Reduce Noise in K-Mean Clustering using DBSCAN Algorithm

Manjur Ahammad
Dept. of Computer Science & Engineering
United International University
Dhaka- 1212, Bangladesh

Faija Juhin
Dept. of Computer Science & Engineering
United International University
Dhaka- 1212, Bangladesh

Dewan Md. Farid
Dept. of Computer Science & Engineering
United International University
Dhaka- 1212, Bangladesh

ABSTRACT

The growth of data mining procedure is increasing day by day. We can extract useful insight from data. For mining data different techniques and tools have been introduced every day. By gaining knowledge from those insight of the data many research paper is being written. Based on the behavior, pattern and the characteristics data are being clustered into different groups. For clustering these massive amount of data we use different types of algorithms and techniques. The most common types of algorithms that are used in clustering are partitioning, hierarchical, grid-based and model-based algorithms. To handle these data another, type of algorithms are K-means clustering, density-based algorithm, similarity-based algorithms etc. the agenda off these algorithms are different. Some performs well for nominal data, some for categorical or ordinal data, contrariwise some can remove duplicate or noisy data and some can't do so. In this paper a method has been showed that how can we cluster a dataset and remove the noisiness of that particular dataset at the same time.

General Terms

Algorithm

Keywords

Big Data, K-Means Clustering, DBSCAN, OPTICS

1. INTRODUCTION

Data mining is a procedure of removing and noticing patterns in large data sets including methods at the intersection of machine learning, statistics, and database systems. There are few methods which are not useful for dataset. The high streaming data that are produced every minute by the applications of Internet of Things (IoT) is playing a vital role in everyone's life. From these tremendous amount of data high quantity of pattern, behavior and characteristics can be extracted. [1], [2]. The capacity of storing these huge amount of data is still a challenging task as well as time consuming. Sometimes these data are being called networking data because of the dependency with one data to another. There are various aspects such as biotechnology, machine learning, IoT and other sources where these data can be used. [3]. Basically these clustering formula handle these large amount of data. Based on the similarity and dissimilarity of data, it's been clustered for analyzing and for finding the hidden behavior of the data [4]. For example, when someone researches on marketing for a particular product, that individual will be able to gain some information about how frequently consumers are buying that particular product and the if the product is capable enough to hold the popularity among the buyers and that's how these hidden pattern help people in taking decisions in their business. There are various types of clustering algorithms exist such as K-means clustering, Similarity-based clustering, Density-based clustering, Distanced-based clustering, Hierarchical clustering etc. K-means clustering is one of the most popular algorithms among all these clustering methods.

K-means clustering algorithm calculates the centroids and repeats until we it catches optimal centroid. The data points are allocated to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum, in this algorithm. However, K-means clustering is time consuming because of the computational steps of the algorithm. On the other hand, it's impossible to remove duplicate data or noisy data in K-means clustering technique. In this paper authors are going to share an idea of clustering a dataset and removing noisy data at the same time in K-means clustering.

The paper is arranged as: the second section signifies the challenges of Big Data clustering. A view of the various clustering techniques has been provided in the third section. Fourth section demonstrates the proposed idea. Finally, fifth section concludes the paper.

2. BIG DATA CLUSTERING CHALLENGES.

On the web data is growing unbelievably because of the use of the internet since the number of internet users are growing. With the number of new data sources, the complexity of data is also increasing. There are three categories for this huge streaming data– structured data, unstructured data and semi-structured data. There are 6Vs that define big data. *Volume:* the size of data is growing in the blink of eyes. Today the size of those data is larger than terabytes and petabytes.

Variety: this defines the types of data. A data can be anything, such as- audio, video, image, written document etc. *Velocity:* it's the generating and processing speed of a data. *Value:* it's the information a person can obtain from the data. *Veracity:* the trustworthiness or the reliability of the data. *Validation:* it shows if the purpose of the data has been served or not.

