ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

**Vol. 4, Issue 1, January 2016**

# Survey on K-Means and Its Variants

Akanksha Choudhary

M. Tech. Scholar, Gurukul Institute of Engineering & Technology, Kota, Rajasthan, India

**ABSTRACT:** Clustering is the process of placing objects in groups. Each group is termed as a cluster. Object assignment to clusters is based on the fact that objects within same cluster are more similar to one another and dissimilar to the objects belonging to different clusters. Dissimilarity is computed on the basis of values of attributes which describe objects. K-means is most widely used method of clustering. It randomly select k objects which act as initial centroids for k clusters and then iteratively assign rest of the objects to these k clusters on the basis of similarities between the objects of clusters. The goal of this survey is to study research and development work done on k-means clustering method.

**KEYWORDS**: Clustering, cluster, k-means, centroid

## I. INTRODUCTION

Many computing applications incorporate data analysis either in their design phase or as part of their on-line operations. The key element in data analysis procedures is the grouping or classification of measurements based on either goodness-of-fit to a postulated model or natural groupings(clustering) revealed through analysis. Cluster analysis or data clustering is the formal study of algorithms and methods for grouping, or clustering, objects according to measured or perceived intrinsic characteristics or similarity **[1]** such that the objects within a cluster are more similar to one another than to the objects belonging to another cluster. Clusters can differ from each other in terms of shape, size and density. An ideal cluster is set of points that is compact and isolated. Cluster analysis does not use category labels or class labels which differentiate it from classification. Data Mining, machine learning, image processing, statistics, biology are the application areas of clustering which has led the development of thousands of clustering algorithms.

## II. CLASSIFICATION OF CLUSTERING METHODS

Categorization of clustering methods is neither straightforward, nor canonical, in reality groups below overlap **[2] :**

Partitioning methods: These methods classify data into partitions on the basis of some criterion function. After creating initial partitions, iterative relocation technique is used that attempts to improve the partitioning by relocating objects from one cluster to another. Generally adopted criterion for good partitioning is that objects in the same cluster should be similar to each other and dissimilar to the objects of different cluster.

Hierarchical methods: These methods create a hierarchy or a tree type decomposition of data objects. It can be of two types agglomerative and divisive. Agglomerative is bottom-up approach and starts with each object forming a separate cluster. It then successively merges clusters together until a stopping criterion is satisfied. The divisive approach is top-down approach and begins with all the objects in the same cluster. In each iteration, a cluster is broken up into smaller clusters until a stopping criterion is met.

Density based methods: These methods allow a cluster to grow until number of objects or data points exceeds some threshold. Unlike most of the partitioning methods that can find spherical shaped clusters only, these methods can discover arbitrary shaped clusters.

Grid based methods: These methods quantize object space into finite number of cells that form a grid structure. All the clustering operations are performed on this quantized space.

Constraint based clustering: Here clustering is performed on the basis of user specified or application oriented constraints. A constraint provides information about user's expectation or required properties of the clustering result.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

## III. K-MEANS

The k-means is a popular clustering method which falls in the category of partitioning based clustering method. The name comes from representing each cluster by the mean (called as centroid) of its data points. The objective of the method is to partition set of say *n* objects into say *k* ( required number of clusters which is given as input to the method) clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. To measure similarity, mean value of the objects within a cluster is used. After randomly picking up *k* initial cluster centroids, the method iteratively relocate the objects within clusters by moving them from one cluster to another, on the basis of the Euclidean distance between the cluster centroid and the object. After this reassignment of objects within clusters, for each cluster its mean value is recomputed which gives new cluster centroid. This process is repeated until the criterion function converges. Generally the square error criterion is used which can be defined as:

$$E = \sum_{i=1}^{k} \sum_{X \in C_i} |X - M_i|^2$$

Where $(X_1, X_2, X_3, \ldots\ldots, X_n)$ is set of objects with $m$ dimensions for each object, $M_i$ is the mean of cluster $C_i$,then k-means method aims to partition these $n$ objects into say $k$ clusters so as to minimize square error criterion. This criterion tries to make k clusters as compact as possible.

**Algorithmic steps of k-means [3]**

Input: *k* – Number of clusters
*D* – Dataset containing *n* objects
Output: Set of *k* clusters
Method:

1) Dynamically choose *k* objects from *D* as initial centroid;
2) Repeat
a. (Re)assign each object to cluster to which the object is most similar,
Based on the mean value of the objects in the cluster;
b. Update cluster mean
3) Until no change;

The wide popularity of k-means is well deserved as it is simple, straightforward and easy to implement. But it also suffers from all the usual suspects :

1. It is a local search heuristics and is sensitive to initial centroids and so may find a local minimum solution rather than a global one**[4]**. So it is recommended to do multiple runs and average the result.
2. Number of clusters is required to be feed as input even if it is unknown. Therefore multiple trials are necessary to find the best amount of clusters **[5]**.
3. Method is sensitive to outliers as it is based on mean of object values. Outliers can distort the distribution of data.
4. Method does not define sample mean for nominal values **[4]**.
5. It works only for clusters that are spherical in shape**[4].**

To overcome these limitations and to make it more accurate and efficient, many developers and researchers have proposed their variants for the original k-means method.

