International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

Two Phase Clustering Approach in Data Mining : A Review

Vashali[1], Shagun[2]
[1]M. Tech. Scholar, [2]Asstt. Professor
[1,2]Department of Computer Science & Engineering Manav Institutes of Technology & Management, Haryana, India

ABSTRACT

In clustering, dissimilarity is measured between gadgets by way of measuring the Euclidean distance among every pair of items. Analysis of cluster is a descriptive assignment that perceive homogenous group of objects and it is also one of the fundamental analytical method in facts mining. mean Shift clustering does now not depend upon a priori information of the wide variety of clusters while ok-way algorithm is an unmanaged clustering set of rules that classifies the enter records factors into multiple clusters k based on their inherent distance from each other. On this assessment, the outline of mean shift and okay-way algorithm is provided.

Keywords : Clustering, okay-manner, Euclidean distance, imply

I. INTRODUCTION

Clustering is the project of grouping a set of items in the sort of manner that items inside the equal group are more similar to each other than to those in other agencies. It's far the computational undertaking to partition a given enter into subsets of equal characteristics. these subsets are generally referred to as clusters. it's miles a primary project of exploratory facts mining, and a common method for statistical information analysis, used in many fields, which includes picture analysis, pattern reputation, system learning, facts retrieval and bioinformatics. Many variations of k-means clustering set of rules have been developed recently. [1]

Cluster evaluation is an iterated procedure of information discovery and it is a multivariate statistical technique which identifies groupings of the information objects based totally at the inter-object similarities computed via a designated distance metric. In clustering, we degree the dissimilarity among objects with the aid of measuring the space among each pair of objects. these measures include the Euclidean, ny and Minkowski distance [7]. Clustering algorithms can be classified into categories: Hierarchical clustering and Partitioned clustering [8]. The partitioned clustering algorithms, which differ from the hierarchical clustering algorithms, are normally to create a few units of clusters at start and partition the statistics into similar groups after every new release. Partitional clustering is extra used than hierarchical clustering because the dataset may be divided into more than two subgroups in a unmarried step however for hierarchy approach, always merge or divide into 2 subgroups, and don't need to finish the dendrogram[9].

Standard Clustering Algorithm

D. Małyszko[10], proposed the stairs for the clustering algorithm as follows:

(i) Initialize step with centers C.

(ii) For each statistics point xi, compute its minimum distance with every center cj.

(iii) For every center cj, recomputed the brand new center from all data factors xi belong to this cluster.

(iv)Repeat steps 2 and 3 till convergence.

Mean Shift Clustering Algorithm

Shift Clustering algorithm

The suggest shift manner turned into at first provided in 1975 with the aid of Fukunaga and Hostetler. Primary imply-shift clustering algorithms keep a fixed of data factors the same size as the input data set. to begin with, this set is copied from the enter set. Then this set is iteratively changed via the imply of these factors within the set which are inside a given distance of that factor.

suggest shift set of rules clusters an n-dimensional information units. For each factor imply shift computes its related peak through first defining a round window on the information point of radius r and computing the imply of points that lie inside the window. Algorithm then shifts the window to the suggest and repeats till convergence [11]. At each generation, the window will shift to a extra densely populated part of facts set till height is reached wherein data is equally disbursed.

An instance illustrating mean shift procedure is proven in Fig.1[11]. The shaded and black dots denote the facts points of an image and successive window centres respectively. imply shift system begins at factor Y1, by means of defining spherical window of radius r round it, set of rules then calculates the suggest of records factors that lie inside the window and shifts the window to the imply and iterates the equal method till height is reached. At every iteration, window is shifted to the more densely populated vicinity.