To cluster these massive data according to their respective group is not an easy task. The potentialchallenges have been identified as follows:

The identification of distance measure: Euclidian, Manhattan, and maximum distance measure can be used for numerical attributes as a distance measure technique. But for categorical elements this identification measure is problematic. *Lack of class labels:* the distribution of data has to be done to understand where the data labels are, for real dataset. Structure of data: it’s not real that the dataset will be structured all the time. They may not always contain identifiable clusters. Also the order of arranged data may affect the result of the algorithm. Missing values may exist in a dataset. If a dataset has missing values then the attribute has to be removed, described in [5]. Also, based on the mean-and-mode method in [6] three cluster-based algorithms to deal with missing values have been proposed. *Types of attributes in database:* it’s not necessary that the database will only contain distinctively numerical and categorical attributes. They may contain other types also such as: nominal, ordinal, binary etc. *Choosing the initial cluster:* it’s one of the most difficult part during clustering a dataset, because if the initial cluster is not chosen correctly, then after a few repetitions it is found that clusters may even be left void.

3. CLUSTERING METHODS

The clustering technique can be used for both labelled and no labelled data. The main role of clustering is to group data based on their similarity [7]. The simple Clustering are divided by four categories: i) density-based method, ii) grid-based method, iii) partitioning method, iv) hierarchical method. The partitioning method concepts k clusters of the given set of N instances, where k ≤ N. Traditional distance measure uses for mutually special clusters shape [8]. Mean or median are used to find the cluster center and use recurrent formula to proceed the clustering by affecting instances from one cluster to another (i.e k-means clustering) [9]. In high dimension data the partitioning algorithm are fail to cluster. A hierarchical dividing of N instances occurs for the hierarchical approaches Top-down and bottom-up approaches use in hierarchical clustering. In top-down approach N instances within a single cluster divide into small cluster and repeat it for each iteration until a conclusion condition holds. In bottom up approach each single instance are unique cluster. Merges the closes cluster with another until all clusters merge into one cluster. The density-based methods cluster instances constructed on the distance between instances, which can find arbitrarily shaped clusters.

It can cluster instances as dense regions in the data space, separated by sparse regions. The grid-based methods use a multi-resolution grid data structure. It’s fast processing time that typically self-determining of the number of instances, yet dependent on the grid size. Clustering hasbeen widely used in many real world submissions such as human genetic clustering, medical imaging clustering, market research, field robotics, crime analysis, and pattern recognition. In this section, the basic clustering algorithms that applied in many real life clustering problems have been discussed.

3.1 K- Means clustering

To identify the structure in data set, k-means algorithm is an effective and useful formula. The dataset divides into k number of subsets in k-means clustering. Each cluster is defined by centroid in k-means clustering and average point is used to measure the centroid. We can find out useful result on dataset using k-mean algorithm [10], [11]. At the Same time noise and outliers affects to get the best output from algorithm. In k-mean we can divide n instance of a dataset to k clusters and distance is used in k-means clustering algorithm finds the nearest cluster for each point.

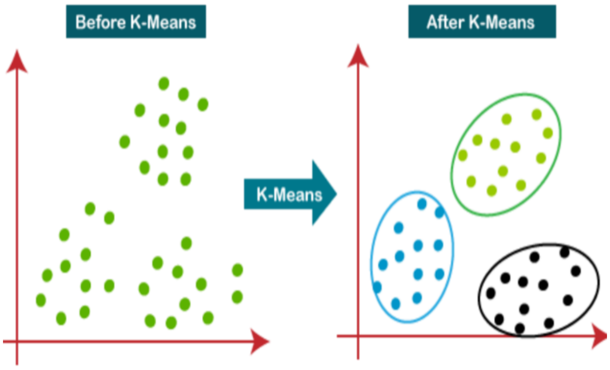


Fig 1: K-means clustering

The basic algorithm is very simple and the steps are
Step-1: Choose k to find no of clusters.
Step-2: Choose random k points.
Step-3: To find predefined k clusters, put each point totheir nearest centroid.
Step-4: Define the variance and place a new centroid of each cluster
Step-5: Repeat the step 3, that means reassign each pointto the closet centroid of each cluster.
Step-6: Go to step 4 if have any reassignment.
Step-7: The model is ready