## IV. VARIANTS FOR K-MEANS

**Variants For Centroid Initialization :** As mentioned earlier, traditional k-means randomly picks up initial centroids. When chosen centroids are close to the final solution, k-means has high possibility of finding correct cluster center. Otherwise it will lead to incorrect result. Because of random selection of centroids, it does not guarantee unique clustering results and may find a local minimum solution. To overcome this drawback of traditional k-means several variants have been proposed. A brief description of some of them is given below:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

Vol. 4, Issue 1, January 2016

To generate accurate initial centroids rather than picking them randomly **Author[6]** used dissimilarity matrix to create Huffman tree. The output of Huffman tree is taken as initial centroids for k-mean. Compared to traditional k-mean, proposed method is found to be more accurate.

**Authors[7]** have proposed hierarchical k-means which combines k-means and hierarchical algorithm. The method executes k-means for some fixed number of times and then apply hierarchical algorithm on centroids obtained as a result from executions of k-means. The centroids thus obtained from hierarchical algorithm are then used as initial centroids for k-means. However, authors have suggested that their method works better (in terms of speed) as compare to traditional k-means for complex clustering task (large numbers of data set and many dimensional attributes).

**Authors[8]** have proposed a heuristic method that takes less computational time than traditional k-means. This method finds weighted average score of the dataset by calculating average of data points and then the method sort data points. The nearest possible data point to the mean is chosen as initial centroid.

**Author's[9]** method picks up initial centroids randomly and the remaining centroids are selected as the data point that has the greatest minimum-distance to the previously selected centroid. This method was originally developed as a 2-approximation to k-centre clustering problem.

k-means++ **[10]** is one of the most popular centroid initialization technique because it makes k-means converge faster ( reducing number of iterations) with better sum of squared distances error. It is a method to initialize the *k* centroids where *k(required number of clusters)* is given as an input to the k-means algorithm. Probabilities depending on the minimum-distance from a point *x* to the previously selected centers are used to select next centre. The greedy version of this method probabilistically selects log(K) centers in each round and then greedily selects the centroid that most reduces the sum of squared error. This modification aims to avoid the unlikely event of choosing two centers that are close to each other.

**Variants Based Distance Metrics :** k-means method is based on Euclidean distance or Euclidean metric, which is commonly used to evaluate proximity between objects. It works well when a dataset has compact clusters. Because of the usage of this distance metric k-means finds spherical or ball-shaped clusters in data**[1]**. Below is the brief description of few of the variants of k-means in which developers have experimented with distance metrics other than Euclidean distance metric.

**Authors[11]** have proposed a novel dissimilarity measure, named density sensitive distance metric which causes the method to identify complex non-convex clusters thus generalizing the application area of original k-means algorithm.

**Author[12]** has used three types of distance metrics namely Euclidean, Manhattan, Minkowski and have found out that k-means with Euclidean distance metric gives best result. Dummy dataset have been used for experiment.

**Author[13]** has experimented with Euclidean, Manhattan, Cosine and Correlation distance metrics on Iris and Wine datasets. In terms of computational time, Manhattan distance based k-means was found to be best and cosine distance based k-means was worst.

**Variants for Improving k-means Accuracy :** Developers have employed various techniques to make k-means more accurate, below is the description of few of them :

By incorporating background knowledge **Author[14]** proposed constraint based k-means. Author has suggested that in most of the cases the experimenter has some background knowledge about which instances should be and should not be grouped together. This information is expressed in terms of instance-level constraints and is given as input. The modified algorithm assign an object to a cluster only when none of the constraints are violated. If a legal cluster is not found for an object then method returns empty cluster. From the result of the experiment one can say that incorporation of background knowledge has improved accuracy of traditional k-means significantly.

As per **Author[15]** applying standardization before clustering leads to better quality, efficient and accurate cluster result. Author has experimented on min-max, z-score and decimal scaling techniques and concluded that among the three techniques, z-score provides best result for infectious diseases dataset with improved accuracy over traditional k-means. However author has commented that the selection of standardization technique should be done in accordance with the nature of the chosen dataset.

**Author's[16]** method first normalizes(min-max normalization) the data to improve effectiveness and accuracy of the result, and then picks up those data points as initial centroids which are nearest to the mean value.

**Authors[17]** have applied normalization to pursue the goal of equalizing the size or magnitude and the variability of proximity indices such as Euclidean distance. This can also be seen as a way to adjust the relative

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

**Vol. 4, Issue 1, January 2016**

weighting of the attributes. Author's method applies normalization on five clustering algorithms namely, single linkage, complete linkage, average linkage, k-means and shared nearest neighbor and draws the conclusion that best results are obtained by applying some sort of normalization on datasets.

**Few other popular variants:**

**k-medians** is a variation of k-means where in place of calculating mean to compute centroid of the cluster, median is computed. This method uses Manhattan or 1-norm distance. Outliers have lesser impact on this method as compare to k-means which uses 2-norm or Euclidean distance**[18]**.

