ANALYSIS OF DBSCAN CLUSTERING TECHNIQUE ON DIFFERENT DATASETS USING WEKA TOOL

Iftikhar Hussain Babur [*], Jawad Ahmad*[*], Bilal Ahmad, Muhammad Habib

Department of Computer Science, Lahore Leads University, Lahore

Contact: *iftibabur@gmail.com [**], jawadahmad412@gmail.com

ABSTRACT— *Data mining is used to extract hidden information pattern from a large dataset which may be very useful in decision making. The main task of exploratory data analysis and data mining applications is clustering. In clustering, the data is divided in groups containing similar data that is dissimilar to the data in other groups. Various techniques of clustering have been suggested by researchers based on low distances or probability, etc. In this paper, an analysis of well known clustering algorithm DBSCAN on different datasets using WEKA clustering tool has been presented. The results show that this clustering algorithm overloads the user in choosing the input parameters carefully for proper clustering.*

Keywords— Clustering, density-based clustering, DBSCAN

1. INTRODUCTION

Data Mining or "Knowledge Discovery in data (KDD)" is a method for extracting valuable information from the huge volume of data. Different data mining techniques are classification, prediction, clustering, summarization, association rules and sequence discovery [1]. The techniques and algorithms of data mining have been drawn from the fields of Statistics, Machine Learning and Data Base Management Systems. Due to blistering increase in the storage of data, the stake in the discovery of hidden information in the databases has exploded in the last decade. It is something like a big bang explosion in databases. Particularly, the clustering of time series has attracted the interest of the researchers.

1.1. Learning Approaches

In Data mining two learning approaches are used to mine the data. They are supervised and unsupervised learning.

Supervised Learning specifies the relationship between the dependent and the explanatory variables. These relationships are then used to calculate values of the dependent variable in future data instances.

Unsupervised Learning deals all the variables in the same way. So the explanatory and the dependent variable is similar in unsupervised learning [1]. The data is investigated to find some structures in them.

1.2. Clustering Techniques:

In Clustering we split the data into groups of similar objects. Each group is known as a cluster. The intra-cluster similarity is high while inter-cluster similarity index is low. It is a very important technique in data mining. Traditionally it is seen as part of unsupervised learning. Different types of clusters as reported in the literature are [2,3,8,9,10,11,12]:

Well Separated Clusters: Every node in this type of cluster is much similar to every other node in the cluster, but different from any other node not in the cluster.

Centre-Based clusters: Every object in the cluster is more similar to the centre also called the centroid than to the centre of any other cluster.

Contiguous clusters: A node in a cluster is nearest (or more alike) to one or more other nodes in the cluster as compared to any node that is not in the cluster.

Density based clusters: A cluster is a thick region of points, which is separated by according to the low-density regions, from other regions that is of high density

Conceptual clusters: A conceptual cluster shares some common feature, or indicates a particular thought.

1.3. Use of Clustering and Methods

Clustering has wide applications in Image Processing, Document Classification, Pattern Recognition, Spatial Data Analysis, Economic Science and Cluster Web log data to discover similar web access patterns, etc. Various Methods of clustering have been reported in literature [4][5]:

Partitioning method: In literature different Partitioning methods reported are: K-mean method, K-Medoids method (PAM), Farthest First Traversal k-center (FFT), CLARA, CLARANS, Fuzzy K-Means, Fuzzy K-Modes, K-Modes, Squeezer, K-prototypes and OOLCAT etc.

Hierarchical Methods: Agglomerative Nesting (AGNES), Divisive Analysis (DIANA), Clustering using Representatives (CURE), Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) are some of the hierarchical methods.

Grid Based: Some of the Grid based clustering methods are STING, Wave Cluster, CLIQUE and MAFIA.

Density Based Methods: Density based clustering methods include DBSCAN, GDBSCANS, OPTICS, DBCLASD and DENCLUE.

