Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Implementation of Density Based Clustering Algorithm for ILPD Dataset

Prof. M.S. Prasad Babu, K. Swapna, V.Gayathri
Dept. of CS & SE, Andhra University,
Andhra Pradesh, India

Abstract— Density Based clustering is a process of clustering which improves the capabilities of dataset by dividing the data into no of clusters based on the attributes and the selection. Here DBSCAN algorithm is used to cluster the ILPD dataset. ILPD data set contains the patient data who are suffering from different types of liver diseases. An expert system is developed based on the obtained cluster data on liver diagnosis system which is more user-friendly can obtain better results. Introduction merely deals with DBSCAN algorithm and ILPD dataset.

Keywords: Data mining, clustering, DBSCAN, ILPD, Expert system

I. INTRODUCTION

Knowledge Discovery in Databases (KDD) was the term coined by Gregory Piatetsky-Shapirino 1989 in connection with his studies on large databases. But most press liked the term "data mining" for the same. It involves the integration of techniques from multiple disciplines such as statistics, artificial intelligence, machine learning, neural networks and pattern reorganization etc. It predicts future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. Classification, Clustering, Regression and Association rules are some of the important techniques used in Data mining. Classification is the process of extracting and categorising data into models to predict the future data trends. In the literature various classifiers are used to build predefined classes or concepts. Regression is used to fit an equation to a dataset to predict or forecast the unknown values. Association rules deals with identifying the objects by uncovering relationship between seemingly unrelated data in databases. Clustering is the process of grouping the objects into clusters with more similarity. Clustering algorithms have been used in a large variety of applications [Jain and Dubes 1988; Rasmussen 1992;Oehler and Gray 1995; Fisher et al.1993][3]. Partitioning , hierarchical, density methods are some of the methods used in clustering. DBSCAN algorithm was one of the important density based algorithm, used to find spherical-shaped clusters with some similarity. ILPD dataset, a dataset collected, analyzed and donated to UCI Machine learning repository by the author [8],[9],[10], is a dataset consisting different attributes for liver patients. In this paper DBSCAN algorithm is implemented on the ILPD Dataset to identify different clusters with some similarities.

II. DBSCAN ALGORITHM

DBSCAN algorithm (Density-based spatial clustering of applications with noise algorithm) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996[]. Finding clusters of arbitrary shape such as the "S" shape and oval shape clusters is very difficult task since these algorithms groups the objects according to the specific density objective functions. Density is usually defined as the number of objects in a particular neighbourhood of a data objects[2]. Figure.1 illustrates clusters of data into arbitrary shapes may be in accurately identifying convex regions, where noise or outliers are included in the clusters.

Fig:1. Clusters of arbitrary shape.

Here a user-specified parameter $\epsilon > 0$ is used to specify the radius of a neighborhood under consideration for every object. The ϵ-neighborhood of an object o is the space within a radius ϵ centered at o. The density of a neighborhood can be measured simply by the number of objects in the neighborhood. To determine whether a neighborhood is dense or not, DBSCAN uses another user-specified parameter, *MinPts*, which specifies the density threshold of dense regions. An object is a core object if the ϵ-neighborhood of the object contains at least *Min Pts* objects. Core objects are the pillars of dense regions. Given a set, *D*, of objects, all core objects can be identified with respect to the given parameters, ϵ and *Min Pts*. The clustering task is therein reduced to identifying the points in dense regions, using core objects and their

