Experimental Approach of K-Means Clustering

Shrikant A. Shinde[1], Abhilasha V. Biradar[2]
[1]Dept of Computer Engg
[2]Dept of Information Technology
[1, 2]SPPU, Pune, India

Abstract- *Clustering is the process of grouping data elements based on some characteristics of relationship between the elements in the group. Clustering has various applications such as data firmness, data mining, pattern recognition, machine learning and there are different clustering methods. Among many clustering algorithms, the K-means clustering algorithm is broadly used because of its simple computation and fast convergence. However, the K-value of clustering is to be known in advance and the choice of K-value has direct effect on the convergence result. In this paper, we examine what are the clustering algorithms and what are problems to split a cluster of k-means clustering. We have used Manhattan distance formula for calculation of distance between two points. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. We have mentioned advantages and limitations of k-means clustering algorithm.*

Keywords- Clustering, data mining, distance measure, k-means, machine learning.

I. INTRODUCTION

Data Mining

There is a large amount of information accessible in the Information Industry. This information is of no use until the point that it is changed over into useful data. Extraction of data is not only the procedure we have to perform; information mining additionally includes different processes, for example, Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. When every one of these procedures are finished, we would have the capacity to use this data in numerous applications, for example, Fraud Detection, Market Analysis, Production Control, Science Exploration, and so on. At the end, we can state that information mining is the technique of mining learning from information[9].

a. Association

One of the well known techniques of data mining is association rules which are used to find out the relationship or association between various items. The problem of finding relation between items is often termed as market basket analysis. In this problem the presence of items within baskets is identified so that the customers buying nature can be analyzed. This technique is used in inventory management, sales promotion etc [6][10]. The realization of association rules is primarily dependent on finding the frequent sets. This needs multiple passes through the database. The aim of the algorithm is to reduce number of passes by generating a candidate set which should turn out to be frequent sets. Different algorithms are developed to find out the association rules. The algorithm differs on the basis of how they handle candidate sets and how they reduce number of scans on the database. Some of the recent algorithms of association rule mining do not create candidate set. Practically the frequent sets generated are very large in number and this can be restricted by choosing only those items in which the user is interested[8]. Let us consider a set of items and a transaction database which is again a set of transactions. The association rule takes the following form for a transaction database: X=>Y, where X and Y are the sets of items called item sets. There are two important metrics for association rules, support(s) and confidence(c). Since the database is large and users concern about only those.

b. Classification

In general, classification is a two step method. In the first step, known as the learning step, a model that defines a planned set of categories is constructed by analyzing a set of coaching info instances. Every instance is assumed to belong to a predefined category. Within the second step, the model is tested employing a totally different knowledge set that calculate the classification accuracy of the model. If the accuracy of the model is acceptable, the model will be used to classify future instances for that the category label is unknown[4]. At the end, the model acts as a classifier within the call making method. There square measure many techniques that will be used for classification like call tree, Bayesian strategies, rule based algorithms, and Neural Networks. Decision tree classifiers square measure quite widespread techniques because the construction of tree will not need any domain expert data or parameter setting, and is applicable for exploratory knowledge discovery[5]. Call tree will turn out a model with rules that square measure human readable and explainable. Decision Tree has the benefits of straightforward interpretation and understanding for call manufacturers to compare with their domain data for validation and justify their call[3].

c. Clustering

Clustering is one type of unsupervised learning technique in machine learning. Clustering is useful when we do not have labelled data. K-means clustering is one of the many clustering algorithms. Clustering is the process of dividing a set of data objects into clusters so that objects within same cluster are similar to one another resulting dissimilarities with objects within other clusters. A cluster is defined as a collection of data points having certain similarities.

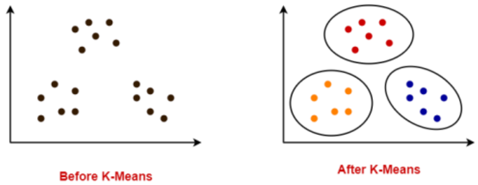


Fig 1: Example of Clustering

In the above figure 1, we can easily identify the three clusters into which the data can be divided; here the similarity criteria is distance: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case we consider geometrical distance). This is called distance-based clustering [7].

Another type of clustering is conceptual clustering; two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures [2].

Clustering partitions the data set such that-

- Each data point belongs to a cluster with the nearest mean.
- Data points belonging to one cluster have high degree of similarity.
- Data points belonging to different clusters have high degree of dissimilarity.