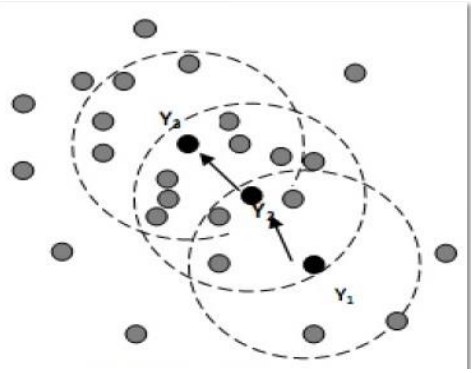

Fig.1. Mean Shift Procedure

K-mean clustering algorithm

-imply is a unsupervised, non-deterministic, numerical, iterative technique of clustering. It was first off proposed by means of Mac Queen in 1967. Here, a fixed of n item is partitioned into ok cluster in order that inter cluster similarity is low and intra cluster similarity is excessive. Similarity is measured in time period of suggest cost of objects in a cluster and assigns the gadgets to the closest cluster centre, such that the squared distances from the cluster are minimized. Okay centroids may be described, one for every cluster. these centroids should be positioned in such a way that cluster label of the images does now not exchange anymore.

Procedure of K-mean Algorithm

K-mean distributes all objects to K number of clusters at random;

1) Calculate the mean value of each cluster, and use this mean value to represent the cluster;

2) Re-distribute the objects to the closest cluster according to its distance to the cluster center.

3) Update the mean value of the cluster, say, calculate the mean value of the objects in each cluster.

4) Calculate the criterion function E, until the criterion function converges.

Usually, the K-mean algorithm criterion function adopts square error criterion, defined as:

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

In which, E is total square error of all the objects in the data cluster, p is given data

object, mi is mean value of cluster Ci (p and m are both multidimensional.

The function of this criterion is to make the generated cluster be as compacted and independent as possible[12].

Analysis of the Performance of the K-mean Algorithm

Advantages

1) It is a classic algorithm to resolve cluster problems; this algorithm is simple and fast.
2) For large data collection, this algorithm is relatively flexible and highly efficient, because the complexity is O (ntk), among which, n is the number of all objects, k is the number of cluster, t is the times of Iteration. Usually, k<<n and t<<n. The algorithm usually ends with local optimum.
3) It provides relatively good result for convex cluster.
4) Because of the limitation of the Euclidean distance[13], it can only process the numerical value, with good geometrical and statistic meaning.

Disadvantages

1) Sensitive to the selection of initial cluster centre, usually end without global optimal solution, but suboptimal solution.
2) There is no applicable evidence for the decision of the value of K (number of cluster to generate), and sensitive to initial value, for different initial value, there may be different clusters generated.
3) This algorithm is easy to be disturbed by abnormal points; a few of this abnormal data will cause extreme influence to the mean value.
4) Sometimes the result of cluster may lose balance.

Methods to improve k-means algorithm's performance

a) Methods for Initial Point Selection

i. Refining Initial Points Algorithm

In partitioned clustering algorithm, the first step we have to get preliminary seed points (cluster facilities). To pick excellent preliminary factors will improve solutions and decrease execution time. Refining initial factors set of rules is proposed [9].

For a start, randomly pick a few subsets within same number of samples from large data units. Secondly, partitioned algorithm is carried out to every subset to get each centre sets of the subsets. Thirdly, accumulate these middle sets and apply the partitioned algorithm once more to reap the maximum right middle set. for buying properly initial seed factors, definitely we repeat the partitioned algorithm 2 times through fewer sample units. Eventually run the partitioned algorithm with the most viable middle set as the preliminary seeds and authentic big facts units.

The algorithm steps are:

1) Randomly build J sample subsets. Si is a random subset of data (i = 1...J and the size of Si is S_size).
2) Use modified algorithm to find center Ci of each Si. Gather all Ci (i = 1...J) into C _Total .
3) For each set Ci (i = 1...J), run paritional algorithm with initial points Ci and data set C _Total to get another center set FCi .
4) For each FCi(i = 1...J), calculate sum Sumi of the distance between each point in C_Total to the closest center point in FCi .
5) Find minimum of Sumi (i =1...J). If Sum is minimum, take FCp as final initial points.

ii. Cluster Centroid Decision Method

This technique proposed a method to assign the data point to appropriate cluster's centroid, we calculate the space among every cluster's centroid and for every centroid take the minimum distance from the last centroid and make it 1/2, denoted with the aid of dc(i) i.e., half of the minimum distance from its cluster's centroid to the last cluster's centroid. Now take any facts factor to calculate its distance from its centroid and compare it with dc(i). If it is much less than or equal to dc(i) then facts point is assigned to the its cluster otherwise calculate the distance from the alternative centroid. Repeat this system till that statistics point is assigned to any of the closing cluster. [14].