neighborhoods to form dense regions. For a core object *q* and an object *p*, *p* is said to be directly density-reachable from *q* (with respect to ϵ and *Min Pts*) if *p* is within the ϵ-neighborhood of *q*. Clearly, an object *p* is directly density-reachable from another object *q* if and only if *q* is a core object and *p* is in the ϵ-neighborhood of *q*. Using the directly density-reachable relation, a core object can "bring" all objects from its ϵ neighborhood into a dense region. In DBSCAN, *p* is density-reachable from *q* (with respect to ϵ and *Min Pts* in *D*) if there is a chain of objects $p1, \ldots, pn$, such that $p1 = q$, $pn = p$, and $pi + 1$ is directly density-reachable from *pi* with respect to ϵ and *Min Pts*, for $1 \leq i \leq n$, $pi \in D$. Note that density-reach ability is not an equivalence relation because it is not symmetric. If both $o1$ and $o2$ are core objects and $o1$ is density-reachable from $o2$, then $o2$ is density-reachable from $o1$. However, if $o2$ is a core object but $o1$ is not, then $o1$ may be density-reachable from $o2$, but not vice versa. To connect core objects as well as their neighbors in a dense region, DBSCAN uses the notion of density-connectedness. Two objects $p1, p2 \in D$ are density-connected with respect to ϵ and *MinPts* if there is an object $q \in D$ such that both $p1$ and $p2$ are density-reachable from *q* with respect to ϵ and *MinPts*. Unlike density-reach ability, density-connectedness is an equivalence relation. It is easy to show that, for objects $o1$, $o2$, and $o3$, if $o1$ and $o2$ are density-connected, and $o2$ and $o3$ are density-connected, then so are $o1$ and $o3$. DBSCAN works with noise resistant for handling clusters of different shapes and sizes [5]. It requires that the density in a neighbourhood for an object should be high enough if it belongs to a cluster. It creates a new cluster from a data object by absorbing all objects in its neighbourhood [4]

DBSCAN algorithm:

```
Input: DBSCAN (Set of Points, Eps, Min Pts) // Set Of Points is UNCLASSIFIED
Cluster Id := next Id(NOISE);
FOR i FROM 1 TO Set Of Points .size DO
Point: = Set Of Points.get (i);
IF Point. C  Id = UNCLASSIFIED THEN
IF Expand Cluster (Set Of Points, Point,
Cluster Id, Eps, Min Pts) THEN
Cluster Id := next Id(Cluster Id)
END IF
END IF
END FOR
END; // DBSCAN
```

Complexity: If a spatial index is used, the computational complexity of DBSCAN is $O(n \log n)$, where *n* is the number of database objects. Otherwise, the complexity is $O(n2)$. With appropriate settings of the user-defined parameters, ϵ and *Min Pts*, the algorithm is effective in finding arbitrary-shaped clusters.

2.1 Indian Liver Patient Dataset (ILPD)

A. Liver disease:

Liver is an important organ, compared to the other organs, in a human body. When the liver becomes diseased, it may have many serious consequences. Liver disease (also called hepatic disease) is abroad term describing number of diseases affecting the liver. Many are accompanied by Jaundice caused by increased levels of bilirubin in the system. The bilirubin results from the breakup of the haemoglobin of dead red blood cells; normally, the liver removes bilirubin from the blood and excretes it through bile [12].

B. Description of ILPD

The Indian liver patient dataset (ILPD) dataset contains 583 liver records with 10 attributes that are eight simple blood tests usually referred as liver function tests, age, and gender and group label [6]. The present study is conducted on clustering the ILPD dataset. In this dataset the liver function tests are total bilirubin, direct bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and alkphos. This data set contains 416 liver patient records and 167 non liver patient records. The attributes are simple blood tests used to measure the levels of enzymes, proteins and bilirubin levels in the blood that help us to detect the liver damage. Proteins are large molecules that are needed for the overall health. Enzymes are protein cells that help important chemical reactions to occur in the body. Bilirubin helps the body break down and digest fats. ALT (SGPT), AST (SGOT), ALP and GGT are enzymes made by the liver. The ALT, AST, ALP, and GGT are liver enzymes tests that measure the level of ALT, AST, ALP, and GGT in blood respectively**.** High levels of ALT and AST in the blood can be a sign of liver damage. High levels of ALP and GGT can be sign of bile duct damage.

Table 1: ILPD dataset attributes

Attribute	Type
Gender	Categorical
Age	Real number
TB	Real number
DB	Real number
TP	Real number
ALB	Real number

A /G Ratio	Real number
SGPT	Integer
SGOP	Integer
Alkphos	Integer

ILPD dataset attribute table from [8].The description of ILPD dataset attributes is represented below.