3.2 Density-Based clustering

The density based clustering is useful to remove noisy data. Traditional methods are not suitable to cluster density based clustering. Idea of data density is the main concept of density-based algorithm. Using density based method various kinds of algorithm are made i.e DBSCAN which is used to improve order of data collection structure. DBSCAN uses computing the number of points to identify density and linked those points within each other’s neighborhood. Two user defined types like a and b use in Density based algorithm [12], [13]. Neighborhood of a point identify by a and minimum points of neighborhood identify by b. Using b parameters DBSCAN can make arbitrary shaped clusters. It can also handle noise and outliers. A point denoted by i is a density reachable from a point j with respect to Eps, MinPts if there is a sequence chain of a point i1,…., in, i1 = j, pn = i such that $i_{i.}$ + 1 is directly density reachable from $i_{i.}$DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It depends on a density-based notion of cluster. It also identifies clusters of arbitrary size in the spatial database with outliers.

The basic algorithm is very easy to understand which is given bellow:

*Step-1:*Define a point K.

*Step-2:*Select all density reachable from K.

*Step-3:*If K is a main point then a cluster is created.

*Step-4:*If K is a margin point, then no point is density-reachable and DBSCAN forward to the next point.

*Step-5:*Mentioned process are continued until all the points are completed.

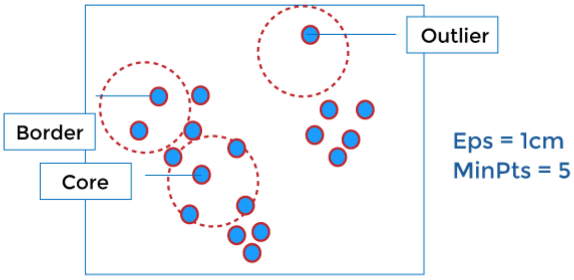


Fig 2: Density-based clustering

4. PROPOSED METHODS

Earlier the K-means clustering technique has been illustrated, but k-means clustering and removing noisy data at the same time is very difficult and time consuming. That's why a method is being shown which might be fruitful to do both at the same time.

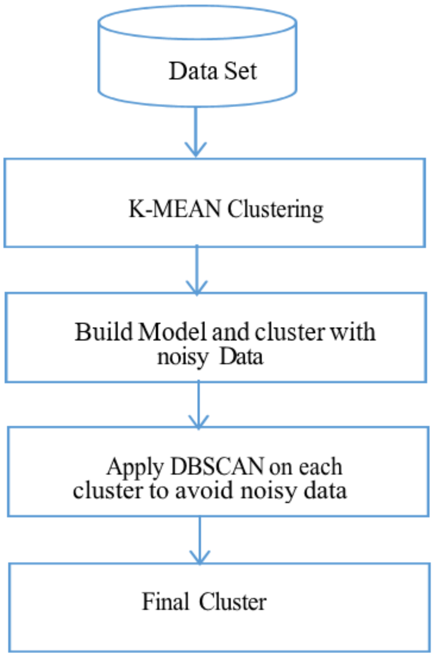

Data Set
K-MEAN Clustering
Build Model and cluster with noisy Data
Apply DBSCAN on each cluster to avoid noisy data
Final Cluster

