Study of Simple K-Means and DBSCAN Clustering Algorithm Using Hypothyroid Dataset

[1]Abhaya Kumar Samal, [2]Subhendu Kumar Pani, [3]Jitendra Pramanik

[1]Professor, [2]Asso. Professor, [3]Research Scholar
[1]Dept. of Computer Science & Engineering,
[1]Trident Academy of Technology, Bhubaneswar, India

***Abstract*— In data mining and knowledge discovery, data clustering is considered as one of the major basic research topics that put similar data into groups. A clustering algorithm divides a data set into several groups based on the standard rule of maximizing the intra-class similarity and minimizing the inter-class similarity. In this work, we analyze the performance of K-Means, and Density based clustering. Performance of the 2 techniques are presented and compared using a clustering tool WEKA. These 2 clustering algorithms have applied on hypothyroid dataset which is available at UCI machine learning repository.**

***Index Terms*— K-means algorithms, Data mining algorithms, Weka tools, Density based clustering algorithm**

I. INTRODUCTION

Data mining is considered to be an emerging technology that has made revolutionary change in the information world. The term `data mining' (often called as knowledge discovery) refers to the process of analyzing data from different perspectives and summarizing it into useful information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system. Technically, "data mining is the process of finding correlations or patterns among dozens of fields in large relational databases". Therefore, data mining consists of major functional elements that transform data onto data warehouse, manage data in a multidimensional database, facilitates data access to information professionals or analysts, analyses data using application tools and techniques, and meaningfully presents data to provide useful information.

According to the Gartner Group, "data mining is the process of discovering meaningful new correlation patterns and trends by sifting through large amount of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques" [3] . Thus use of data mining technique has to be domain specific and depends on the area of application that requires a relevant as well as high quality data. More precisely, data mining refers to the process of analysing data in order to determine patterns and their relationships. It automates and simplifies the overall statistical process, from data source(s) to model application. Practically analytical techniques used in data mining include statistical methods and mathematical modeling. However, data mining and knowledge discovery is a rapidly growing area of research and application that builds on techniques and theories from many fields, including statistics, databases, pattern recognition, data visualization, data warehousing and OLAP, optimization, and high performance computing [1]. Worthy to mention that online analytical processing (OLAP) is quite different from data mining, though it provides a very good view of what is happening but cannot predict what will happen in the future or why it is happening. In fact, blind applications of algorithms are not also data mining. In particular, ``data mining is a user-centric interactive process that leverages analysis technologies and computing power, or a group of techniques that find relationships that have not previously been discovered" [2, 4]. So, data mining can be considered as a convergence of three technologies -- viz. increased computing power, improved data collection and management tools, and enhanced statistical algorithms.

Data and information have become major assets for most of the organizations. The success of any organisation depends largely on the extent to which the data acquired from business operations is utilised. In other words, the data serves as an input into a strategic decision making process, which could put the business ahead of its competitors. Also, in this era, where businesses are driven by the customers, having a customer database would enable management in any organisation to determine customer behaviour and preference in order to offer better services and to prevent losing them resulting better business. The data needed that will serve as an input to organizational decision-making process is generated and warehoused. It is being collected via many sources, such as the point of sales transactions, surveys, through the internet logs – cookies, etc. This has resulted in huge databases which have valuable knowledge hidden in them and may be difficult to extract.

Data mining has been identified as the technology that offers the possibilities of discovering the hidden knowledge from these accumulated databases. Techniques such as pattern recognition and classification are the most important in data mining [4, 5]. The task of recognition and classification is one of the most frequently encountered decision making problems in daily activities [6, 7]. A classification problem occurs when an object needs to be assigned into a predefined group or class based on a number of observed attributes, or features, related to that object. Humans constantly receive information in the form of patterns of interrelated facts, and have to make decisions based on them. When confronted with a pattern recognition problem, stored knowledge and past experience can be used to assist in making the correct decision. Indeed, many problems in various domains such as financial, industrial, technological, and medical sectors can be cast as classification problems. CLUSTERING is a data mining technique to

group the similar data into a cluster and dissimilar data into different clusters. I am using Weka data mining tools for clustering. Techniques and Algorithms are discussed in section 2. In section 3, we describe our study on findings. Finally the paper concludes in section 4.

II. Techniques and Algorithms

In this section, we describe various techniques and algorithms used in this research.

K-means Clustering:

The term "k-means" was first applied y by James MacQueen in 1967 [8], though the idea goes back to 1957 [9]. The standard algorithm was first developed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published until 1982. K-means is a widely used partitional clustering technique in the industries. The K-means algorithm is the most commonly used partitional clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time.

