A COMPARATIVE STUDY OF K-MEANS, DBSCAN AND OPTICS

Hari Krishna Kanagala
Asst. Prof
Department of MCA
Vignan's Lara Institute of Technology & Sci.
Vadlamudi, Guntur Dist, India.
harikanagala@gmail.com

Dr. V.V. Jaya Rama Krishnaiah
Assoc. Prof.
Dept. of Computer Science & Engineering
ASN Women's Engineering College
Tenali, Guntur Dist, India
jkvemula@gmail.com

Abstract - In view of today's information available, recent progress in data mining research has lead to the development of various efficient methods for mining interesting patterns in large databases. It plays a vital role in knowledge discovery process by analyzing the huge data from various sources and summarizing it into useful information. It is helpful for analyzing the volumes of data in different domains like Marketing, Health, Science and Technology. Cluster analysis is widely used approach to notice the trends in the volumes of data. In this paper, we evaluated the performance of the different clustering approaches like as K-Means, DBSCAN, and OPTICS in terms of accuracy, outlier's formation, and cluster size prediction.

Keywords: Clustering, k-means, dbscan, optics

I. INTRODUCTION

Clustering is the process of partitioning the set of objects or data into a set of classes of similar objects. Clustering having maximum similarity in between the data objects of the same class and minimum similarity in between the objects of different classes. The quality of the clustering result depends on the similarity measure used by the method. The similarity measure is expressed in terms of a distance function. The distance functions are very different for interval-scaled, Boolean, categorical, ordinal and ratio variables. Distance measure will determine the similarity of the two elements and it will influence the shape of the clusters. Many distance measures are used such as Euclidian distance, Manhattan distance, Minkowski distance. This paper presents different clustering algorithms such as K-Means, DBSCAN, OPTICS and the performance evaluation of those algorithms.

II. K-MEANS

The input parameters for the K-MEANS algorithm is the number of clusters, k, and partitions a set of n objects into k clusters containing data points so as to minimize the sum of the squared error criterion by iteratively. Cluster similarity is measured by using the mean value of the cluster objects. First k number of objects each represents a cluster mean or center is randomly selected. For each of the remaining objects, an object is assigned to the cluster by using the distance measurement in between the object and cluster mean. For each of the cluster, a new mean value to be computed. This process iterates until the criterion minimizes the sum of the squared error minimized. The sum of the squared error is defined as

$$E = \sum_{i=1}^{k} \sum_{x \in ci} | x - m_i |^2 \qquad (1)$$

K-Means works as follows

1. The algorithm arbitrarily selects k number of data objects initially as the cluster mean or cluster centers.
2. Compute the distance measurement such as the Euclidian distance between each data object and the cluster center, each data object is assigned to the closest cluster.
3. Recompute each cluster center as the average of the data objects in that cluster.
4. Repeat the steps 2 and 3 until no change in clusters.

III. DBSCAN (Density Based Spatial Clustering of Applications with Noise)

Density-based clustering identifies regions of high density that are separated from one and other by regions of low density. The density is defined as a minimum number of objects within a certain distance of each other. The DBSCAN approach is to create clusters with a minimum size and density. In the center based approach, a point to be classified as a core point or border point or noise point. A core point has more than a specified number of points (MinPts), within the specified radius (Eps). A border point has fewer than MinPts within Eps but is in the neighborhood of a core point. A noise point is neither a core point nor a border point. An object is called the $\in$-neighborhood of the object if the neighborhood is within the radius of a given object. An object is called a core object if the ε-neighborhood of an object contains at least a minimum number, MinPts, of

978-1-4673-6680-9/16/$31.00 ©2016 IEEE

objects. An object p is directly density-reachable from object q if p is within the ε-neighborhood of q, and q is a core object. A point is said to be density-reachable from another point if there is a chain of points from one to the other, which contains only points that are directly density-reachable to each other. An object p is density-connected to object q if there is an object o such that both p and q are density-reachable from o. The algorithm defines a cluster as the maximal set of density-connected points. [1]

The DBSCAN works as follows

1. Label all points as Core, Border or Noise points
2. Eliminate Noise points.
3. Put an edge between all core points that are within neighborhood of each other
4. Make each group of connected core points into a separate cluster.
5. Every border point is assigned to one of the clusters of its associated core points.