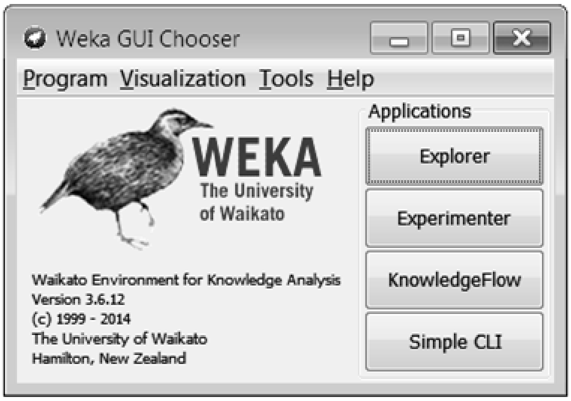
Model Based method: Model based methods are divided into two approaches: Statistical approach includes AutoClass method while Neural Network Approach includes Competitive learning and Self-organizing feature maps.

Quartile Clustering [6]: Business and technical communities across organizations are using many of the data mining packages to unearth meaningful information from structured and unstructured data. The most important observation is that these widely used packages do not include all of the algorithms. Some reasons for this might be:

1. Implementation challenges associated with few of the algorithms
2. It is not standard enough so that we can apply it across different application domain
3. The clusters generated are not intuitive to business analysts.

1.4. WEKA:

WEKA is a data mining software workbench developed by the University of Waikato in 1997 that implements data mining algorithms using Java. WEKA is a modern facility which can be used for developing machine learning (ML) methods to apply on data mining problems of real-world nature. It is a set of algorithms for machine learning for data mining tasks. These algorithms are employed directly to a single database table called a dataset. WEKA implements algorithms for pre-processing of data, its classification, clustering, regression and association rules; it also includes visualization tools. This package can also develop new machine learning schemes. WEKA is open source software provided under the GNU General Public License [7]. Its GUI for manipulating data files and visualizing results is displayed below in figure 1.1



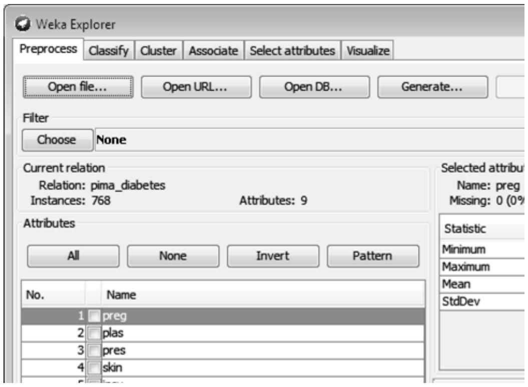
Waikato Environment for Knowledge Analysis
Version 3.6.12
(c) 1999 - 2014
The University of Waikato
Hamilton, New Zealand

Figure 1.1 GUI of WEKA

2. DATA SET

For analyzing the density-based clustering techniques in data mining, four data sets were selected with different size of instances and attributes. The first data set diabetes.arff[13] consists of 768 instances and 9 attributes; the second data set soyabeen.arff[13] contains 683 instances and 36 attributes and the third data set vote.arff[13] contains 435 instances and 17 attributes whereas fourth data set BC.arff[14] (Breast Cancer Data) contains 286 instances and 10 attributes. I will apply a density-based clustering algorithm available in WEKA i.e. DBSCAN and predict a result for the new users and new researchers.

1.5. Density-based Clustering Methods:



None
Relation: pima_diabetes
Instances: 768
Attributes: 9
Name: preg
Missing: 0 (0%

No.	Name
1	preg
2	plas
3	pres
4	skin

Figure 1.2 View of WEKA Explorer

DBSCAN (Density Based Spatial Clustering of Application with Noise) is a clustering algorithm based on density. It is using the concept of “density reachability” and “density connect ability”, both of which depends upon input parameter- size of epsilon neighbourhood e and minimum terms of local distribution of nearest neighbours. Here parameter e controls the size of the neighbourhood and size of clusters. It begins with a random starting point that has not been visited. DBSCAN algorithm is an important part of clustering technique which is mainly used in scientific literature. Density is calculated by the number of objects which are nearest to the cluster.

Density-based clustering has been shown in Figure 2.1. Here minpts=3. The core points are all the points except B, C and N. As at least three points are surrounding it in the radius of size eps. As they are all reachable from each other, they have been clustered as similar points.