N0: Number of data point
K: Number of cluster's centroid
Ci: its cluster
Some equations used in algorithm are:

$|Ci,Cj| = \{d(mi,mj) : (i,j)\varepsilon [1,k] \& i \neq j\}$

Where |Ci,Cj| is the distance between cluster Ci and Cj.

$dc(i) = \frac{1}{2}(\min\{|Ci,Cj|\}$

where dc(i) is the half of the minimum distance from ith cluster to any other remaining cluster.

iii. Cluster Seed Selection

When calculating the K turn of clustering seeds with the improved algorithm, those data in the cluster having a great similarity to the K-1 category seeds should be adopted to calculate their mean points (geometrical centre) as the clustering seed of the K tom and the specific calculation method is below as[15]:

1) For the cluster Ci(k-1) obtained through the K-1 tom of clustering, the minimum similarity sum _ mini(K-1) of the data in the cluster to the clustering seed Si(k-1) of the cluster is calculated;

2) The data in the cluster Ci(k-l) is calculated that has a similarity of more than 1-β* (1-sim _ mini(k-I))to the clustering seed Si(k-I) (among, β is a constant between 0-1), and the data set is recorded as cni(k-1).

3) The mean points of the data in cni(k-l) are calculated as the clustering seed of the K tom.

b) Methods to Define no of Clusters

i. Initialization Method

This method depends on the data and works well to find the best number of cluster and their centroids values. It starts by reading the data as 2D matrix, and then calculates the mean of the first frame.

Then, it keeps the value of means in an array called means array even at the end of the data matrix. After that it sorts the values in the means array in an ascending manner. In cases where the values are similar, they are removed to avoid an overlap. In other words, only one value is kept. It will then calculate the number of elements in the means array: this number is the number of clusters and their values are the centroids values as indicated in the steps below as[16]:

1) Read the data set as a matrix.

2) Calculate the means of each frame depending on the frame size and putting them in the means array.

3) Sort the means array in an ascending way.

4) Comparing between the current element and the next element in the means array. If they are equal, then keep the current element and remove the next, otherwise, keep both.

5) Repeat step 4 until the end of the means array.

6) Count how many elements remain in the means array. These are equal to the number of clusters and their values.

ii. The Encoding Method

According to the characteristics of K-mean cluster algorithm, to find the optimum cluster, the optimum K value should be found, the value of K is the learning object of the genetic algorithm, the encoding it encoding to K value. In general situation, to the class issue, there is always a maximum number of classes “MAXC Lassnum” for the cluster, this value is input by the user. So K is a integral between 1 and MAXC Lassnum, can be indicated in a binary string. In this experiment, using a byte to express K value, that is 255 classes maximum. This value is enough for normal cluster problem[15].

1) Chooses n number of chromosome from the original n chromosome Using the roulette wheels selection of the traditional genetic algorithm.

2) Crossover method is applied on selected Chromosome in the matting pool.

3) Mutation is applied over chromosomes in the mating pool.

4) Form a new generation of chromosome with the original chromosome.

5) Design the fitness function to evaluate K value by the quality of sample cluster result. The Fitness function is:

$$Fitness = \omega_1 \frac{Dis\ \ of\ \ class}{1 + Dis\ \ in\ \ class} + \omega_2 \frac{1}{NumDifference}$$

Distance between classes is:

$$Dis\ \ of\ \ class = \frac{2\sum_{i=0}^{k}\sum_{j=i+1}^{k} dis(center_i, center_j)}{k(k-1)}$$

The centeri is classic of cluster centre, dis(x,y)is the Euclidean distance between x,y.