IV. OPTICS

OPTICS is the Density Based clustering by creating an ordering of the points that allows the extraction of clusters with arbitrary values for ε. The parameter ε is a distance, it is the neighborhood radius. Therefore, in order to generate a set or ordering of density-based clusters, we provide a set of distance parameter values. To construct the different clustering's simultaneously, the objects should be processed in a specific order. This order selects an object that is density-reachable with respect to the lowest value so that clusters with higher density (lower ε) will be finished first. The generating distance $\in$ is the largest distance considered for clusters. Clusters can be extracted for all ε_i such that $0 \le \varepsilon_i \le \varepsilon$. Based on this idea, two values need to be stored for each object core distance and the reachability distance. The core distance of an object p is the smallest ε' value that makes p as a core object. If p is not a core object, the core distance of p is undefined. The reachability-distance of an object p and another object o is the greater value of the core-distance of p and the Euclidean distance between p and q. If p is not a core object, the reachability-distance between p and q is undefined. [1]

A reachability plot for a simple 2-dimensional data set, which shows the data are clustered. [1]

V. PERFORMANCE EVALUATION

The results obtained after running of the clustering techniques for abalone data set which consists of 4177 instances and 9 attributes such as gender, length, diameter, whole height, whole weight, shucked weight, viscera weight, shell weight and rings. To construct the algorithms, we use Waikato Environment for Knowledge Analysis (WEKA version 3.6.10), an open source data mining tool which was developed at University of Waikato New Zealand. WEKA is an open source application that is freely available under the GNU general public license agreement. This experiment is performed on Duo Core with 2.10 GHz CPU and 4G RAM. The result for each clustering algorithms is shown and described below.

The table I shows the result of K-MEANS clustering algorithm on the Iris dataset for different number of clusters.

TABLE I. The result of K-MEANS on the Iris data set for different number of clusters

K-MEANS		No. of Clusters			
		10	15	20	25
Abalone Dataset No. of Instances: 4177 No. of attributes: 9	No. of Iterations	30	37	63	34
	Sum of Squared Error	165.36	92.73	71.41	62.48
	Clustered Instances	C0-379, C1-296 C2-851, C3-287, C4-78, C5-201, C6-632, C7-373, C8-403, C9-677	C0-239, C1-201,C2-200, C3-287,C4-78, C5-201,C6-412, C7-373,C8-403, C9-475,C10-349, C11-188, C12-503,C13-162 C14-106	C0-149, C1-42,C2-202, C3-205,C4-49, C5-166,C6-314, C7-315, C8-305, C9-358, C10-219,C11-132, C12-360,C13-139 C14-182,C15-299, C16-302,C17-264, C18-102,C19-73	C0-103, C1-42,C2-89, C3-55,C4-46, C5-166, C6-249, C7-301,C8-272, C9-258,C10-222, C11-104,C12-300, C13-127,C14-142, C15-318,C16-189, C17-209,C18-68, C19-39,C20-175, C21-313,C22-182, C23-163,C24-45

When the number of clusters increases, the corresponding sum of squared error is decreased.

Fig. 1 shows the graph which visualizes the cluster assignments in K-MEANS when the input parameter, the number of clusters is 10 for the Abalone dataset.

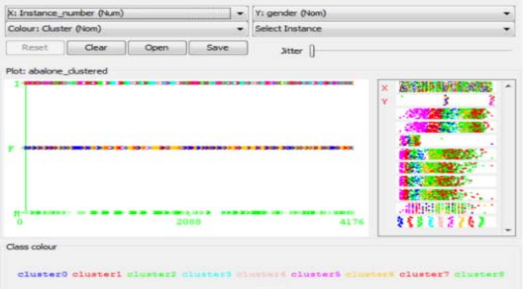

X: Instance_number (Num)
Y: gender (Nom)
Colour: Cluster (Nom)
Plot: abalone_clustered
Select Instance
Reset
Clear
Open
Save
Jitter
I
F
M
0
2088
4176
X
Y
Class colour
cluster0 cluster1 cluster2 cluster3 cluster4 cluster5 cluster6 cluster7 cluster8

Figure 1. Cluster assignments in K-MEANS when the number of clusters is 10

Fig. 2 shows the graph which visualizes the cluster assignments in K-MEANS when the input parameter, the number of clusters is 15 for the Abalone dataset.

