Approximation to the K-Means Clustering Algorithm using PCA

Sathyendranath Malli
Assistant Professor,
MSIS, Manipal

Nagesh H. R.
Professor & Head,
Dept. of ISc &E, AJIET,
Mangaluru

B. Dinesh Rao
Professor,
MSIS, Manipal

ABSTRACT
Healthcare is an emerging domain that produces data exponentially. These massive data contain a wide variety of fields, which lead to a problem in analyzing the information. Clustering is a popular method for analyzing data. Data is split into smaller clusters having similar properties and is then analyzed. The K-Means algorithm [1] is a well-known technique among clustering methods. In this paper, an efficient approximation to the K-means problem targeted for large data by reducing the number of features to one through Principle Component Analysis(PCA) is introduced. This data is clustered in one dimension using the K - means algorithm. Intra-cluster RMS error in the modified algorithm is compared with the K-means algorithm in m dimensions and is found to be reasonable. The time taken by the modified algorithm is significantly less when compared to the K - means algorithm.

Keywords
K-means, RMS error, PCA, Approximation.

1. INTRODUCTION
Clustering is an important method used in data analysis. Clustering is used in areas like bioinformatics, machine-learning, and pattern recognition [4] where partitioning large data into groups is of utmost importance. In other words, the clustering objective is to achieve high intra-cluster similarities. It is difficult and computationally infeasible to assign an instance into a cluster optimally. It requires an exhaustive search before placing the instances into clusters. In K-Means clustering, a set of $k_{1,}$…., k_K centers for the given set of points P of m dimensions are detected, such that the distance between a point p belonging to P and the center k_i, 1<=i<=K is minimum. All the points in P are near to any one of the k_i, 1<=i<=K centers.

In this paper, an approximate K-mean clustering algorithm in which PCA[3] is first applied to reduce the number of dimensions to one, and then k clusters created on one-dimension data is introduced. These clusters in the original feature space are comparable to K-means clustering.

2. MATERIALS AND METHODS
In [4], a new approach is designed to overcome the difficulty in manipulating and analyzing massive data. It is well familiar that the K-means algorithm is capable of handling massive data. The proposed method in [4] is an improved version of the K-means algorithm. The large datasets are recursively partitioned into subsets which are small in number. The individual subset is represented as cluster mass and weight. The weighted K-means algorithm reduces the distance calculation in terms of numbers. The empirical results are extremely well for the proposed approach [4] compare to other techniques i.e. minibatch K-means[12] and the K-means++[2] while comparing the relationship between the quality of approximation and distance computations.

Initialization of centers plays a major role in K - means clustering. Any bad initialization will lead to poor local cluster objects. In [5], an attempt is made to tackle the K-means initialization problem by proposing a new algorithm, i.e., the MinMax K-means algorithm. The algorithm consists of a method that assigns weights to the clusters and clustering is done based on cluster variance. The iterative procedure helps in learning the weights together with the cluster assignments. This approach restricts the limits of the emergence of large variance clusters and also provides high-quality solutions. The effectiveness and robustness of this technique are compared with bad initializations. The experimental results are favorable when compared to K-means' random initializations. The proposed algorithm consists of two routines. The first routine is the minimization step, where instances are assigned to the clusters based on the weighted distances from the cluster centers. The second routine is the maximization step, where the larger variance clusters are expected to decrease as instance distances are far from its center. Apart from weightage and variance, it includes constant factor called the exponent. The value of this constant is between 0 & 1. The changes in the exponent values make changes in weight and variance values. In the MinMax K-means algorithm, initial centers with minimum exponent value are applied to the dataset. The value is increased in each iteration including both minimization and maximization steps.

In [6], the focus is on the assignment step hence they introduce a new approximate K-means algorithm. The assignment step reduces the computational complexity. The points, which exist on or near cluster boundaries, are the most active. These points are altering their cluster assignment at each iteration. These active points are identified and form a group using multiple random spatial partition trees. The group consists of a small number and only these points are considered when allocating a data point to its closest cluster. This approach of clustering is better than the state-of-the-art approximate k-means algorithms that provide better quality and efficiency.

