



## Comparative Analysis of K-Means and the Variants

Dr. Namita Gupta, Astik Gupta  
Maharaja Agrasen Institute of Technology, PSP Area,  
Plot No.1, Sector-22, Rohini, Delhi-110086, India.

**Abstract:** *K-means clustering is a partition-based clustering technique, used to create K-clusters. However, K-means suffers through some drawbacks, which include the fact that it converges in local optimum solution, it requires the user to input the number of clusters, sensitivity to its initial conditions, etc. To solve these issues, various extensions of K-means have been proposed, which include kernel k-means, global k-means, global kernel k-means, X means, K means++. Each of these extensions has their own pros and cons, making each of them useful in separate scenarios of cluster analysis. Thus, the aim of this paper is to study and compare the above mentioned extensions of K-means.*

**Keywords:** *Cluster, K-means, Global K-means, Kernel K-means, Global Kernel K-means, K means++, X-means*

### I. INTRODUCTION

Data mining is a process of collecting, searching through and analyzing a large amount of data sets in a database, as to discover patterns or relationships. These data sets differ from each other in size with respect to number of objects and dimensions, or they contain different data types etc. Clusters help users understand the natural grouping or structuring in a data set.

<sup>[1]</sup>Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. The study of internal structure of a complex data set can be called as cluster analysis. The data sets can be linearly separable or non-linearly separable. For linearly separable, a linear plane can be found that separates the data clusters. Non-linear data sets can be plotted in complex data space dimensions. Cluster analysis is used in Search engines, in some areas of medical analysis like analysis of anti-microbial activity, in classification of various cancerous and non-cancerous data sets and in many more places.

There are majorly two types of clustering methods: Partitioning-based clustering methods and Hierarchical clustering methods. Partition-based clustering algorithms gives a flat partition, as it only provides single set of clusters, with no particular organization or structure within them. But it could easily be the case that some clusters could, themselves, be closely related to other clusters, and more distantly related to others. So sometimes a hierarchical clustering is much more preferable, which is depicted by a tree or dendrogram. The major algorithms to perform hierarchical clustering method are : agglomerative and decisive methods.

Partitioning-based clustering methods are flexible methods based on iterative relocation of data points between clusters. This basically means that with each iteration, the value of criterion function is reduced until the convergence is obtained. The famous algorithm of 'K-means' and related algorithms fall under this category. K-means and its variants algorithms like X-means, K means++, kernel K-means, global K-means and kernel global K-means are discussed in detail in next section.

A partitioning method first creates an initial set of k partitions, where parameter k is the desired number of clusters as output. It then uses an iterative relocation technique that attempts to improve the partitioning of the data points. <sup>[2]</sup>This iterative method is referred to as K-means technique. However, it suffers from two major drawbacks. First, the solution of K-means depends on the initial positions of the cluster centres, resulting in poor minima. Second, it can only find linearly separable clusters. This gave rise to some changes needed in K-means clustering technique.

First limitation is overcome by the method of global K-means. Here, the cluster centres are allocated dynamically. However, its major drawback is that the algorithm takes high computational time.<sup>[2]</sup> Also the second limitation of K-means clustering still persists.

Second limitation is overcome by the method of kernel K-means. This provides a major advantage over standard k-means and allows to cluster points, if a positive definite matrix of similar values is given. But this method does not overcome the first limitation of K-means clustering.

Global kernel K-means technique is able to overcome both the limitations of K-means, together. It uses property of global K-means to initialise clusters and is able to solve the problem of non-linearly separable clusters using the property of Kernel K-means.

K-means++ is another algorithm which follows partitioning-based clustering method to produce more efficient results. <sup>[4]</sup> Different results in different number of iterations are possible in different runs.<sup>[4]</sup> To obtain good results in less number of iterations, the k-means++ has to be run number of times.

K-means algorithm renders useless in case the value of 'k' is not provided. <sup>[5]</sup>To overcome this problem, algorithm of X-means was proposed which quickly estimates the appropriate value of 'k' and provides efficient results.

Each cluster formed in these algorithms has centroid of their own, each having some distance between them. There are various methods used to find the distance between the clusters. Few of the commonly used methods used to calculate distance between these clusters are Euclidian distance measure, Manhattan distance measure, Minkowski distance measure, etc. These clusters also have solution for the given data set. If the solution obtained is optimum within neighbouring set of data, it is called as a local optimum solution. If the solution is optimum among all possible solutions of the data set, then it is called global optimum solution. Contributions of different authors to improve K-means clustering are discussed in next section.