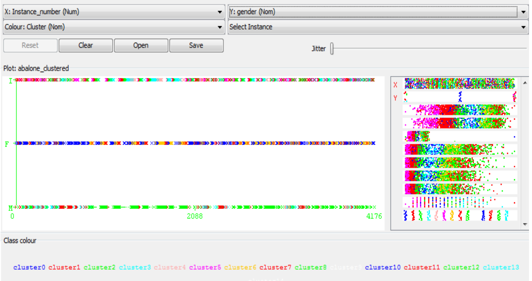

X: Instance_number (Num)
Y: gender (Nom)
Colour: Cluster (Nom)
Plot: abalone_clustered
Select Instance
Reset
Clear
Open
Save
Jitter
I
F
M
0
2088
4176
X
Y
Class colour
cluster0 cluster1 cluster2 cluster3 cluster4 cluster5 cluster6 cluster7 cluster8 cluster10 cluster11 cluster12 cluster13

Figure 2. cluster assignments in K-MEANS when the number of clusters is 15

Fig. 3 shows the graph which visualizes the cluster assignments in K-MEANS when the input parameter, the number of clusters is 20 for the Abalone dataset.

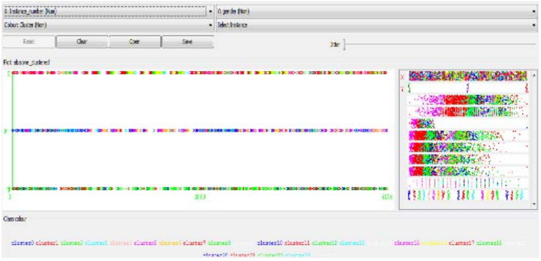


Figure 3. Cluster assignments in K-MEANS when the number of clusters is 20

Fig. 4 shows the graph which visualizes the cluster assignments in K-MEANS when the input parameter, the number of clusters is 25 for the Abalone dataset.

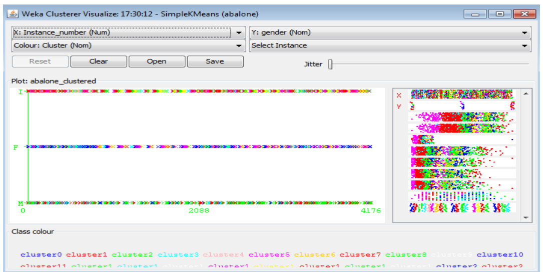

Weka Clusterer Visualize: 17:30:12 - SimpleKMeans (abalone)
X: Instance_number (Num)
Y: gender (Nom)
Colour: Cluster (Nom)
Plot: abalone_clustered
Select Instance
Reset
Clear
Open
Save
Jitter
I
F
M
0
2088
4176
X
Y
Class colour
cluster0 cluster1 cluster2 cluster3 cluster4 cluster5 cluster6 cluster7 cluster8 cluster10

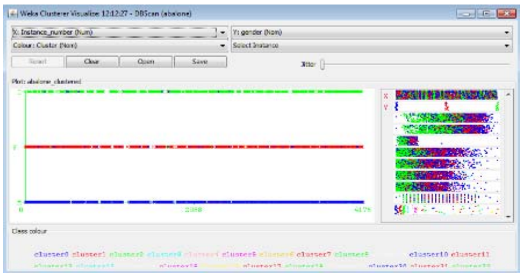
Figure 4. Cluster assignments in K-MEANS when the number off clusters is 25

The table II is the result of DBSCAN on Abalone data set which contains of 9 attributes and 4177 instances on different ∈ and MinPts parameters.

TABLE II. The result of DBSCAN on abalone data set with different ∈ and MinPts parameters.

