Journal of Engineering and Applied Sciences 12 (21): 5485-5489, 2017
ISSN: 1816-949X

Efficient Data Clustering Algorithm Designed for 2-D Dataset

Himanika, Vishesh Mehta and Poonam Nandal
Department of Computer Science and Engineering, Faculty of Engineering and Technology,
Manav Rachna International University, Fridabad, Haryana, India

Abstract: Extraction of information from a database is a major issue these days. There is huge amount of information available in web in the form of web pages which is used to extract as per the need of the user to perform a vital task. To overcome this issue of information retrieval various techniques are known today like clustering, classification, natural language processing techniques etc. In this study, we have discussed various clustering methods algorithms with various features to classify the data. k-means clustering algorithm is majorly used to cluster the data which is also focussed in this study. The capability of k-means clustering algorithm is due to its computational competence. k-means is a clustering technique in which similar data points are grouped into clusters. In this study, we have proposed a clustering algorithm based on the density of data points and used Manhattan distance for grouping the data points into a cluster. It has been empirically found that the results of proposed clustering algorithm provide better clusters as compared to existing clustering algorithms.

Key words: Clustering, partition, hierarchical, agglomerative, divisive, k-means

INTRODUCTION

In general, the process of analyse/examine the relevant data from different extents is called data mining. It is an interdisciplinary subfield of computer science which focuses on the retrieval of relevant information from the large databases. It also focuses the change of retrieved data into understandable and readable manner. It is a basic process of exploring the data from different areas to use it for various different purposes like data managing in big areas.

Data classification techniques can be supervised or unsupervised. One of the most widely used techniques is to classify the data set into different clusters commonly referred to as clustering. Clustering is grouping of similar data or information. Clustering can be categorized as follows: hierarchical, spectral clustering, grid based, density based and partitioning based clustering (Maulik and Bandyopadhyay, 2002).

Hierarchical clustering algorithm data is classified in the form of a tree. It is further categorized in two types, i.e., agglomerative and divisive. Spectral clustering algorithm is an algorithm in which data points are portioned by means of similarity matrix. This research in three stages, i.e., pre-processing which focus on building of similarity matrix, construction of Eigen vectors which is done by spectral mapping, post processing which deals with grouping data points.

Grid based clustering algorithm is an algorithm in which operations are done on grids and that grids are formed by the objects space. The major advantage of this algorithm is that this does not need the computation of distance and further the clustering is done based on the obtained summarized data points, the complexity of this algorithm is (O).

Density based clustering algorithm in an algorithm in which a cluster is continuously growing till the density in the region surpasses the threshold. It requires only a single scan of the input data sets and parameters associated with density which are to be initialized.

Elavarasi *et al*. (2011) partitioning clustering algorithms divide the data points into k partition and each constructed partition signifies a cluster. It has two properties each group should contain an object and each object should belong to one group. Partitioning clustering is also known as non-hierarchical clustering as every instance is positioned in precisely one of k commonly exclusive clusters. In this clustering, the user needs to input the preferred count of clusters k as only a single set of cluster is the outcome of a typical partitioned clustering algorithm. One of the utmost

Corresponding Author: Himanika, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Manav Rachna International University, Fridabad, Haryana, India

frequently used partitioned clustering is k-means. As discussed, in this clustering the user needs to give the count of clusters (k) and from computation point of view the algorithm initiates the centres also called centroids of the k partitions.

In this study, we focussed on k-means clustering which can be applied on data like numeric, categorical and mixed data (Lim *et al*., 2012). In general, k-means clustering data in divided into different clustering by selecting centroids for clusters. Initially algorithm takes two inputs, i.e., the dataset having n number of objects and k number of clusters that are going to be created. Firstly, the centroids are selected randomly and then data points which were input are allocated to clusters by measuring the Euclidean distance. Next, when all the available input data points are allocated to some number of clusters, the first iteration is executed and the steps are repeated until the desired objective function is attained. The computationally time complexity of this is (nkl) where n is the input of data points, k be the count of clusters and l are the number of iterations needs to be performed. In the proposed algorithm, an approach to systematically selecting the initial centroids has been proposed. In this initially, the given data points are plotted in 2D. All the data points should have positive values and if negative value then first is converted into positive value this is necessary because distance is calculated from the origin.