Attribute	**Attribute Information**
Age	Age of patient
Gender	Gender of patient
TB	Total Bilirubin
DB	Direct Bilirubin
Alkphos	Alkaline Phosphotase
SGPT (ALT)	Serum GlutamicPyruvic Transaminase
SGOP (AST)	Serum Glutamic Oxaloacetic Transaminase
TP	Total Protein
ALB	Albumin
A/G Ratio	Albumin and Globulin Ratio

Table 2: Sample ILPD dataset

AGE	GENDER	TOTAL_ BILIRUBIN	DIRECT_ BILIRUBIN	ALKPHOS	SGPT	SGOT	TOTAL_ PROTIENS	SERUM_ ALBUMIN	AG_ RATIO
62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74
62	Male	7.3	4.1	490	60	68	7	3.3	0.89
58	Male	1	0.4	182	14	20	6.8	3.4	1
72	Male	3.9	2	195	27	59	7.3	2.4	0.4
46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3
26	Female	0.9	0.2	154	16	12	7	3.5	1
29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1
17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2
55	Male	0.7	0.2	290	53	58	6.8	3.4	1
57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8
72	Male	2.7	1.3	260	31	56	7.4	3	0.6
64	Male	0.9	0.3	310	61	58	7	3.4	0.9
74	Female	1.1	0.4	214	22	30	8.1	4.1	1
61	Male	0.7	0.2	145	53	41	5.8	2.7	0.87

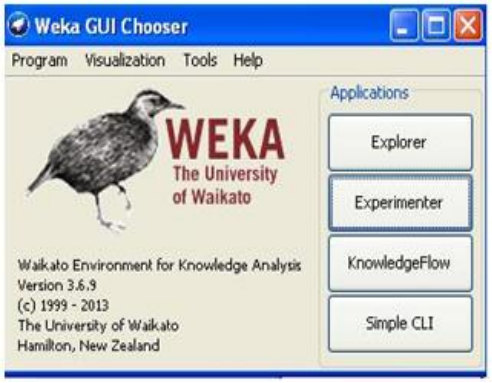
2.2. Data Mining Tool for Clustering:

WEKA is the one of the data mining tool that can be implemented to various classifications, clustering, association, regression algorithms. In present paper WEKA tool is used and implement DBSCAN clustering algorithm for ILPD data set.

III. IMPLEMENTATION OF DBSCAN ALGORITHM TO ILPD DATASET

Weka is a collection of machine learning algorithms for data mining. The algorithms can either be applied directly to a data set or called from our own java code. Weka contains tools for data pre-processing, classification, Regression, Clustering, Association Rules and Visualization. It is also well suited for developing new machine learning schemes.

Launching weka: the weka GUI Chooser provides a starting point for launching weka's main GUI applications and supporting tools The GUI chooser consists of four buttons –one for each of the four major weka applications and four menus



The buttons can be used to start the following applications:
1. Explorer: An environment for exploring data with WEKA
2. Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes.
Fig:2. WEKA GUI chooser
3. Knowledge Flow This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.
4. Simple CLI: Provides a simple command line interface that allows direct Execution of WEKA commands for operating systems that do not provide their own command line interface.

3.1 IMPLEMENTION:

Steps for Executing DBSCAN Clustering on ILPD dataset

3.1.1 Step1: Download the ILPD dataset which is in .CSV (Comma Separated Format) and convert it into .ARFF (attribute relation file format) which is best suitable for clustering task is shown in following figure.3.

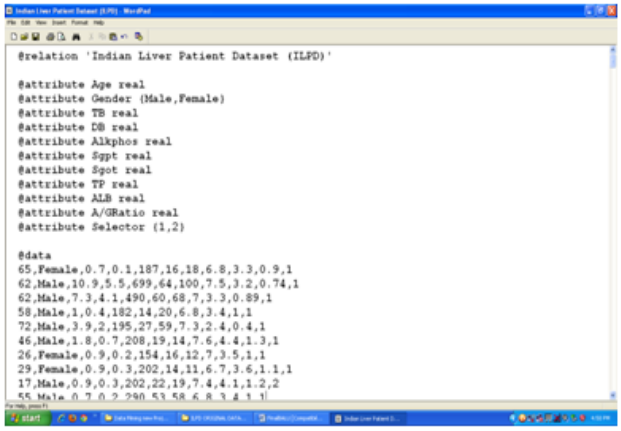


Fig:3. Conversion of .CSV format to .ARFF file format

3.1.2 Step2: Load the ILPD dataset with .ARFF file in the weka explorer GUI and remove the selector attribute in the existing data is shown in following figure:4.