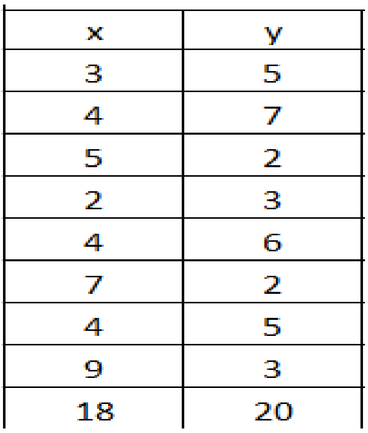
Fig 3: Proposed methods

The steps are drawn below:
Step-1: Choose k to find no of clusters.
Step-2: Choose random k points.
Step-3: To find predefined k clusters, put each point to their nearest centroid.
Step-4: Define the variance and place a new centroid of each cluster.
*Step-5:*Repeat the step 3, that means reassign each point to the closet centroid of each cluster.
*Step-6:*Go to step 4 if have any reassignment.
Step-7: The model is ready with noisy data and apply DBSCAN algorithm on this model to avoid noisy data.
Step-8: Arbitrary select a point P for each cluster of noisy model.
Step-9: Select all point which is density reachable from K
Step-10: If K is a core point then a cluster is made.
Step-11: If K is a marginal point, then there is no point which is density-available and DBSCAN forward to the next point.
Step-12: Mentioned process are continued until all the points are completed.
Step-13: Repeat the steps 10 to 13 for each cluster
Step-14: The model is ready.

4.1 Experimental analysis

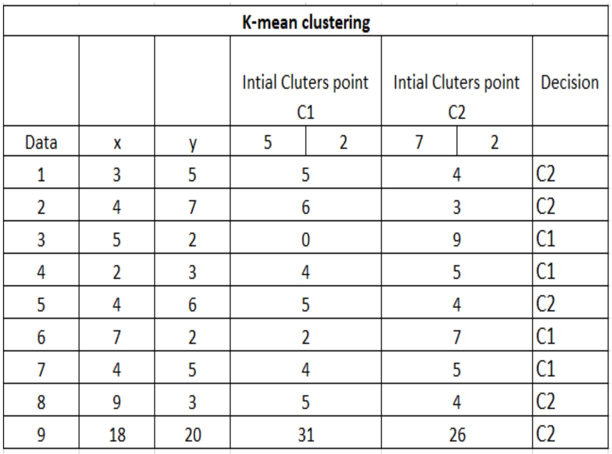
Here a structured data set has been taken on which the proposed method will be applied.

Table 1. Structured data set for the experiment



x	y
3	5
4	7
5	2
2	3
4	6
7	2
4	5
9	3
18	20

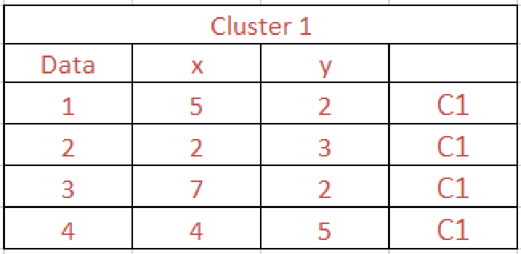
Table 2. Applying K-means clustering on the data set



K-mean clustering							
			Intial Cluters point C1		Intial Cluters point C2		Decision
Data	x	y	5	2	7	2	
1	3	5	5		4		C2
2	4	7	6		3		C2
3	5	2	0		9		C1
4	2	3	4		5		C1
5	4	6	5		4		C2
6	7	2	2		7		C1
7	4	5	4		5		C1
8	9	3	5		4		C2
9	18	20	31		26		C2

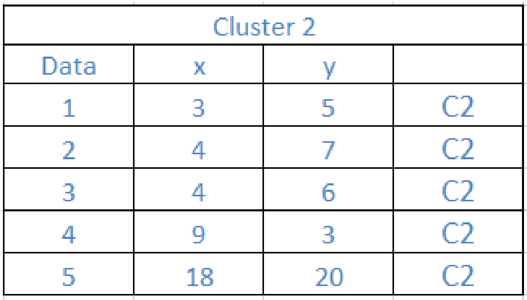
In table 2 is showing that k-means clustering has been applied on the given data set. Where, C1 and C2 are two clusters and for C1(5,2) are the initial cluster point and for C2 cluster (7, 2) are the initial cluster point. Here,Euclidean distance formula has been used to find the distance. And in the decision section it's clearly shown that C1 cluster contains (3,4,6,7) number of data and C2 cluster contains (1,2,5,8) number of data.