DBSCAN	∈ = 0.1, MinPts = 2	∈ = 0.2, MinPts =2	∈ = 0.3, MinPts = 6	∈ = 0.4, MinPts = 5	∈ = 0.8, MinPts = 5	∈ = 1.0, MinPts = 5
No. of generated clusters	28	4	3	3	3	3
No. of Unclustered instances	179	11	6	1	1	0
Elapsed Time	24.66	20.72	23.27	26.5	25.58	26.38
Clustered Instances	C0-1422 C1-1197 C2-1307 C3-4,C4-2 C5-7, C6-6 C7-2, C8-4 C9-3, C10-2 C11-2,C12-2 C13-2,C14-2 C15-2,C16-2 C17-2,C18-2 C19-8,C20-4 C21-2,C22-2 C23-2,C24-2 C25-2,C26-2 C27-2	C0-1524 C1-1300 C2-1339 C3-3	C0-1527 C1-1304 C2-1340	C0-1528 C1-1306 C2-1342	C0-1528 C1-1306 C2-1342	C0-1528 C1-1307 C2-1342

Fig. 5 shows the graph which visualizes the cluster assignments in DBSCAN when the input parameter, ∈ = 0.1, MinPts = 2 for the Abalone dataset.



Weka Clusterer Visualize: 12:12:27 - DBScan (abalone)

X: Instance_number (Num)

Y: gender (Nom)

Colour: Cluster (Nom)

Select Instance

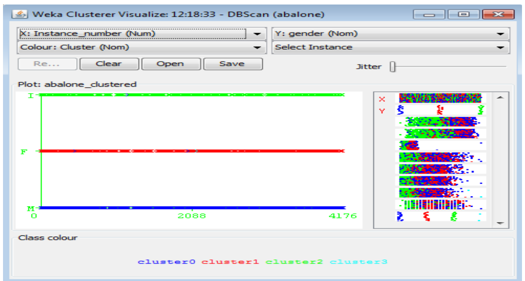
Plot: abalone_clustered

Class colour

Reset
Clear
Open
Save
Jitter
0
2088
4176
X
Y
cluster0 cluster1 cluster2 cluster3 cluster4 cluster5 cluster6 cluster7 cluster8 cluster10 cluster11

Figure 5. Cluster assignments in DBSCAN when ∈ = 0.1, MinPts = 2 for the Abalone dataset.

Fig. 6 shows the graph which visualizes the cluster assignments in DBSCAN when the input parameter, ∈ = 0.2, MinPts = 2 for the Abalone dataset.



Weka Clusterer Visualize: 12:18:33 - DBScan (abalone)

X: Instance_number (Num)

Y: gender (Nom)

Colour: Cluster (Nom)

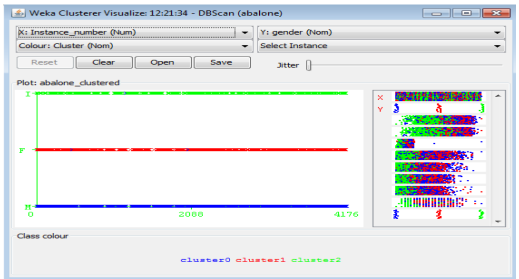
Select Instance

Plot: abalone_clustered

Class colour

Figure 6. Cluster assignments in DBSCAN when ∈ = 0.2, MinPts = 2 for the Abalone dataset.

Fig. 7 shows the graph which visualizes the cluster assignments in DBSCAN when the input parameter ∈ = 0.3, MinPts = 6 for the Abalone dataset.



Weka Clusterer Visualize: 12:21:34 - DBScan (abalone)

X: Instance_number (Num)

Y: gender (Nom)

Colour: Cluster (Nom)

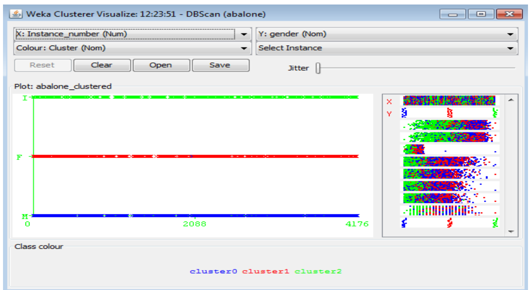
Select Instance

Plot: abalone_clustered

Class colour

Figure 7. Cluster assignments in DBSCAN when ∈ = 0.3, MinPts = 6 for the Abalone dataset.