**k-medoids** tries to eliminate k-means's sensitivity towards outliers which can distort the distribution of data. This method chooses an object per cluster as representative of that cluster and rest of the objects are assigned to cluster on the basis of their similarity with the representative object. Then partitioning is performed on the basis of dissimilarities between rest of the objects and the representative objects. This is also termed as absolute error criterion. The algorithm iterates until each representative object actually becomes medoid or is the central most object of the cluster.

**k-modes** method is a technique of clustering categorical data. K-modes modifies k-means by replacing Euclidean distance metric with simple matching dissimilarity measure, using modes to represent cluster centers and updating modes with the most frequent categorical values in each iteration of clustering process.

## V. CONCLUSION

k-means is one of the most extensively used data mining algorithm and so is an important topic of research. In this paper we have presented a survey of a few of research work done, to modify and enhance k-means. The enhancements are done to overcome its shortcomings and to make it efficient and accurate in term of cluster quality. However k-means is still at the stage of exploration and development and there exist areas of improvements for efficient and universally accepted methods to initialize centroid, to enhance cluster quality and to achieve more accuracy.

## REFERENCES

1. Anil K. Jain, " Data Clustering: 50 Years Beyond K-Means", Department of Computer Science & Engineering, Michigan State University, East Lansing, Michigan 48824 USA
2. Pavel Berkhin , "Survey of Clustering Data Mining Techniques", Technical Report, Accure Software, San Jose, CA, 2002
3. Jiawei Han, Micheline Kamber, "Data Mining : Concepts and techniques", 2nd Edition.
4. Sami Ayramo, Tommi Karkkainen " Introduction to partitioning-based cluster analysis methods with a robust example", Reports of the Department of Mathematical Information Technology; Series C: Software and Computational Engineering, No. C. 1/2006.
5. Laurence Morissette, Sylvain Chartier, "The k-means clustering technique: General considerations and implementation in Mathematica" , Tutorials in Quantitative Methods for Psychology, Vol. 9, Issue 1, pp. 15-24,2013.
6. Wang Shunye "An Improved K-means Clustering Algorithm Based on Dissimilarity", IEEE International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC),pp. 2629-2633, 2013, Shenyang, China
7. Kohei Arai, Ali Ridho Barakbah, "Hierarchical K-means: an algorithm for centroids initialization for K-means ", Reports of the Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007
8. Md Sohrab Mahmud, Md. Mostafizer Rahman, Md. Nasim Akhtar, " Improvement of K-means Clustering algorithm with better initial centroids based on weighted average", IEEE 7th International Conference on Electrical and Computer Engineering, 2012, pp. 647-650.
9. T. Gonzalez, "Clustering to minimize the maximum intercluster distance". *Theoretical Computer Science*, Vol. 38,pp. 293–306, 1985.
10. D. Arthur, S. Vassilvitskii, "k-means++: The advantages of careful seeding", *Proceedings of the 18th annual ACM-SIAM symposium on discrete algorithms*, pp. 1027–1035, 2007.
11. Ling Wang, Liefeng Bo, and Licheng Jiao, "A Modified K-Means Clustering with a Density-Sensitive Distance Metric", Institute of Intelligent Information Processing, Xidian University Xi'an 710071, China .
12. Archana Singh, Avantika Yadav, Ajay Rana, "K-means with Three different Distance Metrics", International Journal of Computer Applications, Vol. 67, No.10, 2013
13. Dibya Jyoti Bora, Dr. Anil Kumar Gupta,"Effect of Different Distance Measures on thePerformance of K-Means Algorithm: An Experimental Study in Matlab", International Journal of Computer Science and Information Technologies, Vol. 5 (2) , pp. 2501-2506, 2014.
14. Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schroedl, "Constrained K-means Clustering with Background Knowledge", ICML Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577–584, 2001.
15. Ismail Bin Mohamad, Dauda Usman, "Standardization and Its Effects on K-Means Clustering Algorithm", Research Journal of Applied Sciences, Engineering and Technology, Vol. 6, 2013.
16. Deepali Virmani, Shewta Taneja, Geetika Malhotra, "Normalization based K means Clustering Algorithm" , International Journal Of Advanced Engineering Research and Science, Vol. 2, Issue 2,2015.
17. Marcilio C. P. de Souto, Daniel A.S. Araujo, Ivan G. Costa, Rodrigo G.F. Soares, Teresa B. Ludermir, Alexander Schliep, "Comparative Study On Normalization Procedures for Cluster Analysis of Gene Expression Datasets", IEEE World Congress on Computational Intelligence , pp. 2792-2798, 2008.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

**Vol. 4, Issue 1, January 2016**

18. P. S. Bradley, O. L. Mangasarian, W. N. Street, "Clustering via Concave Minimization", Advances in Neural Information Processing Systems, Vol. 9,1997.

## BIOGRAPHY

**Akanksha Choudhary** is an M. Tech. Scholar from Gurukul Institute of Engineering & Technology, Kota, Rajasthan. She is pursuing her M. Tech. in computer science. She has received her B. Tech. From S.G.S.I.T.S. Indore, India..