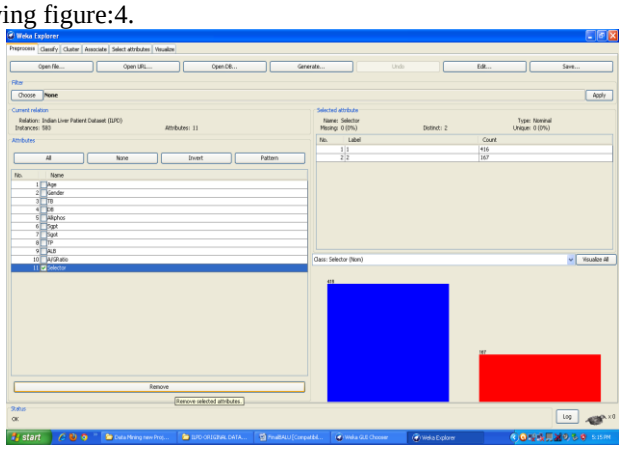


Fig :4. Remove selector attribute for existing data

3.1.3 Step3: Applying clustering algorithms, select the desired DBSCAN clustering technique from the available algorithms shown in the following figure :5.

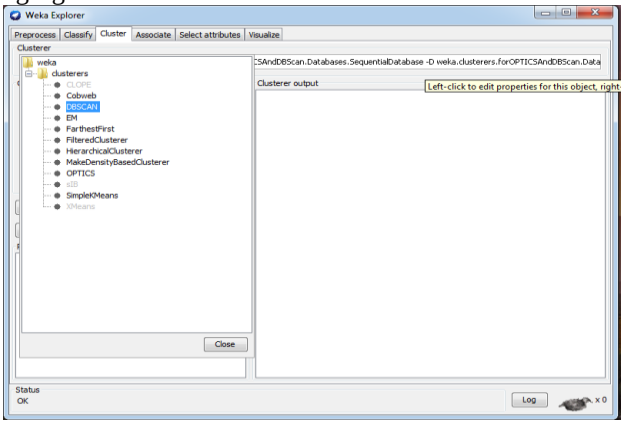


Fig :5. Select the desired DBSCAN clustering algorithm

3.1.4. Step4: Setting the clustering parameters, database _type, datadase_distance type, epsilon, minpoints
OPTIONS
data base_ Type -- used database
data base_ distance Type -- used distance-type
epsilon -- radius of the epsilon-range-queries
min Points -- minimum number of Data Objects required in an epsilon-range-query shown in the following figure :6

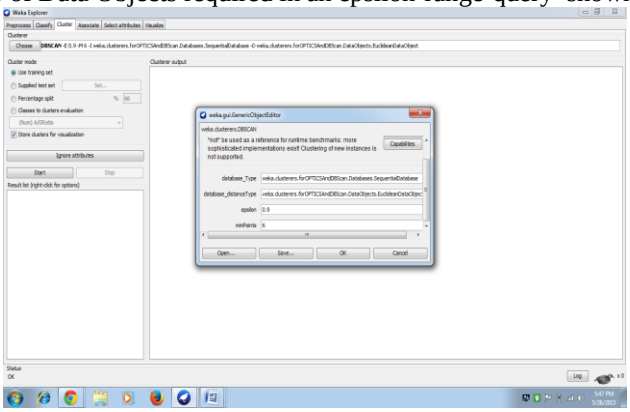


Fig :6. Select the parameters for DBSCAN clustering.