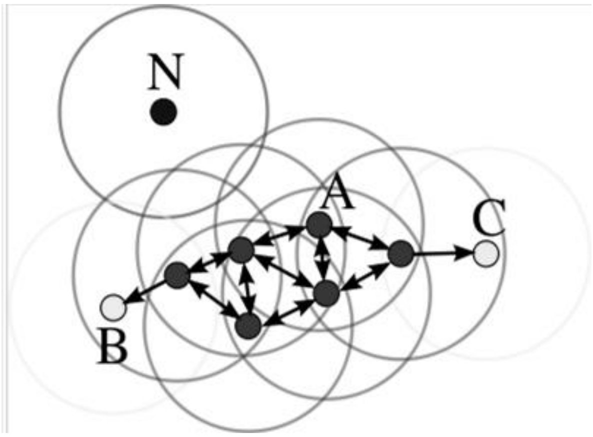


Figure 2.1 Density based clustering

B and C are not considered as core points because at least three points are not surrounding them, but they are reachable from other core points so they have also been included in the same cluster. While N is neither density reachable nor a core point. So it is just a noise point.

OPTICS (Ordering Points to Identify Clustering Structure) is an algorithm for spatial data to find density based clusters. DBSCAN burdens the user from choosing the input parameters. Moreover, different parts of the data could require different parameters. One of the DBSCAN’S major weaknesses i.e. of detecting useful clusters in data of changing density has been overcome by this algorithm.

Steps involved in DBSCAN [2]

- Random select a point s
- Get back all the points density-reachable from s with respect to *Eps* and *MinPointts*.
- If point s is a core object, a cluster is formed.
- If point s is a border object, no nodes are density-reachable from s and DBSCAN visits the next node of the data base.
- The process continues unless all of the nodes are processed.

Core Object: A node with at least MinPoints nodes within a radius ‘Eps-neighborhood’

Border Object: A point that on the border of a cluster.

1.6. Pros and Cons of Density-Based Algorithm

The main advantage of density-based clustering algorithms is that they do not require theoretical specification and able to identify noisy data while clustering. It does not work well in case of data having high dimensionality [2].

3. EXPERIMENTAL RESULTS

Here I used the selected datasets in WEKA version 3.6.12 and tabulated the results in table 3.1 and table 3.2. According to the results obtained it is obvious that the default values Eps and Minpts is not always providing better results in clustering the data. The user has to carefully select the values of epsilon and minpoints. It is also observed that if the dimensionality of data is high the results are not so much accurate. In table 3.1, the data sets used were diabetes.arff and soyabean.arff. The results of the clustering the data using WEKA tool’s DBSCAN facility reveals that if the default values of Eps and Minpts are used the Diabetes data set is clustered into 1 cluster i.e., tested_negative. So 268 nodes were clustered incorrectly. Whereas on using Eps=0.32 and minpts=10 the results are better. Same is observed with soyabean.arff data set. It is not clustered and produces error on default values. So on changing the values 18% nodes were incorrectly clustered.

Table 3.1: Experimental data of two data sets Diabetes and Soyabean

	Diabetes		Soya bean	
Instances	768		683	
Attributes	9		36	
Eps	0.9	0.32	0.9	2.0
Minpts	6	10	6	5
No of Clusters	1	2	No Clusters Error generated	18
Elapsed time sec	0.42	0.4		0.75
Not Clustered Instances	0	54		185
Incorrectly clustered	268 (34.90%)	240 (31.25 %)		123 (18%)
Clusters (few)	Tested_negative 768	Tested_negative 694 Tested_positive 20		Charcoal-rot 20 Downy-mildew 5 Alternarialleaf-spot 167
Class count in relation	Tested_negative 500 Tested_positive 268		Charcoal-rot 20 downy-mildew 20 Alternarialleaf-spot 91	

In Table 3.2, the experimental results of other two data sets have been tubulised. The default values for the two data sets either do not produce the clusters or a great number of nodes are incorrectly clustered. Whereas upon selecting the suitable Eps and Minpts values, the results are far better.