Table 3. Data set of cluster 1



Cluster 1			
Data	x	y	
1	5	2	C1
2	2	3	C1
3	7	2	C1
4	4	5	C1

Table 4. Data set of cluster 2



Cluster 2			
Data	x	y	
1	3	5	C2
2	4	7	C2
3	4	6	C2
4	9	3	C2
5	18	20	C2

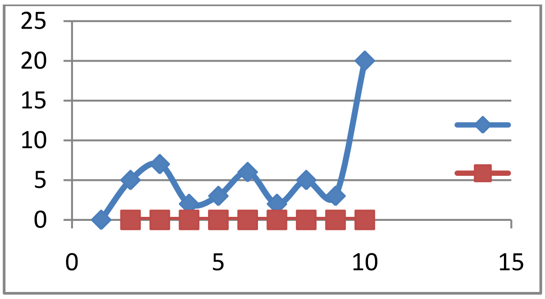
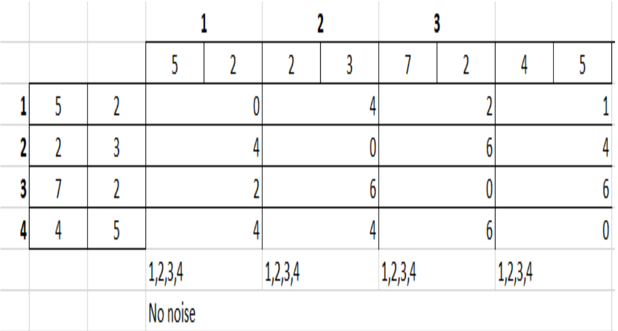


Fig 4. Graph after applying k-means clustering

Now DBSCAN algorithm will be applied to remove noise from each cluster.

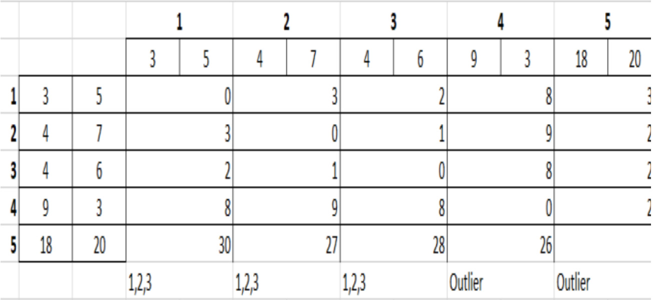
Table 4. DBSCAN on cluster 1



			1		2		3			
			5	2	2	3	7	2	4	5
1	5	2	0		4		2		1	
2	2	3	4		0		6		4	
3	7	2	2		6		0		6	
4	4	5	4		4		6		0	
			1,2,3,4		1,2,3,4		1,2,3,4		1,2,3,4	
			No noise							

In table 4 the number of elements are within distance 6 and minimum point is 2. After calculation it’s clearly visible that no noise has been found from the data of table 3 after applying DBSCAN algorithm.

Table 5. DBSCAN on cluster 2



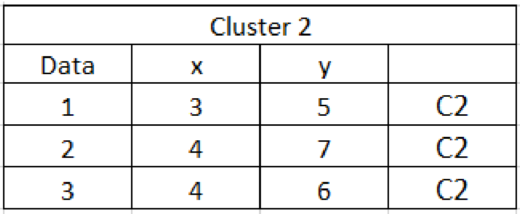
			1		2		3		4		5	
			3	5	4	7	4	6	9	3	18	20
1	3	5	0		3		2		8		3	
2	4	7	3		0		1		9		2	
3	4	6	2		1		0		8		2	
4	9	3	8		9		8		0		2	
5	18	20	30		27		28		26			
			1,2,3		1,2,3		1,2,3		Outlier		Outlier	

In table 4 the number of elements are within distance 3 and the minimum point is 2. And after applying DBSCAN algorithm on the data of cluster 2, two outliers have been found.