Fig. 8 shows the graph which visualizes the cluster assignments in DBSCAN when the input parameter, ∈ = 0.4, MinPts = 5 for the Abalone dataset.



Weka Clusterer Visualize: 12:23:51 - DBScan (abalone)

X: Instance_number (Num)

Y: gender (Nom)

Colour: Cluster (Nom)

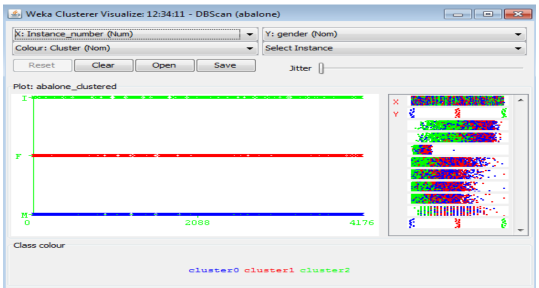
Select Instance

Plot: abalone_clustered

Class colour

Figure 8. Cluster assignments in DBSCAN when ∈ = 0.4, MinPts = 5 for the Abalone dataset.

Fig. 9 shows the graph which visualizes the cluster assignments in DBSCAN when the input parameter, ∈ = 0.8, MinPts = 5 for the Abalone dataset.



Weka Clusterer Visualize: 12:34:11 - DBScan (abalone)

X: Instance_number (Num)

Y: gender (Nom)

Colour: Cluster (Nom)

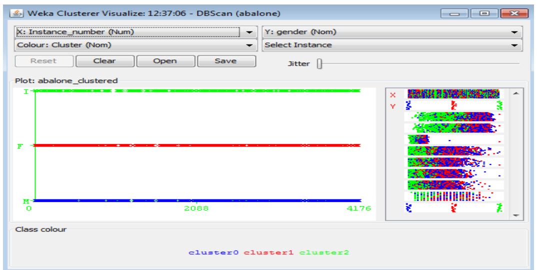
Select Instance

Plot: abalone_clustered

Class colour

Figure 9. Cluster assignments in DBSCAN when ∈ = 0.8, MinPts = 5 for the Abalone dataset.

Fig. 10 shows the graph which visualizes the cluster assignments in DBSCAN when the input parameter, ∈ = 1.0, MinPts = 5 for the Abalone dataset.



Weka Clusterer Visualize: 12:37:06 - DBScan (abalone)

X: Instance_number (Num)

Y: gender (Nom)

Colour: Cluster (Nom)

Select Instance

Plot: abalone_clustered

Class colour

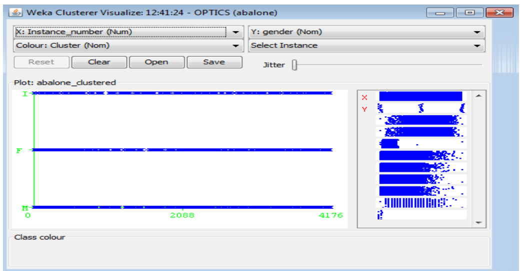
Figure 10. Cluster assignments in DBSCAN when ∈=1.0, MinPts = 5 for the Abalone dataset.

The table III is the result of OPTICS clustering on Abalone data set which consists of 9 attributes and 4177 instances on different ∈ and MinPts parameters.

Table III. The result of OPTICS on abalone data set with different ∈ and MinPts parameters.

OPTICS	∈ = 0.1, MinPts = 2	∈ = 0.2, MinPts =2	∈ = 0.3, MinPts = 6	∈ = 0.4, MinPts = 5	∈ = 0.8, MinPts = 5	∈ = 1.0, MinPts = 5
No. of generated clusters	0	0	0	0	0	0
No. of Unclustered instances	4177	4177	4177	4177	4177	4177
Elapsed Time	25.71	24.38	20.89	30.24	18.32	31.62

Fig. 11 shows the graph which visualizes the cluster assignments in OPTICS when the input parameter, ∊ = 0.1, MinPts = 2 for the Abalone dataset.