## II. RELATED WORK

Khaled Alsabti, Sanjay Ranka and Vineet Singh proposed an efficient algorithm to perform K-clustering<sup>[6]</sup>. Groups in K-means are originally identified by minimizing the clustering error, which is defined by the sum of the squared Euclidean distances between each data set point and their corresponding cluster centre. Clustering obtained in this procedure is called centroid method of clustering. It is improved by inculcating the logic of trees in clusters such that the calculation has to be performed in internal nodes and not the patterns itself, which has significantly reduced the time utilised in calculations for the next iterations.

Aristidis Likas, Nikos Vlassis and Jakob J.Verbeek proposed global K-means<sup>[8]</sup> technique in 2001. This method proceeds as follows: As the optimal solution to the 1-clustering problem corresponds to the data set centroid, it is begun by solving the 2-clustering problem. K-means is run N times, each time starting with the following initial cluster centres: one cluster centre is always placed at the optimal position of the 1-clustering problem and the other, during the nth, is initially placed at ' $x_n$ '. The solution with the lowest clustering error is kept as the solution of the 2-clustering problem. The above procedure is repeated until ' $k=i$ ', where 'i' is the number used in the iteration to run the loop.

Inderjit Dhillon, Yuqiang Guan and Brian Kulis, provided another improvement in kernel K-means<sup>[3]</sup> technique and compared it to be mathematically equivalent to weighted graph partitioning objective. It is implicitly mapping points to a higher-dimensional space, kernel k-means can discover clusters that are non-linearly separable in input space, using the kernel matrix. It can be shown that any positive semi-definite matrix (where,  $z^T M z$  is positive for every non-zero column vector  $z$  in a symmetric matrix  $M$ )  $M$  can be thought of as a kernel matrix.

Grigorios F. Tzortzis and Aristidis C. Likas propose a Global Kernel K-means<sup>[2]</sup> technique which overcame both the major limitations of K-means. Global kernel K-means maps the data set points from input space to a higher dimensional feature space with the help of a kernel matrix, similar to the kernel K-means technique. Global kernel K-means finds solutions to the clustering problem by incrementally and deterministically adding a new cluster at each stage of the loop and by using kernel K-means as a local search procedure instead of initializing all clusters at the beginning of the execution, similar to global K-means.

Karteeka Pavan, Allam Appa Rao, A.V. Dattatreya Rao and G.R. Sridhar proposed a Single Pass Seed Selection(SPSS)<sup>[4]</sup> technique based algorithm for K-means++. In SPSS technique, minimum probable distance of clusters is calculated, based on the point which was closer to more number of other points in the data set.

Dan Pelleg and Andrew Moore propose an X-means<sup>[5]</sup> algorithm with some modifications. This algorithm goes into action after each run of K-means, making local decisions about which subset of the current centroid should split themselves in order to better fit the data. This splitting decision is done by computing the Bayesian Information Criterion (BIC).

All the variants of K-means mentioned above improve K-means clustering but, none of them are perfect on its own. Each variant has its own limitations. Next section includes the comparison of the most popular K-means variants with respect to few parameters.

## III. COMPARISON OF K-MEANS AND VARIANTS

K-means, Global K-means, Kernel K-means, Global kernel K-means, K-means++ and X-means are compared in the table below in accordance to the following points:

1. Prerequisite Values: Values which are required to be programmed before the start of the program.
2. Inputs: Values which are required to be entered by the user in program.
3. Nature of Clusters: This is referred to whether the final cluster obtained is local or global. If the solution obtained is optimum within neighbouring set of data, it is called local optimum solution. If the solution is optimum in all possible solutions, it is called global optimum solution.
4. Dimensionality of Clusters: This is referred to whether the final cluster obtained is linearly separable or not. Linearly separable data sets can be plotted on linear planes whereas, non-linear separable are data sets which are plotted in complex planes.
5. Time Complexity proportional to clusters: This is referred to the time taken, in proportionality, to compute the clusters of data through various algorithms.