Step5: Start execution to cluster the ILPD dataset is shown in the following figure.7

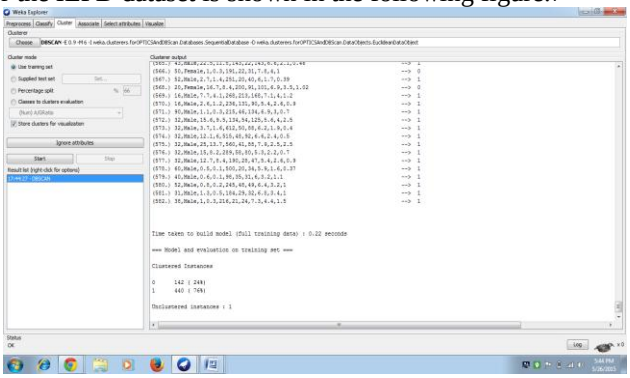


Fig: 7. Start execution to cluster the ILPD dataset

IV. RESULTS AND DISCUSSIONS

4.1 Comparison of Dataset Result:

The results generated using existing ILPD classified dataset alone and the results generated using DBSCAN clustering applied on ILPD dataset the fig 8.1 and 8.2 respectively

Table :3. Output Screen DBSCAN Clustering | Output Screen actual Clustering

0,65,Female,0.7,0.1,187,16,18,6.8,3.3,0.9,**cluster0**	65,Female,0.7,0.1,187,16,18,6.8,3.3,0.9,**1**
1,62,Male,10.9,5.5,699,64,100,7.5,3.2,0.74,**cluster1**	62,Male,10.9,5.5,699,64,100,7.5,3.2,0.74,**1**
2,62,Male,7.3,4.1,490,60,68,7,3.3,0.89,**cluster1**	62,Male,7.3,4.1,490,60,68,7,3.3,0.89,**1**
3,58,Male,1,0.4,182,14,20,6.8,3.4,1,**cluster1**	58,Male,1,0.4,182,14,20,6.8,3.4,1,**1**
4,72,Male,3.9,2,195,27,59,7.3,2.4,0.4,**cluster1**	72,Male,3.9,2,195,27,59,7.3,2.4,0.4,**1**
5,46,Male,1.8,0.7,208,19,14,7.6,4.4,1.3,**cluster1**	46,Male,1.8,0.7,208,19,14,7.6,4.4,1.3,**1**
7,29,Female,0.9,0.3,202,14,11,6.7,3.6,1.1,**cluster0**	26,Female,0.9,0.2,154,16,12,7,3.5,1,1
8,17,Male,0.9,0.3,202,22,19,7.4,4.1,1.2,**cluster1**	29,Female,0.9,0.3,202,14,11,6.7,3.6,1.1,**1**
9,55,Male,0.7,0.2,290,53,58,6.8,3.4,1,**cluster1**	17,Male,0.9,0.3,202,22,19,7.4,4.1,1.2,**2**
10,57,Male,0.6,0.1,210,51,59,5.9,2.7,0.8,**cluster1**	55,Male,0.7,0.2,290,53,58,6.8,3.4,1,**1**
11,72,Male,2.7,1.3,260,31,56,7.4,3,0.6,**cluster1**	57,Male,0.6,0.1,210,51,59,5.9,2.7,0.8,**1**
12,64,Male,0.9,0.3,310,61,58,7,3.4,0.9,**cluster1**	72,Male,2.7,1.3,260,31,56,7.4,3,0.6,**1**
13,74,Female,1.1,0.4,214,22,30,8.1,4.1,1,**cluster0**	64,Male,0.9,0.3,310,61,58,7,3.4,0.9,**2**
14,61,Male,0.7,0.2,145,53,41,5.8,2.7,0.87,**cluster1**	74,Female,1.1,0.4,214,22,30,8.1,4.1,1,**1**
15,25,Male,0.6,0.1,183,91,53,5.5,2.3,0.7,**cluster1**	61,Male,0.7,0.2,145,53,41,5.8,2.7,0.87,**1**
16,38,Male,1.8,0.8,342,168,441,7.6,4.4,1.3,**cluster1**	25,Male,0.6,0.1,183,91,53,5.5,2.3,0.7,**2**
17,33,Male,1.6,0.5,165,15,23,7.3,3.5,0.92,**cluster1**	38,Male,1.8,0.8,342,168,441,7.6,4.4,1.3,**1**
18,40,Female,0.9,0.3,293,232,245,6.8,3.1,0.8,**cluster0**	33,Male,1.6,0.5,165,15,23,7.3,3.5,0.92,**2**
19,40,Female,0.9,0.3,293,232,245,6.8,3.1,0.8,**cluster0**	40,Female,0.9,0.3,293,232,245,6.8,3.1,0.8,**1**
20,51,Male,2.2,1,610,17,28,7.3,2.6,0.55,**cluster1**	40,Female,0.9,0.3,293,232,245,6.8,3.1,0.8,**1**
Cluster Label:1-Liverpatient,0-NonLiver patient.	**Class Label:1-Liverpatient,2-Non-Liver patient**

