrossRef]
6. Ahmed, M. Collective anomaly detection techniques for network traffic Analysis. Ann. Data Sci. 2018, 5, 497–512. [CrossRef]
7. Tondini, S.; Castellan, C.; Medina, M.A.; Pavesi, L. Automatic initialization methods for photonic components on a silicon-based optical switch. Appl. Sci. 2019, 9, 1843. [CrossRef]
8. Ahmed, M.; Mahmood, A.N.; Islam, M.R. A survey of anomaly detection techniques in financial domain Future Gener. Comput. Syst. 2016, 55, 278–288.
9. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA, 1 January 1967; pp. 281–297.
10. Su, M.C.; Chou, C.H. A modified version of the k-means algorithm with a distance based on cluster symmetry. IEEE Trans. Patternanal. Mach. Intell. 2001, 23, 674–680.
11. Cabria I.; Gondra, I. Potential-k-means for load balancing and cost minimization in mobile recycling network. IEEE Syst. J. 2014, 11, 242–249. [CrossRef]
12. Xu, T.S.; Chiang, H.D.; Liu, G.Y.; Tan, C.W. Hierarchical k-means method for clustering large-scale advanced metering infrastructure data. IEEE Trans. Power Deliv. 2015, 32, 609–616. [CrossRef]
13. Qin, J.; Fu, W.; Gao, H.; Zheng, W.X. Distributed k-means algorithm and fuzzy c-means algorithm for sensor networks based on multiagent consensus theory. IEEE Trans. Cybern. 2016, 47, 772–783. [CrossRef]
14. Liu, H.; Wu, J.; Liu, T.; Tao, D.; Fu, Y. Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence. IEEE Trans. Knowl. Data Eng. 2017, 29, 1129–1143. [CrossRef]
15. Adapa, B.; Biswas, D.; Bhardwaj, S.; Raghuraman, S.; Acharyya, A.; Maharatna, K. Coordinate rotation-based low complexity k-means clustering Architecture. IEEE Trans. Very Large Scale Integr. Syst. 2017, 25, 1568–1572. [CrossRef]
16. Jang, H.; Lee, H.; Lee, H.; Kim, C.K.; Chae, H. Sensitivity enhancement of dielectric plasma etching endpoint detection by optical emission spectra with modified k-means cluster analysis. IEEE Trans. Semicond. Manuf. 2017, 30, 17–22. [CrossRef]
17. Yuan, J.; Tian, Y. Practical privacy-preserving mapreduce based k-means clustering over large-scale dataset IEEE Trans. Cloud Comput. 2017, 7, 568–579.
18. Xu, J.; Han, J.; Nie, F.; Li, X. Re-weighted discriminatively embedded k-means for multi-view clustering. IEEE Trans. Image Process. 2017, 26, 3016–3027. [CrossRef]
19. Wu, W.; Peng, M. A data mining approach combining k-means clustering with bagging neural network for short-term wind power forecasting. IEEE Internet Things J. 2017, 4, 979–986. [CrossRef]
20. Yang, J.; Wang, J. Tag clustering algorithm lmmsk: Improved k-means algorithm based on latent semantic analysis. J. Syst. Electron. 2017, 28, 374–384.