Weka Clusterer Visualize: 12:41:24 - OPTICS (abalone)

X: Instance_number (Num)

Y: gender (Nom)

Colour: Cluster (Nom)

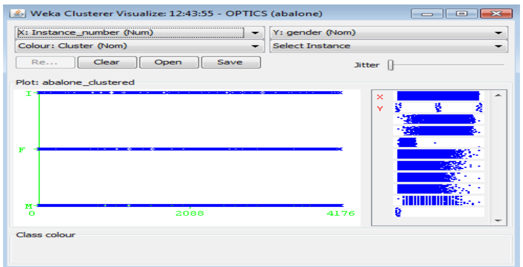
Select Instance

Plot: abalone_clustered

Class colour

Figure 11. Cluster assignments in OPTICS when ∊ = 0.1, MinPts = 2 for the Abalone dataset.

Fig. 12 shows the graph which visualizes the cluster assignments in OPTICS when the input parameter, ∊ = 0.2, MinPts = 2 for the Abalone dataset.


Weka Clusterer Visualize: 12:43:55 - OPTICS (abalone)

X: Instance_number (Num)

Y: gender (Nom)

Colour: Cluster (Nom)

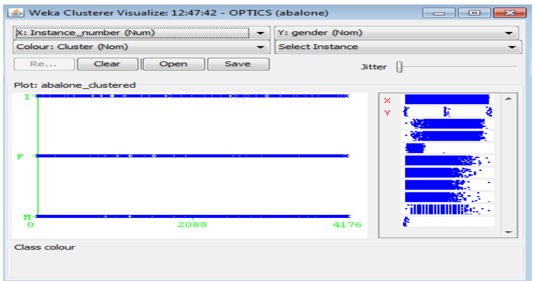
Select Instance

Plot: abalone_clustered

Class colour

Figure 12. Cluster assignments in OPTICS when ∊ = 0.2, MinPts = 2 for the Abalone dataset.

Fig. 13 shows the graph which visualizes the cluster assignments in OPTICS when the input parameter, ∊ = 0.3, MinPts = 6 for the Abalone dataset.


Weka Clusterer Visualize: 12:47:42 - OPTICS (abalone)

X: Instance_number (Num)

Y: gender (Nom)

Colour: Cluster (Nom)

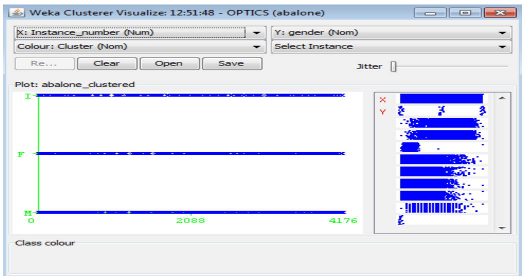
Select Instance

Plot: abalone_clustered

Class colour

Figure 13. Cluster assignments in OPTICS when ∊ = 0.3, MinPts = 6 for the Abalone dataset.

Fig. 14 shows the graph which visualizes the cluster assignments in OPTICS when the input parameter, ∊ = 0.4, MinPts = 5 for the Abalone dataset.


Weka Clusterer Visualize: 12:51:48 - OPTICS (abalone)

X: Instance_number (Num)

Y: gender (Nom)

Colour: Cluster (Nom)

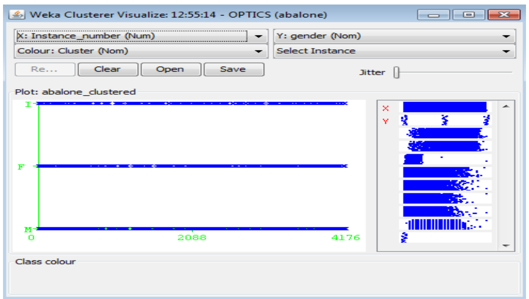
Select Instance

Plot: abalone_clustered

Class colour

Figure 14. Cluster assignments in OPTICS when ∊=0.4, MinPts=5 for the Abalone dataset.

Fig. 15 shows the graph which visualizes the cluster assignments in OPTICS when the input parameter, ∊ = 0.8, MinPts = 5 for the Abalone dataset.