LITERATURE REVIEW

Goyal and Kumar (2014) have given an algorithm in which centroids are selected randomly and they used Euclidean distance metric to assign data points to the random clusters. After assignment of all the input data points to some cluster first iteration is completed and then the same process will start in clusters too and this process needs 3 iterations. This is computationally expensive its time complexity is very high in terms of input dataset, clusters and iterations and result for this algorithm depends on the input it can vary for multiple runs.

Arockiam *et al*. (2012) explained the concepts of clustering using Hierarchical method which is further divided in to two parts Agglomerative and Divisive algorithms; Partitioning Methods which is further divided into 4 parts relocation, probabilistic, k-means, density-based which is further divided into two parts connectivity and functions clustering. k-mean is part of partitioning clustering which partitions a data set into a cluster. This is how a k-means help us to divide input dataset into clusters. It considers the input of clusters to group data into and the dataset. It constructs the k initial clusters from the dataset by selecting k rows of datasets randomly. For example, if there exists 10,000 number of rows in the dataset, then for the first step k = 3 initial clusters will be constructed. Each of these three initial clusters designed will consists of one row of datasets.

Vij and Kumar (2012) has proposed a 2D algorithm in which centroids is not selected randomly. It was basically improved k-means algorithm in which initial centroids are selected using the researchers proposed algorithm. When the data sets containing the negative values then firstly that negative value is converted to positive. Next, the minimum values are computed for all x and y-axis. So, this will make all the data points have positive value now these values form the boundary of rectangle which is divided into k clusters. After selecting the centroids distance of each centroid is computed in comparison to each centroid.

MacQueen in 1967 proposed the k-means algorithm which is well known method for clustering. However, the research illustrated by various indicates the result of k-means is quite sensitive to initial selection of random centres. When the centre closes to the final solution, it efficiently assigns the data to the appropriate cluster centre. Otherwise, k-means will get incorrect clustering results and have weak performances. After that further many methods have been proposed to deal with cluster initialization for k-means. After comparing many of the methods (Lan *et al*., 2015).

Shehroz and Ahmad introduced a Cluster Center Initialization (CCIA) for k-means. The basic idea of CCIA is observing that some of the pattern are very similar to each other and that is why they have same cluster membership which makes it independent on initial cluster centres. In this study, the authors proposed an algorithm for centres initialization for k-means based on density peaks (CIDP).

Elavarasi *et al*. (2011) proposed the review on the partition clustering algorithms. In this study, the researchers describes the operational performance, the procedures to be monitored and the restrictions which affects the enactment of the algorithm. The authors also discussed the various different types of clustering algorithms.

Various researchers have given different approaches for clustering. Although, they have covered many applications but there are some issues that are still

need to be challenged. In our proposed algorithm we will give an algorithm for clustering the data set.

K-MEANS CLUSTERING

As discussed above the k-means clustering needs two inputs the data set and the count of clusters required. This helps us to classify n number of data points into k clusters. Similarity of clusters is known by measuring the Euclidean distance between the objects. Various distant metrics can also be used along with Euclidean distance like Manhattan distant metric, Minkowski distance metric, Mahalanobis metric, etc. Now take the mean value of clusters as centre of gravity. Firstly, the centroids are selected randomly as the centre of cluster and every data point is allocated to a given cluster by computing the Euclidean distance for considering the computational efficiency. When all the input data points are assigned to some clusters then first iteration is done. Then algorithm starts new iteration and then again we find the new centroids and finally a situation will come when the algorithm will attain its objective function due to which the centroids or the data point do not change their cluster which illustrates the convergence measure for clustering. The k-means approach is given in Algorithm 1 as:

Algorithm 1:

Input: ν = ν1, ν2, ν3, ...,νn
Output: k = The count of preferred clusters
Method:-

i ¥ each data point compute Euclidean distance as:

$$dis((x, v)), (a, b) = \sqrt{(x-a)^2+(y-b)^2}$$

ii Compute mean (η) till the convergence is met

This algorithm is easy to implement on large datasets but it has some limitations too. This algorithm is applicable only for numeric data only this can't be applied for categorical data.