4.2 Result and discussion

Here, an idea of using two algorithms has been shared to remove the noisy data. After taking a structured data set K-means clustering is applied on that and then DBSCAN algorithm is applied. After applying k-means clustering algorithm Table 3 and Table 4 have been found. When DBSCAN applied on cluster 1 there was no noise but when the same algorithm applied on cluster 2 there two outliers have been found and later because of the distance and minimum point the outliers are removed and the final output of cluster 2 is shown in Table 6.

Table 6. Final output after applying DSCAN algorithm



Cluster 2			
Data	x	y	
1	3	5	C2
2	4	7	C2
3	4	6	C2

5. CONCLUSION

In this paper, authors proposed a method by combining K-means and density-based clustering. In most cases, some noisy and duplicate data are being found during each cluster which might not be useful in further procedure, that’s DBSCAN clustering technique has been applied on each cluster to remove noisy and unwanted data. Here authors also share experimental analysis for removing the noise of the data. Since the proposed method has only been applied on numericalstructured data set.In future authors hope to apply the same proposed method on categorical and semi structured data set to remove the noise from the data set.

6. ACKNOWLEDGMENT

The authors want to thank professor Dr. DewanMd.Farid who contributed to, reviewed and helped to write this paper.

7. REFERENCES

[1] L.-J. Zhang, J. Zhang, and H. Cai, Services computing. Springer, 2007.

[2] F. Curbera et al. Unraveling the Web Services: An Introduction to SOAP, WSDL, and UDDI. IEEE Internet Computing, Mar/ Apr issue 2002.

[3] D.A. Menasce, “QoS issues in web services,” IEEE Internet Compute., vol. 6, pp. 72–75, 2002.

[4] Tao Yu, Yue Zhang, and Kwei-Jay Iin, “Efficient Algorithms for Web Services Selection with End-to-End QoS Constraints”, ACM Transactions on the Web, Vol. 1, No. 1, Article 6, Publication date: May 2007.

[5] Kaufman, L. and Rousseeuw, P. (1990). Finding Groups in Data—An Introduction to Cluster Analysis. Wiley Series in Probability and Mathematical Statistics. NewYork: JohnWiley& Sons, Inc.

[6] Fujikawa, Y. and Ho, T. (2002). Cluster-based algorithms for dealing with missing values.In Cheng, M.-S., Yu, P. S., and Liu, B., editors, Advances in Knowledge Discovery and Data Mining, Proceedings of the 6th Pacific-Asia Conference, PAKDD 2002, Taipei,Taiwan, volume 2336 of Lecture Notes in Computer Science, pages 549–554. New York:Springer.

[7] S. Y. Hwang, H. Wang, J. Tang, and J. Srivastava, “A probabilistic approach to modeling and estimating the QoS of web-servicesbased workflows,” Info. Sci., vol. 177, pp. 5484–5503, 2007.

[8] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, “Automated variable weighting in k-means type clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 5, pp. 657–668, May 2005.

[9] A. K. Jain, “Data clustering: 50 years beyond k-means,” Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, June 2010.

[10] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, ‘‘Basic Concepts and Taxonomy of Dependable and Secure Computing,’’ IEEE Trans. Dependable Secure Comput., vol. 1, no. 1, pp. 11 /33, Jan.-Mar. 2004.

[11] Marin Silic, GoranDelac, Ivo Krka and SinisaSrbljic, “Scalable and Accurate Prediction of Availability of Atomic Web Services”, IEEE transaction on service computing, 2014.

[12] L. Shao, J. Zhang, Y. Wei, J. Zhao, B. Xie, and H. Mei, “Personalized QoS prediction for Web services via collaborative filtering, ”in Proc. 5th International Conference on Web Services (ICWS 2007), 2007, pp. 439446.

[13] T.Miranda Lakshmi, R.JosephineSahana, V.PrasannaVenkatesan. Review on Density-Based Clustering Algorithms for Big Data p.15, p.17.