K-means algorithm is effective, but selecting proper initial centers is not always possible. This is even more sensitive especially when the number of clusters increases. In [7], the authors propose an iterative approach. This helps the K-means algorithm to obtain a good quality of the solution. In the proposed algorithm, each iteration removes one cluster(minus) and divides another one(plus) to improve quality. The experiment is conducted for a larger dataset and

with a larger number of clusters. The runtime is compared with the K-means algorithm which is less than 1.22 times. In [8] this paper effort has been done to compare eight initialization methods. These are the most commonly used linear time complexity initialization methods. The experiment is done on real-world and synthetic datasets. As part of the effectiveness criteria, the methods are evaluated after executing 100 times and collecting statistic parameters like minimum, mean, and standard deviation. In [9], a new cost function and distance measurements are introduced based on the co-occurrence of values. The significance of all the attributes is taken into account while calculating the measures during the clustering process. The performance of the proposed modified K-means algorithm for clustering mixed datasets is highly encouraging.

The following algorithm is an approximation algorithm for the K-Means problem. The m dimensional data is reduced to one dimension. Experiments have been conducted with various sizes and types of data sets. K-means algorithm is used in one dimension to get the k clusters.

Algorithm: Modified K-means

Input: D= $\{x_1, x_2. . . x_n\}$ in R^d where R^d is real data in d dimensions.

Output: k mutually disjoint clusters $C_1... C_k$ such that union of the cluster elements is D

Step 1: Perform PCA on D to get One-dimensional data

Step 2: Apply the K-means algorithm on One-dimensional data to get the k clusters

Output: Indices of k clusters.

3. EXPERIMENTS AND RESULTS

The proposed modified K-means clustering method generates K number of clusters effectively. The main feature of the proposed method is to reduce the large dimension data into one dimension using Principal Component Analysis. By applying the K-means algorithm on these datasets, effective clusters are generated. The data objects in each cluster are then mapped into the original dataset. Intra-cluster Root Mean Square (RMS) value ratio of the K-means method and the proposed modified K-means method is calculated.

Three different types of large real-world data sets have been selected to conduct experiments on the proposed algorithm. The first type of data set used is the original MNIST dataset which includes 10 handwriting digits and contains 60,000 training and 10,000 test patterns of 784 dimensions. Using the proposed method, dimensions have been reduced to one using Principal Component Analysis and then data has been clustered into K-clusters. The data objects in each cluster are then mapped into the original feature space of the dataset. Intracluster RMS values for each of the K-clusters are calculated and its ratio with the intracluster RMS values found in the original K- means algorithm is determined.

Table 1 (using original MNIST dataset) shows the comparison of time taken and the accuracy of the K-means algorithm and the modified K-means algorithm. The accuracy is good when there are large databases and fewer number of clusters.

Cases have been considered where PCA returns two dimensions and subsequently, K-means clustering has been applied on two dimensions. Table 2 (using original MNIST dataset) shows the results of two-dimension clustering versus K-means clustering in n dimensions. Accuracy improves in two-dimension clustering when compared to one-dimension clustering. Correspondingly, time taken for one-dimension clustering is less when compared to two-dimension clustering.

Table 1: Results of real-world datasets (MNIST train and test)

Data set (N)	Clusters (K)	**K-means (d-dimension)**	**Modified K-means (PCA = 1)**	**Ratio**
		Time (milliseconds)	Time (milliseconds)	
60,000	2	49.83	0.28	**0.999**
10,000	2	9.176	0.07	**0.998**
60,000	3	68.38	0.6	**0.981**
10,000	3	10.37	0.117	**0.978**
60,000	4	72.49	0.99	**0.967**
10,000	4	11.29	0.157	**0.96**
60,000	5	71.53	1.32	**0.94**
10,000	5	11.76	0.158	**0.93**
60,000	6	107.5	1.8	**0.927**
10,000	6	16.1	0.28	**0.914**

Table 2: Results of real-world datasets (MNIST train and test)

Data set (N)	Clusters (K)	**K-means (d-dimension)**	**Modified K-means (PCA = 2)**	**Ratio**
		Time (milliseconds)	Time (milliseconds)	
60,000	2	49.83	0.62	**0.99**
10,000	2	9.176	0.2	**0.99**
60,000	3	68.38	1.13	**0.98**
10,000	3	10.37	0.166	**0.99**
60,000	4	72.49	2.21	**0.988**
10,000	4	11.29	0.21	**0.99**
60,000	5	71.53	3.09	**0.96**
10,000	5	11.76	0.24	**0.97**
60,000	6	107.5	3.35	**0.955**
10,000	6	16.1	0.3	**0.95**