The above k-means clustering is done by using the Euclidean distance and as discussed we can use multiple metric for computing the distance k-means like Manhattan, Minkowski, etc. (Sinwar and Kaushik, 2014).

K-means using Manhattan distance metric: Manhattan distance is used for calculating the complete difference between the two points as distance xy = $|x_1-x_2|+|y_1-y_2|$ (Algorithm 2).

Algorithm 2:

Input set of data points v and c clusters ν = ν1, ν2, ν3, ..., νn//data points and c = c1, c2, c3, ..., //clusters

i ¥ every data point and selected centroid compute Manhattan distance as:

$$xy = |x_1-x_2|+|y_1-y_2|$$

ii Calculate new centre using the formula:

$$\left(\frac{1}{ci}\right)\sum_1^{cl} -x_1$$

iii Re-compute distance using Manhattan between new cluster and each data point

iv Repeat until (¥ data points)-> cluster (ci)

K-means using Minkowski distance metric: The Algorithm 3 describes the steps followed by k-means using Minkowski distance metric:

Algorithm 3:

Input set of data points and clusters. ν = ν1, ν2, ν3, ..., νn //data points and c = c1, c2, c3, ..., cn //clusters

i Choose cluster centre 'v_i' randomly

ii ¥ data point and selected centroid Compute Minkowski distance as:

$$dist\ (xy) = x_{ik}-x_{jk}$$

iii Calculate new centre using the formula:

$$\left(\frac{1}{ci}\right)\sum_1^{ci} x1$$

iv Assign data point with min (dist (xy) to the cluster

v Re-compute the distance between each available data point and the newly created cluster

In conclusion by analysing the results of different distance metric, it is noted that k-means is done using the Euclidean distance because it gives the most efficient result and moreover, it is space oriented result for k-means using Manhattan and Euclidean is almost same it's just that Manhattan gives the more distortion (Singh *et al*., 2013). In next subsection, we will discuss the algorithm related for two-dimensional datasets.

ALGORITHM FOR 2D DATA

In 2D clustering algorithm, 2 dimensional dataset is taken as input like numeric data having both positive and negative values. If the data sets containing the negative values then firstly that negative value is converted to positive value so that all input data points lie on the identical plane. The min value for x-axis will be x_{min} and min value for y-axis will be y_{min}. Then all the data points from the data sets are subtracted from the minimum values. Now, all the data points have positive value now these values form the boundary of rectangle which is divided into k clusters. After selecting the centre data point distance of each data point is computed with respect to each centroid. Steps for improved 2D algorithm are given in Algorithm 4 as follows:

Table 1: Sample 2D dataset

Data points	X	Y	Computed X	Computed Y
D_1	5	2	10	13
D_2	5	-3	5	13
D_3	4	3	11	12
D_4	4	-4	4	12
D_5	-5	-3	5	3
D_6	-4	6	14	4
D_7	7	5	13	15
D_8	3	-2	11	6
D_9	-4	-2	4	6
D_{10}	4	6	12	14
D_{11}	-6	8	2	16
D_{12}	-5	-8	3	0
D_{13}	-8	-3	0	5
D_{14}	1	6	9	14
D_{15}	1	-6	9	2

Table 2: Comparative analysis of sample dataset

Clustering algorithm	Count of iterations
k-means	3
2D algorithm for k-means	2

Algorithm 4:

Input: P = {p_1, p_2, p_3, ..., p_n} where P are 2D points
Output: M = {M_1, M_2, ..., M_n} where M are the formed clusters