Clustering is broadly used in many applications including business intelligence, DNA analysis in computational biology, security, geographical information system, intrusion detection, image retrieval, intelligent transportation system, music sound features analysis, biochemistry, social studies. In this paper, we are interested in only k-means clustering which is a partition-based method.

II. RELATED WORK

The following are the types of clustering algorithm.

a. Partitioning-based

In such algorithms, all clusters are determined promptly. Initial teams are given and reallocated towards a union. In other words, the partitioning algorithms divide knowledge objects into variety of partitions, wherever every partition represents a cluster. These clusters indicates the subsequent requirements to fulfill: every cluster should contain a minimum of one object, and every object should belong to precisely one cluster. Within the K-means formula, as an example, middle is that the average of all points and coordinates representing the expectation. Within the K-medoids formula, objects that are close to the middle represent the clusters. There are different partitioning algorithms like K-modes, PAM, CLARA, CLARANS and FCM[6].

Algorithm K-means

K decides the number of clusters that are needed finally.

Step 1: K non empty subsets of objects are partitioned (randomly)
Step 2: the partitioning seed points of the clusters currently are the clusters centroids.
Step 3: the cluster with the nearest seed point is assigned with an object
Step 4: repeat step 2 until assignment does not change.

b. Hierarchical Clustering Algorithms

Hierarchical clustering is divided into two main types: agglomerate and discordant.

1. Agglomerate clustering: It is referred as AGNES (Agglomerative Nesting). It works in a bottom-up manner. That is, every object is at first thought-about as a single-element cluster (leaf). At every step of the rule, the two that square measure the foremost similar square measure combined into a replacement larger cluster (nodes). This procedure is iterated till all points square measure member of only one single massive cluster (root) (see figure 2). The result is a tree which might be planned as a dendrogram[2].

2. Divisive clustering: It’s referred as DIANA (Divise Analysis) and it works in top-down manner. The rule is

Associate in Nursing inverse order of AGNES. It starts with the foundation, within which all objects square measure enclosed in a very single cluster. At every step of iteration, the foremost heterogeneous cluster is split into 2. The method is iterated till all objects square measure in their own cluster.

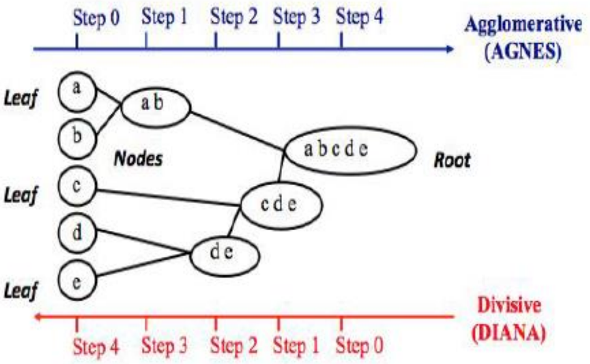


Fig 2: Agglomerative Clustering[2]

As we learned within the k-means, we tend to live the (dis)similarity of observations exploitation distance measures (i.e. euclidian distance, Manhattan distance, etc.) In R, the euclidian distance is employed by default to live the unsimilarity between every combine of observations. As we tend to already understand, it’s straightforward to reason the unsimilarity live between 2 pairs of observations with the get_dist perform.

However, a much bigger question is: however will we live the unsimilarity between 2 clusters of observations? Variety of various cluster agglomeration ways (i.e, linkage methods) are developed to answer to the current question. The foremost common varieties ways are:

•**Maximum or complete linkage clustering**: It calculates all pairwise dissimilarities between parts/the weather in cluster one and therefore the elements in cluster a pair of, and considers the biggest worth (i.e., most value) of those dissimilarities because the distance between the 2 clusters. It tends to provide additional compact clusters.

•**Minimum or single linkage clustering**: It computes all pairwise dissimilarities between parts/the weather in cluster one and therefore the elements in cluster a pair of, and considers the tiniest of those dissimilarities as a linkage criterion. It tends to provide long, “loose” clusters.

•Mean or average linkage clustering: It computes all pairwise dissimilarities between parts/the weather in cluster one and therefore the elements in cluster a pair of, and considers the typical of those dissimilarities because the distance between the 2 clusters.