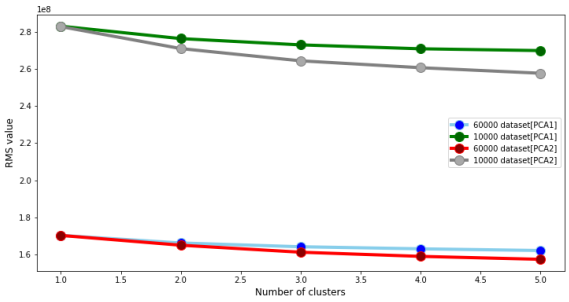

Fig 1. RMS value for modified K-means clustering

In Fig 1 the X-axis represents the number of clusters and Y-axis represents RMS value for each cluster. The experiments are conducted using modified version of k-means clustering algorithm (PCA=1 and PCA=2) for MNIST datasets with sample size 60,000 and 10,000 respectively.

The second type of data set is of Fashion-MNIST [10], a dataset comprising of 70,000 rows from 10 columns. Table 3 and Table 4 show the comparison of time taken and the accuracy of the K-means algorithm with m-dimension and the modified K-means algorithm. The experiment is conducted with PCA values as one and two for the modified K-means algorithm. The experimental results show that accuracy improves in two-dimension clustering when compared to one-dimension.

Table 3: Results of real-world datasets (Fashion-MNIST data)

Clusters (K)	**K-means (d-dimension)**		**Modified K-means (PCA = 1)**		**Ratio**
	Intra cluster RMS value	Time (ms)	Intra cluster RMS value	Time (ms)	
2	2703603377.07	49.35	2718826473.04	0.42	**0.99**
3	2387677619.26	68.44	2545715357.18	0.62	**0.93**
4	2235004570.83	72.84	2493009025.07	1.24	**0.89**

Table 4: Results of real-world datasets (Fashion-MNIST data)

Clusters (K)	**K-means (d-dimension)**		**Modified K-means (PCA = 2)**		**Ratio**
	Intra cluster RMS value	Time (ms)	Intra cluster RMS value	Time (ms)	
2	2703603377.07	49.35	2702545272.28	1.27	**1.00**
3	2387677619.26	68.44	2390077504.51	1.89	**0.99**
4	2235004570.83	72.84	2248711186.99	1.62	**0.99**

The third type of data set is a synthetic 64-dimension data with N=5000 vectors with worm-like shapes[11]. Table 5 and Table 6 show the comparison of time taken and the accuracy of the K-means algorithm with m-dimension and the modified K-means algorithm. The experiment is conducted with PCA values as one and two for the modified K-means algorithm. The results show that accuracy improves in two-dimension clustering when comparing to one-dimension.

Table 5: Results of real-world datasets (Worms)

Clusters (K)	**K-means (d-dimension)**		**Modified K-means (PCA = 1)**		**Ratio**
	Intra cluster RMS value	Time (ms)	Intra cluster RMS value	Time (ms)	
2	6947958332.51	9.85	6955722379.24	0.53	**0.99**
3	6860101250.47	10.84	6920018023.74	1.09	**0.99**
4	6796670693.89	15.14	6912272367.36	2.80	**0.98**

Table 6: Results of real-world datasets (Worms)

Clusters (K)	**K-means (d-dimension)**		**Modified K-means (PCA = 2)**		**Ratio**
	Intra cluster RMS value	Time (ms)	Intra cluster RMS value	Time (ms)	
2	6947958332.51	9.85	6954735596.04	1.35	**0.99**
3	6860101250.47	10.84	6869346536.46	1.15	**0.99**
4	6796670693.89	15.14	6838404449.73	3.13	**0.99**

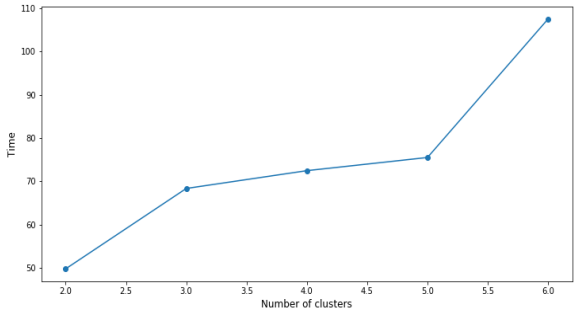

Fig 2. d-dimension data set (Size = 60000)

Fig 2 describes the relationship between the time taken for clustering in m dimension versus the number of clusters. Time taken increases as the number of clusters increases.