1. If ¢ (+ve, -ve) data points in input data set then go to step 2 else go to step 3
2. Compute x_{min} ¥x and y_{min} ¥y cordinate
3. Subtract each data point with the minimum attribute value as obtained in step 2
4. Obtain the rectangle boundary values: compute min(x), max(x), min(y), max(y)
5. Next, construct cluster from the rectangle by dividing it into k parts such that each part represents a cluster
6. ¥ each cluster -> assign data points: compute dis (ə data point). //data points assigned based on distance computed
7. Repeat
8. ¥ data point computed di» centroid cj Π each cluster j -> assign data points.//for each cluster data points are assigned
9. Set cluster d[i] = j
10. Set Dist[i] = d(diΠcj)
11. ¥ each cluster d_i (I< = j < = k), re-compute the centroids
12. If dis <= present nearest distance ,then data point ə same cluster//dis is the distance computed
13. Else
14. ¥ centroid cj(1< = j< = k) compute the dis(diΠcj) //distance computed between di and ci
15. end ¥
16. Repeat until the objective function is met

Illustrative example of 2D algorithm: Table 1, we have taken a sample dataset with 15 points and showing their x and y-axis also x and y computed, respectively. Firstly find the min value on x and y-axis x_{min} ¥x and y_{min} ¥y subtract ¥ data points as shown in computed values of Table 1. Now find the minimum and maximum from this new table x_{min} = 0 x_{max} = 14 y_{min} = 0, y_{max} = 16. Boundary values (14.0) and (0.16).

Now form a rectangle and divide it in to 4 parts with 4 centroids (R_1, R_2, R_3, R_4) R_1 = (3.5, 12), R_2 = (10.5, 12), R_3 = (3.5, 4), R_4 = (10.5, 4) Following are the iterations as:

Iteration 1: R_1->D_2, D_4, D_{11}, R_2->D_1, D_3, D_7, D_{10}, D_{14}, R_3->D_5, D_9, D_{12}, D_{13}, R_4->D_6, D_8 ,D15.

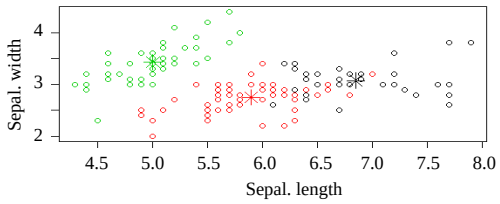

Fig. 1: K-means clustering

Iteration 2: R_1->D_2, D_4, D_{11}, R_2->D_1, D_3, D_7, D_{10}, D_{14}, R_3->D_5, D_9, D_{12}, D_{13}, R_4->D_6, D_8, D_{15}.

Next in Table 2, we provide the comparative analysis of k-means, algorithm for the sample dataset taken in Table 1. Table 2, we give the number of iterations respective to each of the algorithm. It illustrates that that 2D algorithm used for clustering is more efficiently computed to select initial centroid for assigning data points to each of the constructed cluster.

The k-means also stemmed into precise output but it is complex in terms of space and time. In comparison to this existing our proposed algorithm the initial centroids are not considered randomly but computed so that, we move to right direction. The cluster formation of the abovek -means on the sample data set is also shown in Fig. 1.

PROPOSED ALGORITHM

In this study, we will discuss an approach for dividing the 2D data points based on the Manhattan distant metric and systematic selection of centroids using density. The centroids are selected based on density so that different runs on algorithm on the same dataset produce the good quality result. As discussed earlier, to divide 2D clusters we have to convert all data points to some positive value but in proposed algorithm there is no need to take the positive values. We are using Manhattan distance because it gives the low distortion as compared to Euclidean distance. Improved efficient 2D algorithm is given in Algorithm 5 as follows:

Algorithm 5:

Algorithm 5: Input: D = {d_1, d_2, ..., d_n}
Output: C = {C_1, C_2, ..., C_n}
Steps:

i Find the density denoted by a symbol k of each data point as $\delta = x_i/Y_i+0.01\times\alpha$
ii Data points with max(δ)¥P are selected as centroids
iii Compute Manhattan distance ə(data point and cluster centroid) as:

$$dis(xy) = |x_1-x_2|+|y_1-y_2|$$

iv Data point with min (dis(xy) » cluster c_i
v Continue ¥ data point » any cluster c_n