**Table I. Shows the comparison between various K-means variants**

| Description                                 | K-Means                                        | Global K-Means                                      | Kernel K-Means                                 | Global Kernel K-Means                               | K Means++                                      | X-Means                                             |
|---------------------------------------------|------------------------------------------------|-----------------------------------------------------|------------------------------------------------|-----------------------------------------------------|------------------------------------------------|-----------------------------------------------------|
| Prerequisite Values                         | Value for $k$                                  | Value for $k$                                       | Value for $k$                                  | Value for $k$                                       | Value for $k$                                  | Lower bound and upper bound value for $k$           |
| Input(s)                                    | Data set, Initial $k$ clusters centre          | Data set                                            | Data set, Initial $k$ clusters centre          | Data set                                            | Data set, Initial $k$ clusters centre          | Data set                                            |
| Nature of final cluster set                 | Depends on initial cluster set (Local Optimum) | Independent on initial cluster set (Global Optimum) | Depends on initial cluster set (Local Optimum) | Independent on initial cluster set (Global Optimum) | Depends on initial cluster set (Local Optimum) | Independent on initial cluster set (Global Optimum) |
| Dimensionality of Final Cluster Set         | Linear                                         | Linear                                              | Multi-dimensional                              | Multi-dimensional                                   | Linear                                         | Linear                                              |
| Time Complexity of the Algorithm Depends on | N                                              | $N^3$                                               | $N^2$                                          | $N^2$                                               | Log N                                          | N                                                   |

Here, N: Size of the data sets

K: Number of clusters

#### IV. CONCLUSION

In this paper, K-means and its variants clustering are discussed and compared:

- K-means technique depends on initial cluster set and hence, provides with local optimum solution. It is also, not able to solve non-linear data sets.
- Global K-means is independent of the initial cluster positions and hence, produces global optimum solution. However, the problem of linearity still persists.
- Kernel K-means depends on initial cluster positions and thus, provides local optimum solution. However, it solves the problem of linearity i.e. it can solve non-linear data sets.
- Global Kernel K-means technique is independent of initial cluster sets, providing with global optimum solution and also, is able to solve non-linear data sets. However, it has high computational requirements.
- K means++ provides with local optimum solution and is unable to solve non-linear data sets. It has a special use to find the best possible solution for a data set. However, there is a drawback of implementing multiple runs.
- X means provides with global optimum solution but, has is unable to solve non-linear data sets. It also finds a special use to solve cases where, we know the range of K instead of its exact value.

#### REFERENCES

- [1] Sami A'yra'mo and Tommi Karkkainen, "Introduction to partitioning-based clustering methods with a robust example", Reports of the Department of Mathematical Information Technology Series C. Software and Computational Engineering No.C. 1/2006.
- [2] Grigorios F. Tzortzis and Aristidis C. Likas, "The Global Kernel K-Means Algorithm for Clustering in Feature Space", IEEE Transactions on Neural Networks, VOL.20, NO. 7, JULY 2009.
- [3] Inderjit Dhillon, Yuqiang Guan and Brian Kulis, "A Unified View of Kernel k-means, Spectral Clustering and Graph Cuts", UTCS Technical Report #TR-04-25 February 18, 2005.
- [4] K. Karteeka Pavan, Allam Appa Rao, A.V. Dattatreya Rao and G.R. Sridhar, "Single Pass Seed Selection Algorithm for k-Means", Journal of Computer Science 6 (1): 60-66, 2010.
- [5] Dan Pelleg and Andrew Moore, "X-means : Extending K-means with efficient estimation of the number of clusters", School of Computer Science Carnegie Mellon University, Pittsburg, PA 15213 USA.
- [6] Khaled Alsabti, Sanjay Ranka and Vineet Singh, "An Efficient K-Means Clustering Algorithm".
- [7] Dr. M.P.S Bhatia and Deepika Khurana, "Experimental study of Data clustering using K-means and modified algorithms", International Journal of Data Mining & Knowledge Management Process (IJDMP) Vol.3, No.3, May 2013.
- [8] Aristidis Likas, Nikos Vlassis and Jakob J.Verbeek, "The global K-means clustering algorithm", Pattern Recognition 36 (2003) 451 – 461.
- [9] Juanying Xie, Shuai Jiang, Weixin Xie and Xinbo Gao, "An efficient global K-means clustering algorithm", Journal of Computers, vol. 6, NO. 2, FEBRUARY 2011.
- [10] Chin-Tang Chang, Jim Z. C. Lai and Mu-Der Jeng, "A fuzzy K-means clustering algorithm using cluster centre displacement", Journal of Information Science and Engineering, 27, 995-1009 (2011)