Table 3.2: Experimental Data of two datasets Vote and Breast Cancer

	Vote		Breast Cancer	
Instances	435		286	
Attributes	17		10	
Eps	0.9	1.1	0.9	1.2
Minpts	6	7	6	4
No of Clusters	14	2	No Clusters Error generated	3
Elapsed time sec	0.12	0.15		0.05
Not Clustered Instances	313	160		131
Incorrectly clustered	95 (21.8%)	6 (1.4%)		41 (14.33%)
Clusters (few)	Republican 13 Democrate 14	Republican 132 Democrate 143		No_recurrence_events 141 Recurrence_events 05
Class count in relation	Republican 168 Democrate 267		No_recurrence_events 201 Recurrence_events 85	

4. CONCLUSION

Data mining is covering every field of our life. Mainly we are using it in education, business, image processing and banking etc. In this paper, I have provided an overview of the renowned density-based clustering algorithm on different datasets with WEKA tools. The results show that DBSCAN overloads the user form choosing the input parameters carefully for proper clustering otherwise the results obtained are not correct. Moreover the default set of parameters are not always useful for clustering data into meaningful groups.

5. REFERENCES

1. Thinaharan, N., & Vetriselvi, P. An Overview of Clustering Techniques in Data Mining. *International Journal of Innovative Research in Computer and Communication Engineering , 3.2*(2015)., 1269-1275.
2. K.Kameshwanran, & Malarvizhi, K. Survey on Clustering Techniques in Data Mining. *International Journal of Computer Science and Information Technologies , 5* .2 (2014), 2272-2276.
3. Junaid, S., & Bhosle, K. Overview of Clustering Techniques. *International Journal of Advanced Research in Computer Science and Software Engineering , 4.11 (2014)*, 621-624.
4. Suman, & Mittal, M. P. Comparison and Analysis of Various Clustering Methods in Data mining On

Education data set Using the Weka tool. *International Journal of Emerging Trends and Technology in Computer Science , 3.2 (*2014), 240-244.

5. Purviya, R., Tiwari, H., & Mishra, S. Application of Clustering Data Mining Techniques In Temporal Data Sets of Hydeology: A Review. *International Journal of Scientific Engineering and Technology , 3.4* (2014), 359-363.
6. Goswami, S., & A.Chakrabarti, D. Quartile Clustering: A quartile based technique for Generating Meaningful Clusters. *Journal of Computing , 4.2* (2014), 48-55.
7. Waikato, T. U. *Weka 3: Data Mining Software in Java*. Retrieved April 09, 2015, from Machine Learning Group at the University of Waikato: http://www.cs.waikato.ac.nz/ml/weka/
8. Gondaliya, B. REVIEW PAPER ON CLUSTERING TECHNIQUES. International Journal of Engineering Technology, Management and Applied Sciences , 2.7 (2014), 234-237.
9. Sharaf Ansari, S. C. An Overview of Clustering Analysis Techniques use in Data Mining. International Journal of Emerging Technology and Advanced Engineering , 3.12 (2014), 284-286.
10. Dr. Sudhir B. Jagtap, D. K. Census Data Mining and Data Analysis using WEKA. International Conference in "Emerging Trends in Science, (2013) Technology and Management (pp. 35-40). Singapore: Society for Mathematical Development, India.
11. Mohamad Saraee, Najmeh Ahmadian, Zahra Narimani. "Data Mining Process Using Clustering: A Survey." Iran Data Mining Conference 2007. Tehran: Amir Kabir University, 2007. 1-6.
12. Shu-Hsien Liao ., Pei-Hui Chu, Pei-Yuan Hsiao. "Data mining techniques and applications – A decade review from 2000 to 2011." Expert Systems with Applications 39 (2012): 11303-11311.
13. Diabetes, soyabean and Vote datasets, Gary M. Weiss, Ph.D., http://storm.cis.fordham.edu/~gweiss /data-mining/datasets.html, retrieved on 09 April 2015.
14. Breast Cancer Dataset, Håkan Kjellerstrand, http://www.hakank.org/weka/BC.arff, retrieved on 09 April 2015.