International Journal of Engineering & Technology, 7 (2.32) (2018) 111-113

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

An Efficient Density Based Clustering approach for High Dimensional Data

[1]Y. Vijay Bhaskhar Reddy PP COMP.SCI.0560, [2]Dr L.S.S Reddy, [3]Dr.S.S.N. Reddy

[1]Research Scholar, Rayalaseema University, Kurnool,AP.
[2]Vice Chancellor, KL University, Vaddeswaram.
[3]VPrincipal, VCE, Hyd, TS.
**Corresponding author E-mail: vjy.reddyjnr@gmail.com*

Abstract

Data extraction, data processing, pattern mining and clustering are the important features in data mining. The extraction of data and formation of interesting patterns from huge datasets can be used in prediction and decision making for further analysis. This improves, the need for efficient and effective analysis methods to make use of this data. Clustering is one important technique in data mining. In clustering a set of items are divided into several clusters where inter-cluster similarity is minimized and intra-cluster similarity is maximized. Clustering techniques are easy to identify of class in large databases. However, the application to large databases rises the following requirements for clustering techniques: minimal requirements of domain knowledge to determine the input specifications, invention of clusters with absolute shape & certainty of large databases.. The existing clustering techniques offer no solution to the combination of requirements. The proposed clustering technique DBSCAN using KNN relying on a density-based notion of clusters which is accomplished to discover clusters of arbitrary shape.

Keywords: *Clustering, DBSCAN, KNN, Arbitrary Shape.*

1. Introduction

Processing of huge data is very complicated task in the present world. Many users want to store or represent the huge information as data. The important task of the data is to divide or classify the facts into a set of groups. Clustering is the approach in data mining to identify the similar objects and make as one group. There are many issues in clusters to find the group of similar objects of data. In clustering similarity is most widely used to find the similar objects by using a similarity function. Clustering and classification are the two techniques in data mining and there is a difference between these two. In this paper, proposed clustering technique DBSCAN using KNN relying on a density-based notion of clusters which is designed to discover clusters of arbitrary shape. The proposed system, increase the performance of the clustering and this reduces the computation time of the processing of data sets.

In this paper, the proposed system is based on the notion of density. If the clusters are increasing day by day the neighbor clusters are also increased some threshold. Every data point in the present cluster should contain at least the minimum number of points within.

In clustering it is very important to know that the formation of similar data points as a cluster should be meaningful.The proposed density based clustering accepts the finding of arbitrary shape and no need of convex areas of data points that are more generated. Density based clustering does not need the number of clusters beforehand,but relies on a density-based notion of clusters such that for each point of a cluster the neighborhood of a given radius (ε) has to contain at least a minimum number of points (℘). However, finding the correct parameters for standard density based clustering [1] is more of an art than science.

Various algorithms are there for clustering. Many of the clustering algorithms require that the quantity of groups is to be known proceeding the begin of clustering process others decide the clusters themselves as a rule Density-based clustering algorithms are free of earlier learning of a number of a clusters. Such algorithms might be helpful in circumstances where the quantity of cluster is to be resolved effortlessly before the begin of the algorithm.

2. Related Work

The performance of K-means clustering is shown in [2]. This is divided into two phases. In the first phase the basic controls and defined in the better way to generate the clusters with high accuracy. In the second phase the assigning of better data points to the clusters.

To minimize the time complexity of the of k-means clustering techniques the uniform distribution of data points is explained in [3]. This approach is used to reduce the runtime and it generates the better quality of the cluster. To find the basic centroid a better method is adopted to this approach. One more method integrated in this approach is to calculate the distance between every data point.

This is the approach used to reduce the number of iterations from k-mean algorithm and to improve the runtime and reduce the total number of distance calculations [4]. The iteration method enhances the K-means clustering and starts with good point and select the points randomly. This will generate, the better cluster.

It is a most efficient function to assign the data points to clusters. The traditional k-means algorithm calculates the distance between all the data points for each centroid and also the calculation time is also very expensive due to their iterations [5]. This approach will use two distance functions to improve the performance- one same as k-means algorithm and the other is to reduce the number of distance calculations. The issue in this approach is the basic centroids are mentioned randomly, as implemented in k-means algorithm. Thus we can say that for the final clusters there is no guarantee.

To find the accurate and systemic centroids this method is proposed [6]. The centroids formed by using this approach are constant and data are distributed. This will generate the efficient and accurate clusters compare with traditional k-means algorithm. This method not adopted any method for improving the time complexity of the k-means algorithm.

3. KNN:

In data mining, classification and regression are two non-parametric metrics in the k-nearest neighbor algorithm (k-NN) using pattern recognition. At this time the dataset consists of the k-nearest training samples. In this paper, the K-NN is used for classification.

• K-NN is meant for classification and the output is classed membership. This classification is done on objects based on the huge voting of its similarities. Based on the values assigned in k nearest neighbors for example, if k=1, the item is integrated with the single nearest neighbor.

The property value of the object output in K-NN regression. For the k nearest neighbors the value is average of all the values.

The issues in the K-NN are time taking learning or case based learning where the capacity is just approximated locally and all calculation is conceded until order. The k-NN calculation is among the least complex of all machine learning calculations.

In K-NN there are two methods, regression and classification. So that the closer neighbors contribute more to the normal than the more far off ones. For instance, a typical weighting plan comprises in giving each neighbor a weight of 1/s, where s is the separation to the neighbor.