Here Actual data set contain 416 liver patient and 167 non liver patients where as DBSCAN algorithm gives 439 liver patients and 144 non liver patients.To identify the Performance of the proposed DBSCAN clustered ILPD dataset with actual values. DBSCAN clustered ILPD dataset with existing classified ILPD dataset in all iterations are represented in the form of graph as below

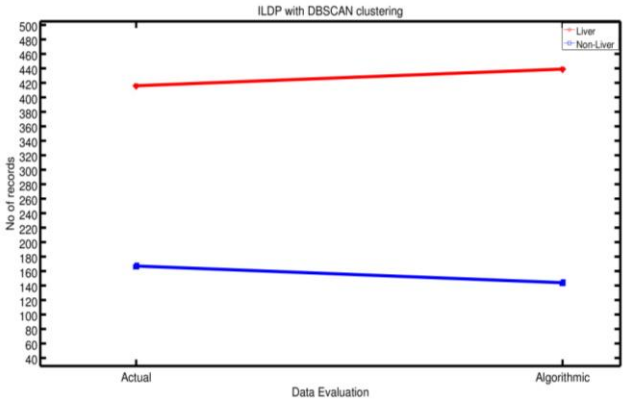


Fig: 8. Graph describing the performance of actual result with algorithm result

Based on the above graph It is concluded that the performance of proposed DBSCAN clustered result on ILPD data using classification, is compared with existing ILPD .The obtained knowledge base for proposed DBSCAN clustering is shown in the above fig.7, The result is 64.2% to the existing ILPD dataset .

4.2 .Experimental Results:

Weka is a data mining tool. It provides so many algorithms in various areas in data mining. Like classification, clustering, association, regression. In this paper we use three clustering algorithms and compare it those algorithms in ILPD data set with three parameters accuracy, time and cluster distance. Those results are mention in below table.

Table .4: Comparative study of ILPD data set with three algorithms

Algorithms	No cluster	Accuracy	Run time in(Seconds)	Cluster Distance
k-mean	2	64.2% 64.2%	0.02 0.03	Euclidean Manhattan
Hierarchical Agglomerative (single link)	2	64.2% 71.2%	2.34 2.43	Euclidean Manhattan
Agglomerative (avgas link)	2	71.2% 71.2%	0.79 0.94	Euclidean Manhattan
Agglomerative (complete link)	2	64.2% 68.2%	0.82 0.74	Euclidean Manhattan
DBSCAN	Epsilon=0.9 Min points=6	64.2%	0.24	Euclidean