Weka Clusterer Visualize: 12:55:14 - OPTICS (abalone)

X: Instance_number (Num)

Y: gender (Nom)

Colour: Cluster (Nom)

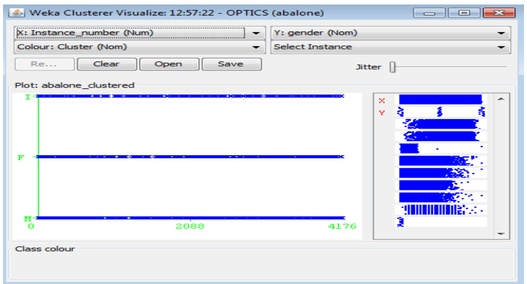
Select Instance

Plot: abalone_clustered

Class colour

Figure 15. Cluster assignments in OPTICS when ∊ = 0.8, MinPts = 5 for the Abalone dataset.

Fig. 16 shows the graph which visualizes the cluster assignments in OPTICS when the input parameter, ∊ = 1.0, MinPts = 5 for the Abalone dataset.


Weka Clusterer Visualize: 12:57:22 - OPTICS (abalone)

X: Instance_number (Num)

Y: gender (Nom)

Colour: Cluster (Nom)

Select Instance

Plot: abalone_clustered

Class colour

Figure 16. Cluster assignments in OPTICS when ∊ = 1.0, MinPts = 5 for the Abalone dataset.

CONCLUSION

K-Means algorithm is only applied when the mean of the cluster is defined. K-Means algorithm produces a quality of clusters when using huge dataset. The number of clusters, K, must be specified, in advance. K-Means will not identify Outliers. DBSCAN can find clusters of arbitrary shape, determine what information should be classified as noise or outliers. It is very fast when compared to other algorithms. In DBSCAN, the user has the responsibility of selecting the parameter values (ε and MinPts). Slightly different parameter settings may lead to different clusters. It has some difficulties in distinguishing separated clusters if they are located too close to each other, even though they have different densities. To overcome this difficulty, OPTICS algorithm was developed. OPTICS ensures good quality clustering by maintaining the order in which the data objects are processed, i.e., high-density clusters are given priority over lower density clusters. OPTICS also requires a parameters (ε and MinPts) to be specified by the user that will affect the result. The efficiency of clustering algorithms can be improved by removing the limitations of the clustering techniques.

REFERENCES

[1] Jiawei Han, MichelineKamber, Jian Pei, “Data Mining Concepts and Techniques” Elsevie Second Edition.

[2] Pang-Ning Tan,Vipin Kumar, Michael Steinbach, “Introduction to Data Mining” Pearson.

[3] Bharat Chaudhari, Manan Parikh “A Comparative Study of clustering algorithms Using weka tools” International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 1, Issue 2, October 2012 ISSN 2319 – 4847 pp 154-158.

[4] V.V.Jaya RamaKrishnaiah, Dr.K.Ramchand H Rao, Dr. R.Satya Prasad “Entropy Based Mean Clustering: A Enhanced Clustering Approach” The International Journal of Computer Science & Applications (TIJCSA) Volume 1, No. 3, May 2012 ISSN – 2278-1080 pp 1-9.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, XiaoweiXu “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise” from KDD-96 proceedings pp 226-231

[6] Narendra Sharma, AmanBajpai, Mr.RatneshLitoriya “Comparison the various clustering algorithms of weka tools” International Journal of Emerging Technology and Advanced Engineering Volume 2, Issue 5, May 2012 ISSN 2250-2459 pp 73-80

[7] Kaushik H. Raviya, KunjanDhinoja “An Empirical Comparison of K-Means and DBSCAN Clustering Algorithm” Paripex – Indian Journal of Research Volume: 2 Issue ISSN - 2250-1991 pp 153-155

[8] Pradeep Rai Shubha Singh “A Survey of Clustering Techniques” International Journal of Computer Applications (0975 – 8887) Volume 7– No.12, October 2010 pp 1-5

[9] P. IndiraPriya, Dr. D.K.Ghosh “A Survey on Different Clustering Algorithms in Data Mining Technique” International Journal of Modern Engineering Research (IJMER) Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274 ISSN: 2249-6645

[10] B.G.Obula Reddy, Dr. Maligela Ussenaiah “Literature Survey On Clustering Techniques” IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 3, Issue 1 (July-Aug. 2012), PP 01-12