•Centroid linkage clustering: It computes the unsimilarity between the center of mass for cluster one (a mean vector of length p variables) and therefore the center of mass for cluster a pair of.

•Ward’s minimum variance method: It minimizes the full within-cluster variance. At every step the combine of clusters with minimum between-cluster distance square measure united.

c. Density-based

Here, data objects are divided by their regions of density, property and border. They're closely associated with point to nearest neighbors. A cluster, outlined as a connected dense element, grows in any direction that density results in. So, density-based algorithms are getting a typical shapes by discovering cluster. Also, this provides a natural protection against outliers. So the general density of a degree is analyzed to work out the functions of datasets that effetcts a selected datum. DBSCAN, OPTICS and DENCLUE are a unit of algorithms are use to such a way of remainder noise (ouliers) and see clusters of individual form[8].

d. Grid-based

The house of the data variables is split into grids. the most advantage of this approach is its quick time interval, as a result of it goes through the dataset once to cypher the applied mathematics values for the grids. The accumulate grid data generate grid-based cluster technique and irregular of the amount of information objects that use a homogenous grid to gather regional applied mathematics data, then perform the clump on the grid, rather than the information directly[4]. The performance of a grid-based approach depends on the scale of the grid,that is typically abundant but the scale of the information. However, for extremely irregular data distributions, employing a single uniform grid might not be easy to get the desired clump quality or fulfill the time demand. Wave Cluster and STING are a part of typical samples on this class[3][7].

III. EXPERIMENTAL APPROACH

The K-means algorithm is a simple iterative clustering algorithm. Using the distance as the metric and given the K clusters in the data set, calculate the distance mean, giving the initial centroid, with each cluster described by the centroid.

K-Means Clustering Algorithm:

K-Means Clustering Algorithm involves the following steps-

1. Choose the number of clusters K.
2. Randomly select any K data points as cluster centers.

Select cluster centers in such a way that they are as farther as possible from each other.

3. Calculate the distance between each data point and each cluster center.

The distance may be calculated either by using manhattan distance function or by using euclidean distance formula.

4. Assign each data point to some cluster.

A data point is assigned to that cluster whose center is nearest to that data point.

5. Re-compute the center of newly formed clusters.

The center of a cluster is calculated by taking mean of all the data points contained in that cluster.

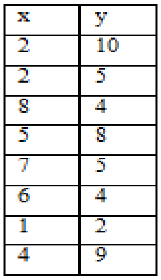
6. Keep repeating the procedure from Step 3 to Step 5 until any of the following stopping criteria is met-
   - Center of newly formed clusters do not change
   - Data points remain present in the same cluster
   - Maximum number of iterations are reached

Step by step example:

Initial data

Cluster the following data into three clusters:

Table 1: Initial Data



x	y
2	10
2	5
8	4
5	8
7	5
6	4
1	2
4	9

- Assume initial cluster centers are: C1(2, 10), C2(5, 8) and C3(1, 2).
- The distance is calculated by using the Manhattan distance formula.
- The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-

  P(a, b) = |x2 – x1| + |y2 – y1|

Use K-Means Algorithm to find the three cluster centers after the second iteration.

We follow the above discussed K-Means Clustering Algorithm-

Iteration-01:

- We calculate the distance of each point from each of the center of the three clusters.
- The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point (2, 10) and each of the center of the three clusters-

Calculating Distance Between C1(2, 10) and (2, 10)-

= |x2 – x1| + |y2 – y1|

= |2 – 2| + |10 – 10|

= 0

Calculating Distance Between C1(2, 10) and (5, 8)-

= |x2 – x1| + |y2 – y1|

= |5 – 2| + |8 – 10|

= 3 + 2

= 5

Calculating Distance Between C1(2, 10) and (1, 2)-

= |x2 – x1| + |y2 – y1|

= |1 – 2| + |2 – 10|

= 1 + 8

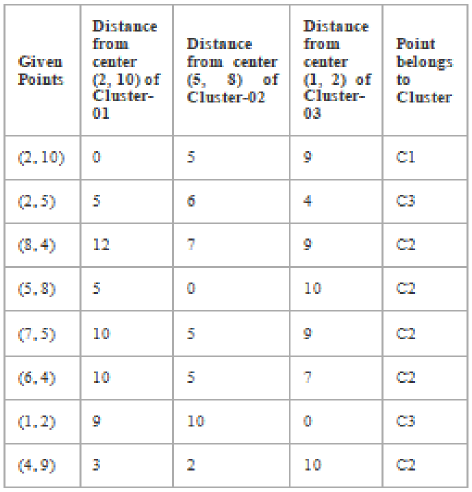
= 9

In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Next,

- We draw a table showing all the results.
- Using the table, we decide which point belongs to which cluster.
- The given point belongs to that cluster whose center is nearest to it.