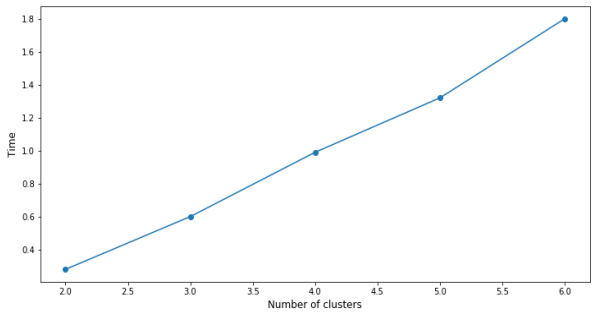

Fig 3. One-dimension data set (Size = 60000)

Fig 3 describes the relationship between the time taken for clustering in one-dimension versus the number of clusters. Time taken increases as the number of clusters increases.

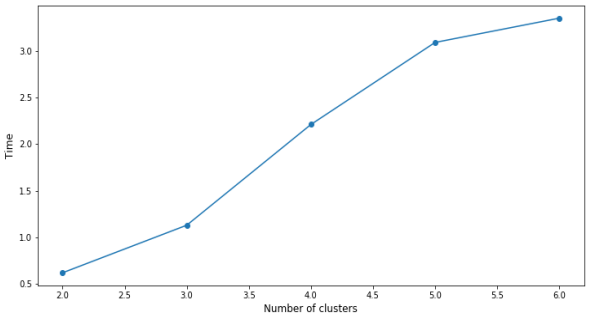

Fig 4. Two-dimension data set (Size=60000)

Fig 4 shows the relationship between the time taken for clustering in two-dimension versus the number of clusters. Time taken increases as the number of clusters increases.

4. CONCLUSION

In this paper, modified K-means method for clustering is proposed. Through the experiments, it is observed that the quality of the results of the proposed modified K-means method is better than the results of the K-means method in terms of time. Accuracy is comparable with that of K means algorithm. The use of two dimensions to cluster increases the time by a small quantity but it also increases the accuracy in clustering. In the experiments, we have calculated the intracluster RMS values and CPU time taken in one-dimensional clustering, two-dimensional clustering, and m dimensional clustering.

5. REFERENCES

[1] S.P Lloyd, Least Squares quantization in PCM, IEEE trans. Inf. Theory 28(2) (1982) 129-136

[2] D. Arthur, S. Vassilvitskii, k-Meansþ þ: the advantages of careful seeding, in ACM-SIAM Symposium on Discrete Algorithms (SODA), 2007, pp. 1027–1035

[3] Hotelling H., Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417–441, and 498–520.

[4] Marco Capóa; An efficient approximation to the K - means clustering for massive data, Knowledge-Based Systems 117 (2017) 56–69

[5] Grigorios Tzortzis n; The MinMax k-Means clustering algorithm, Pattern Recognition 47(2014)2505–2516

[6] Jing Wang; Fast Approximate k-Means via Cluster Closures, 978-1-4673-1228-8/12/2012 IEEE.

[7] Hassan Ismkhan; I-k-means−+: An iterative clustering algorithm based on an enhanced version of the k-means, Pattern Recognition 79 (2018) 402–413

[8] M. E. Celebi, Hassan A, Patricio; A comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Systems with Applications 40 (2013) 200–210

[9] Amir Ahmad, Lipika Dey, A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering 63 (2007) 503–527

[10] Han Xiao, Kashif Rasul, Roland Vollgraf; Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms, https://www.researchgate.net/publication/319312259, 2017

[11] S. Sieranoja and P. Fränti, "Fast and general density peaks clustering", Pattern Recognition Letters, 128, 551-558, December 2019

[12] D. Sculley. Web-scale k-means clustering. In Proceedings of the 19th international conference on World wide web, pages 1177–1178. ACM, 2010.