Illustrative example: Suppose data points taken are P_1 (2, 5), P_2(-3, 5), P_3 (3,4), P_4 (-4,4), P_5(-3,5), P_6(6,-2)

i density of each data point by the formula given above P1 = 0.49, P2 = 1.98, P3 = 0.49, P4 = 0, P5 = 3.3

ii Points with highest density as centroids using the formula $\delta X_i/(Y_i+0.01)\times\alpha$ where α is the constant with value 0.5 and δ is the density, i.e., p5, p6

iii Compute dis(xy) for each data point w.r.t centroids (xy) = $|x_1-x_2|+|y_1-y_2|$

Distances from P_5: $P_1 = 15$, $P_2 = 10$, $P_3 = 15$, $P_4 = 16$ and distances from P_6: $P_1 = 13$, $P_2 = 18$, $P_3 = 9$, $P_4 = 18$

iv Now data points will fall into cluster to which they are closer: P2, P4 will fall to cluster P5 and P1, P3 will fall into cluster P6. First iteration will get completed here

v Same points will get repeated now in the clusters P5 and P6

CONCLUSION

In this study, review of clustering tools and techniques is presented and an algorithm for k-means clustering for 2D data set is proposed. The data is divided based on the density to make it more efficient as compared to the other existing algorithms. We have found that after using the density concept, there is no need to differentiate the positive and the negative data points and also Manhattan distance metric gives low distortion value as compared to Euclidean. The major drawback for k-means algorithm for 2D data set was to firstly differentiate the positive and negative points then start clustering. In the proposed research there is no need to differentiate between the positive and negative points as the initial centroids are chosen according to the density of the data points.

ACKNOWLEDGEMENTS

Researchers would like to express the gratitude to Dr. Kiran Khatter, Research Mentor, Accendere Knowledge Management Services Pvt. Ltd. and other Research Mentors from Accendere KMS Pvt. Ltd. for their comments on earlier versions of the manuscript. Although, any errors are our own and should not tarnish the reputations of these esteemed persons.

REFERENCES

Arockiam, L., S.S. Baskar and L. Jeyasimman, 2012. Clustering techniques in data mining. Asian J. Inf. Technol., 11: 40-44.

Elavarasi, S.A., J. Akilandeswari and B. Sathiyabhama, 2011. A survey on partition clustering algorithms. Intl. J. Enterp. Comput. Bus. Syst., 1: 1-14.

Goyal, M. and S. Kumar, 2014. Improving the initial centroids of K-means clustering algorithm to generalize its applicability. J. Inst. Eng. Ser. B., 95: 345-350.

Lan, X., Q. Li and Y. Zheng, 2015. Density k-means: A new algorithm for centers initialization for k-means. Proceedings of the 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), September 23-25, 2015, IEEE, Beijing, China, ISBN:978-1-4799-8352-0, pp: 958-961.

Lim, J., J. Jun, S.H. Kim and D. McLeod, 2012. A framework for clustering mixed attribute type datasets. Proceedings of the 4th International Conference on Emerging Databases (EDB12), August 23-25, 2012, Seoul Selection Publishing, Seoul, Korea, pp: 92-100.

Maulik, U. and S. Bandyopadhyay, 2002. Performance evaluation of some clustering algorithms and validity indices. IEEE. Trans. Pattern Anal. Mach. Intell., 24: 1650-1654.

Singh, A., A. Yadav and A. Rana, 2013. K-means with three different distance metrics. Intl. J. Comput. Appl., 67: 13-17.

Sinwar, D. and R. Kaushik, 2014. Study of Euclidean and Manhattan distance metrics using simple K-means clustering. Intl. J. Res. Appl. Sci. Eng. Technol., 2: 270-274.

Vij, R. and S. Kumar, 2012. Improved k-means clustering algorithm for two dimensional data. Proceedings of the 2nd International Conference on Computational Science, Engineering and Information Technology, October 26-28, 2012, ACM, Coimbatore UNK, India, ISBN:978-1-4503-1310-0, pp: 665-670.