Table 2: Cluster allocation after 1st iteration



Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (5, 8) of Cluster-02	Distance from center (1, 2) of Cluster-03	Point belongs to Cluster
(2, 10)	0	5	9	C1
(2, 5)	5	6	4	C3
(8, 4)	12	7	9	C2
(5, 8)	5	0	10	C2
(7, 5)	10	5	9	C2
(6, 4)	10	5	7	C2
(1, 2)	9	10	0	C3
(4, 9)	3	2	10	C2

From here, New clusters are-

Cluster 1:

First cluster contains points-

- (2, 10)

Cluster 2:

Second cluster contains points-

- (8, 4)
- (5, 8)
- (7, 5)
- (6, 4)
- (4, 9)

Cluster 3:

Third cluster contains points-

- (2, 5)
- (1, 2)

Now,

- We re-compute the new cluster clusters.
- The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster 1:

We have only one point (2, 10) in Cluster-01.

- So, cluster center remains the same.

For Cluster 2:

Center of Cluster-02
= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5)
= (6, 6)

For Cluster 3:

Center of Cluster-03
= ((2 + 1)/2, (5 + 2)/2)
= (1.5, 3.5)

This is completion of Iteration-01.

Iteration-02:

- We calculate the distance of each point from each of the center of the three clusters.
- The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point (2, 10) and each of the center of the three clusters-

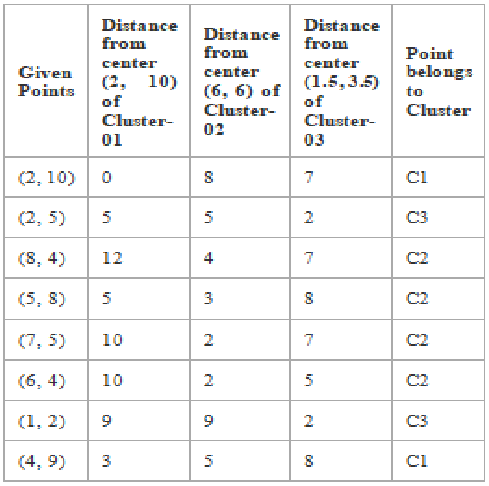
Calculating Distance Between (2, 10) and C1(2, 10)-
= |x2 – x1| + |y2 – y1|
= |2 – 2| + |10 – 10|
= 0

Calculating Distance Between (2, 10) and C2(6, 6)-
= |x2 – x1| + |y2 – y1|
= |6 – 2| + |6 – 10|
= 4 + 4
= 8

Calculating Distance Between (2, 10) and C3(1.5, 3.5)-
= |x2 – x1| + |y2 – y1|
= |1.5 – 2| + |3.5 – 10|
= 0.5 + 6.5
= 7

In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Table 3: Cluster allocation after 2nd iteration



Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (6, 6) of Cluster-02	Distance from center (1.5, 3.5) of Cluster-03	Point belongs to Cluster
(2, 10)	0	8	7	C1
(2, 5)	5	5	2	C3
(8, 4)	12	4	7	C2
(5, 8)	5	3	8	C2
(7, 5)	10	2	7	C2
(6, 4)	10	2	5	C2
(1, 2)	9	9	2	C3
(4, 9)	3	5	8	C1

From here, New clusters are-

Cluster 1:

First cluster contains points-

- (2, 10)
- (4, 9)

Cluster 2:

Second cluster contains points-

- (8, 4)
- (5, 8)
- (7, 5)
- (6, 4)

Cluster 3:

Third cluster contains points-

- (2, 5)
- (1, 2)

Now,

- We re-compute the new cluster clusters.
- The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster 1:

Center of Cluster-01
= ((2 + 4)/2, (10 + 9)/2)
= (3, 9.5)

For Cluster 2:

Center of Cluster-02
= ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4)
= (6.5, 5.25)

For Cluster 3:

Center of Cluster-03
= ((2 + 1)/2, (5 + 2)/2)
= (1.5, 3.5)
This is completion of Iteration 2.