Distance between classes is:

$$Dis\ \ in\ \ class = \frac{1}{k}\sum_{i=0}^{k}\left(\frac{\sum_{j=0}^{num_i} dis(Sample_{ij}, center_i)}{num_i}\right)$$

The num is number of class of i, sample is the sample j of the classic, Num Difference show the statistics of the difference of sample between classes. The ω1, ω1 of Fitness function is the weight of distance between classes.

iii. Tentative Clustering

Clustering uses principal components analysis, to determine a tentative value of count of classes and provide changeable labels for objects. The kernel based clustering approach performs principal component analysis on standard score of a given matrix and thereafter projects the matrix into space of the calculated principal vectors. Count of the employed principal vector is depending on the given number of classes (K-Means algorithm needs 'K' to process). In order to avoid dependence on the number of classes 'K' and to find maximum possible classes, we project the matrix to the space of all principal vectors. After we calculate a probability matrix (P) from result of projection (matrix C) such that Pi,j entry shows probability of connectivity of ith object to jth object. By refining matrix C according to the probability values of matrix P, we find a block matrix that represents groups of objects[15].

II. CONCLUSION

This paper presents an overview of the mean shift and k-means clustering algorithm. K-means clustering is a common way to define classes of jobs within a Dataset. Selection of good initial points will improve solutions and reduce execution time. Therefore the initial starting point selection may have a significant effect on the results of the algorithm. Methods to improve performance of k-means clustering fall into two categories: initial point selection and define number of cluster. Six of these methods, three from each category, are presented in this review. These methods have been implemented in data mining system and can get better results for some practical programs such as character recognition, image processing and text searching.

III. REFERENCES

[1]. F Gibou and R. Fedkiw, "A Fast Hybrid k-Means Level Set Algorithm for Segmentation". 4th Annual Hawaii International Conference on Statistics and Mathematics 2002, pp. 1-11; 2004.

[2]. K Sravya and S. Vaseem Akram, "Medical Image Segmentation by using the Pillar K-means Algorithm", International Journal of Advanced Engineering Technologies Vol. 1, Issue1; 2013.

[3]. G Pradeepini and S. Jyothi, "An improved k-means clustering algorithm with refined initial centroids". Publications of Problems and Application in Engineering Research. Vol 04, Special Issue 01; 2013.

[4]. G Frahling and C. Sohler, "A fast k-means implementation using coresets". International Journal of Computer Geometry and Applications 18(6): pp. 605-625; 2008.

[5]. C Elkan, "Using the Triangle Inequality to Accelerate k-means", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, pp. 147-153; 2003.

[6]. K Wagsta, C. Cardie, S. Rogers and S. Schroedl, "Constrained K-means Clustering with Background Knowledge", Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577-584; 2001.

[7]. J. Yadav and M. Sharma, "A Review of K-mean Algorithm". International Journal of Engineering Trends and Technology. Volume 4, Issue 7, pp. 2972-2976; 2013.

[8]. J. Han and M. Kamber, "Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, CA, 2001.

[9]. J. Gu; J. Zhou and X. Chen, "An Enhancement of K-means Clustering Algorithm". In proceedings of Business Intelligence and Financial Engineering (BIFE '09); 2009.

[10]. D. Mal yszko and S. T. Wierzchon "Standard and Genetic K-means Clustering Techniques in Image Segmentation", (CISIM'07) 0-7695-2894-5/07 IEEE 2007.

[11]. Sumant V. Joshi and Atul. N. Shire, "A Review of an enhanced algorithm for color Image Segmentation", International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 3, pp. 435-440; 2013.

[12]. J. Wang and X. Su; "An improved K-means Clustering Algorithm". In proceedings of 3rd International Conference on Communication Software and Networks (ICCSN), 2011.

[13]. C. Kearney and Andrew J. Patton, "Information gain ranking". Financial Review, 41, pp. 29-48; 2000.

[14]. R.V. Singh and M.P.S. Bhatia, "Data Clustering with Modified K-means Algorithm". In proceeding of International Conference on Recent Trends in Information Technology (ICRTIT), 2011.

[15]. Li Xinwu, "Research on Text Clustering Algorithm Based on Improved K-means". In proceedings of Computer Design and Applications (ICCDA), Vol. 4, pp.: V4-573 - V4-576; 2010.