4. Proposed System

The proposed clustering algorithm DBSCAN adopted with KNN based on density-based notion of clusters. The aim of the proposed work is to generate clusters of arbitrary shape. Finding the cluster is the biggest task with DBSCAN. This starts with arbitrary points and gets back all the density reachable points us to Eps and Min pts. If s is basic point this process gets the cluster. If s is the end point there are no points are dense-reachable from us and DBSCAN visits the other point of the database. Hence the global values are used for Eps and Min Pts, proposed algorithm joins the two clusters in as one in this case. The other two clusters are with different density and that is close. To calculate the distance we have two points’ p1 and p2 and initialize as

dist (p1, px)= min {dist(x,y) p b q C$2}. These sets of points have the lowest density of the cluster and This will separate the cluster from each other based on the distance result between the two sets is larger than Eps. Higher values of Min points for detecting the clusters by using DBSCAN. Thus, we can say that there is no drawback to the DBSCAN which gets the name as efficient algorithm.

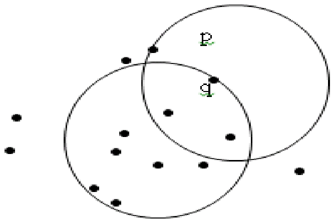
5. DBSCAN: Density Based Clustering:

- The other name of the DBSCAN is density-based algorithm.
- There are two points in this minimum point (Min Pts) and epsilon (Eps). The arbitrary entry point is the starting point. This will
- Finds the nearest points within the distance of Eps of the entry point.
- To form a cluster the no. Of nearest points are higher than or same to Min Pts. All the entry points and its nearest points are merged to this cluster and the entry point is marked as checked.
- The noise is identified based on the no of nearest points is then Min pts.
- Thus the algorithm proceeds for the next looping for finding the unchecked points in the data sets.

6. DBSCAN Algorithm

1. A Graph is created for the points to be clustered.
2. Create an edge for each core-point c to every point in the ε-nearest of c.
3. Set N to the points of the graph.
4. If N does not have any basic points terminate.
5. Select a basic point in N.
6. Let A be the no of nodes that can be reached from s to move forward.

   a. Create a cluster containing X∪U {s}
   b. N=N/(A ∪U {s})
7. Continue with step 4



8. Min Pts: Minimum number of points in any cluster
9. ϵϵ: For each point in the cluster there must be another point in its less than this distance away.
10. ε-Neighbourhood:Points within ε distance of appoint
11. N ϵϵ(p) :{q belongs to D |dist(p,q) <=ϵϵ)
12. Core point: ϵϵNeighbourhood dense enough (Min Pts)
13. Conditions: p belongs to N ϵϵ(q)
    |Nϵϵ(q)| > = Min Pts

7. Comparison

The main features of the proposed algorithm are as follows. Table 1 shows the comparison between existing system RDBC and proposed system DBSCAN with K-NN is implemented in R programming .

- Given parameters.
- Algorithm supports all the data types.
- Clusters shape.
- It solves the noise issues.
- Results of clustering.

The clustering research is done on three datasets iris, cancer, and synthetic shows the comparison between existing system and proposed systems. They are compatible to handle arbitrary shaped collections of points. Artificial data sets were generated from a multivariate normal distribution, whose mean vector and variance of each variable. The Iris flower data set is also a multivariate data set introduced by Sir Ronald Aylmer Fisher. Iris dataset is often referenced in the field of pattern recognition. The same data sets are given as input to the standard RDBC algorithm and the DBSCAN with KNN algorithm. The percentage accuracy and the time taken for each data set are computed and the values are tabulated.

Table 1: Comparison table for RDBC and proposed system on various data sets

Data Sets	RDBC		Proposed System	
	Time Required (in sec)	Accuracy	Time Required (in sec)	Accuracy
Iris	0.7865	77%	0.2432	83.1%
Cancer	0.8765	79.7%	0.5435	91.6%
Synthetic	0.872	71.5%	0.675	87.4%

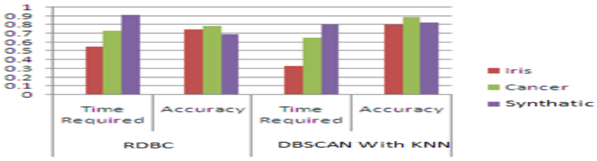


Figure 1: Comparison chart for above table

8. Conclusion

In this paper, the proposed system shows the performance of existing system RDBC and proposed DBSCAN with K-NN. The proposed system starts calculation form arbitrary point and find out the accurate nearest cluster points then well-known algorithm RDBC. For RDBC it has got the constant values of ε and Min Pts and it leads to one, for huge dataset it is not compatible. The proposed DBSCAN with K-NN that attempts to solve the problem of huge clusters by varying ε and Min Pts whenever required.

References

[1] Athman Bouguettaya "On Line Clustering", IEEE Transaction on Knowledge and Data Engineering Volume 8, No. 2, April 1996

[2] K. A. Abdul Nazeer & M. P. Sebastian" Improving the Accuracy and Efficiency of the K-Means Clustering Algorithm" .Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, London, U.K, July 1 - 3, 2009.

[3] D. Napoleon & P. Ganga lakshmi, "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points", IEEE, 2010.

[4] Madhuri A. Dalal & Nareshkumar D. Harale "An Iterative Improved k-means Clustering" Proc. of Int. Conf. on Advances in Computer Engineering, 2011.

[5] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, "An Efficient enhanced k-means clustering algorithm," Journal of ZhejiangUniversity, 10(7):1626–1633, 2006.

[6] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pages 26–29, August 2004.