After second iteration, the center of the three clusters are-

- C1(3, 9.5)
- C2(6.5, 5.25)
- C3(1.5, 3.5)

Next, we go to iteration 3, iteration 4 and so on until one of the conditions from the above algorithm is not met.

In above problem, after 4th iteration data points remains in the same cluster.

After 4th iteration, New clusters are-

Cluster 1:

- (2, 10)
- (5, 8)
- (4, 9)

Cluster 2:

- (8, 4)
- (7, 5)
- (6, 4)

Cluster 3:

- (2, 5)
- (1, 2)

After the 4th iteration, final clusters will be shown using graph. For that, clustering algorithm is executed in python with given dataset.

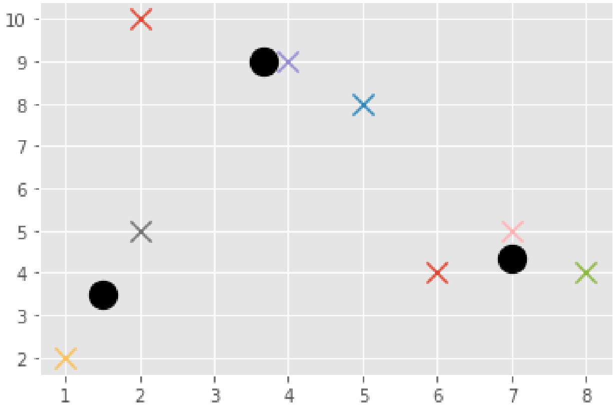


Fig 3: Final result of k-means clustering

The above figure(figure 3) shows the data points contained in three different clusters.

IV. ADVANTAGES AND DISADVANTAGES

- It is efficient with time complexity O(nkt) where-

n = number of instances

k = number of clusters
t = number of iterations

- It often terminates at local optimum.

Techniques such as Simulated Annealing or Genetic Algorithms may be used to find the global optimum.

K-Means Clustering Algorithm has the following disadvantages-

- It needs to specify the number of clusters (k) in advance.
- It can not handle noisy data and outliers.
- It is not suitable to identify clusters with non-convex shapes.

V. CONCLUSION

Clustering is a technique in data mining to group data which is not labeled based on similarity measure. We studied different techniques used in data mining that is association, classification and clustering. In this paper, we have seen different types of clustering algorithms and given algorithm for k-means clustering. Also, step by step solution of example of k-means clustering with graph is given. We have used Manhattan distance formula for calculation of distance between two points. Also, advantages and limitations of k-means clustering is mentioned.

REFERENCES

[1] P. Praveen and B. Rama, "An empirical comparison of Clustering using hierarchical methods and Kmeans," *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Chennai, 2016, pp. 445-449.doi:10.1109/AEEICB.2016.7538328.

[2] Zhang, S. and Chen, Z. The research of hilbert r-tree spatial index algorithm based on hybrid clustering. International Conference on Electronic and Mechanical Engineering and Information Technology, 2011, 3495-3497.

[3] Jiawei Han Micheline Kamber, Data Mining concepts and techniques, 2nd Edition.

[4] A.K.Jain, M.N. Murty, P.J. Flynn, Data clusterin: a review, ACM Computing Surveys 31 (3) (1999) 264-323.

[5] N.Yuruk, M.Mete, X.Xu, T. A. J. Schweiger, A divisive hierarchical structural clustering algorithm for networks, in: Proceedings of the 7th IEEE International Conference on Data Mining Workshops, 2007, pp.441-448.

[6] Fang, H. and Saad, Y. (2008). Farthest centroids divisive clustering. InProc. ICMLA, pages 232–238.

[7] Arun K Pujari, Data Mining Techniques, second edition.

[8] Musa J. Jafar, ―A Tools-Based Approach to Teaching Data Mining Methods‖, Journal of Information Technology Education, Volume9,2010.

[9] De Carvalho, F. A. T. and De Souza, R. M. C. R. (2010). Unsupervised pattern recognition models for mixed featuretype symbolic data. Pattern Recognition Let-ters, 31(5):430–443.

[10] P. Praveen, B. Rama and T. Sampath Kumar, "An efficient clustering algorithm of minimum Spanning Tree," *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Chennai,2017,pp.131-135.doi: 10.1109/AEEICB.2017.7972398.