4.3 ILPD with K-means, Hierarchical, DBSCAN algorithms-Graphs

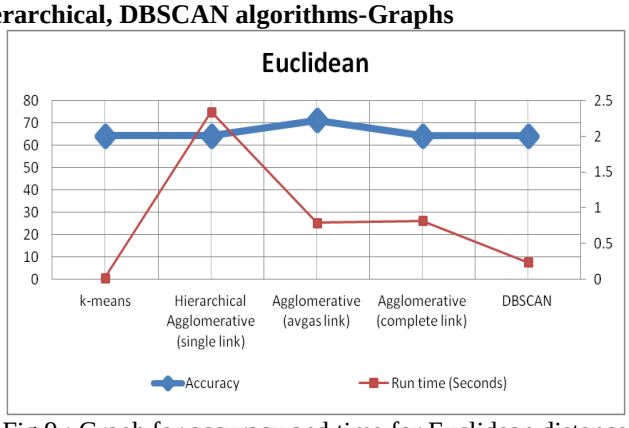


Fig.9 : Graph for accuracy and time for Euclidean distance

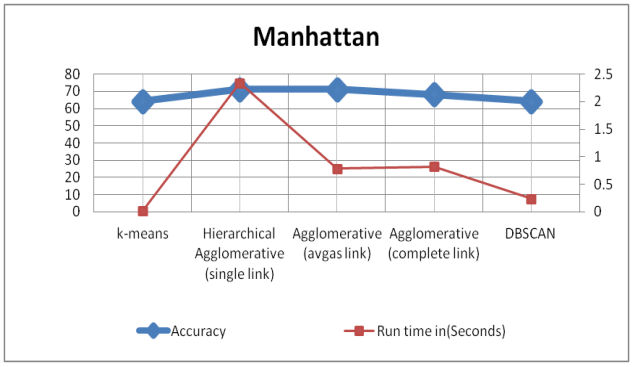


Fig.10 : Graph for accuracy and time for Manhattan distance.

V. CONCLUSIONS & FUTURE SCOPE

The performance evaluation is conducted by comparing the performance of DBSCAN algorithm with other clustering algorithms i.e. k-means, hierarchical algorithms with respect to the performance parameters: accuracy, time and distance. It was found that DBSCAN algorithm gave 90.12 sec less run time than , hierarchical algorithms with Euclidian distances. But 9.83% less accuracy in k-means , hierarchical algorithms. Therefore it may be concluded that DBSCAN is good for time parameter. Generally density based algorithms are used for large datasets where as ILPD data set is a small dataset with 583 records only. Therefore if the size of the data set is increased then DBSCAN algorithm gives better results.

REFERENCES

[1] Jiawei Han and Micheline Kamber *Data Mining: Concepts andTechniques*,3rd edition.

[2] pradeep rai, Shubha Singh "*A Survey of Clustering Techniques*" International Journal of Computer Applications (0975 – 8887)Volume 7– No.12, October 2010

[3] A.K. JAIN, M.N. MURTY, AND P.J. FLYNN ""*Data Clustering: A Review*" ACM Computing Surveys, Vol. 31, No. 3, September 1999

[4] Rui Xu, Student Member, *IEEE* and Donald Wunsch II, Fellow, IEEE *"' Survey of Clustering algorithms*''IEEE transactions on neural networks, vol. 16, no. 3, may 2005.

[5] Margaret H. Dunham S.sridhar "*Data mining introductory and advanced topics*"

[6] The Indian liver patient dataset (ILPD)is downloaded from UCI machine repository in the area of life science. The PD data set is available in following hyper link http://archive.ics.uci.edu/datasets/ILPD+(indian+liver+patient+Dataset)

[7] Weka 3.4.10 jre: *data mining with open source machine learning* software 2002-2005 David Scuse and university of Waikato.

[8] "A *Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis*", Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu, Prof. N. B. Venkateswarlu ,IJCSI International Journal of Computer Science Issue vol. 9,Issue 3, No 2,May 2012

[9] . "A *Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis*", Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu, Prof. N. B. Venkateswarlu ,International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011

[10] A Critical Evaluation of Bayesian Classifier for Liver Diseases Diagnosis Using Bagging and Boosting Methods B .Venkataramana, Prof.M.Surendra Prasad Babu, Prof.N.B. Venkateswarlu, International Journal of Engineering Science and Technology 3 (4), 3422-3426

[11] . "A *Density-Based Algorithm for Discovering Clustersin Large Spatial Databases with Noise*" Martin Ester, Hans-Peter Kriegel, Jorgen Sander, Xiaowei Xu, Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)

[12] *Analysis of Liver Disorder Using Data mining Algorithm*, P.Rajeswari1, G.Sophia Reena, Global Journal of Computer Science and Technology, Vol. 10 Issue 14 (Ver. 1.0) November 2010.