DBSCAN clustering:

Density-based clustering algorithms try to discover clusters based on density of data points in a region. The main idea of density-based clustering is that for every instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (MinPts). One of the most well-known density-based clustering algorithms is the DBSCAN [10].

III. Experimental Study and Analysis

In this section we present details of the basis and approach followed to conduct experimental study of the K-Means and DBSCAN clustering algorithm using hypothyroid dataset.

WEKA Tool:

We use WEKA (www.cs.waikato.ac.nz/ml/weka/), an open source data mining tool for our experiment. WEKA is developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art tool for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, feature reduction, classification, regression, clustering, and association rules. It also includes visualization tools. The new machine learning algorithms can be used with it and existing algorithms can also be extended with this tool.

DBSCAN Dataset Description:

Hypothyroid dataset dataset is selected for this work. Dataset contains 30 attributes, one class attribute and 3772 instances. The dataset is collected from UCI repository. A sample weka file of the dataset is depicted in Figure 3.1.

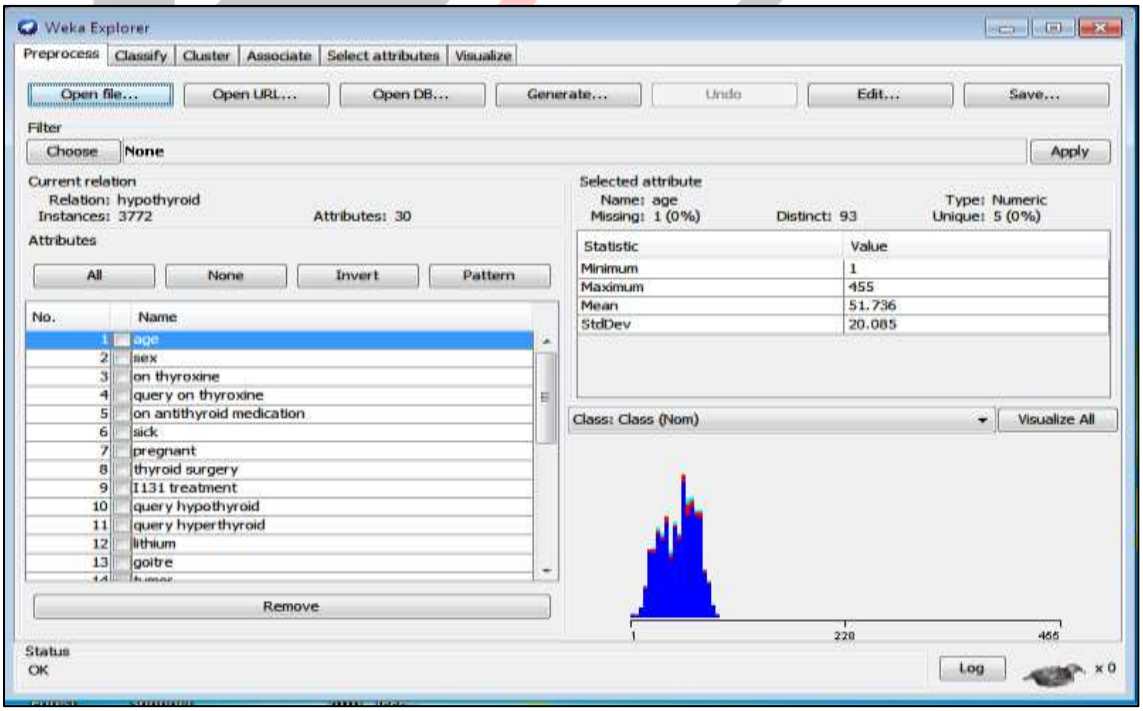

Figure 3.1: A sample weka file

Results Analysis:

Above section involves the study of each of the three techniques introduced previously using Weka Clustering Tool on a hypothyroid data. Clustering of the data set is done with each of the clustering algorithm using Weka tool and the results are shown in Table1.

Table 1: Comparison Result of Clustering Algorithms

Name	No. of clusters	Cluster Instances	Time taken to build model	Un-clustered Instances
Simple K-means	02	0:892(24%) 1:2880(76%)	0.47 Sec	0
DBSCAN	80	5:481(15%) 3:399(12%) 16:298(9%)	14.21	545

Figure 3.2 shows Visualizing Cluster assignments using Simple K-Means, Figure 3.3 shows Visualizing Cluster assignments using DBSCAN and the Performance Graph of Simple K-Means and DBSCAN algorithms are shown in Fig 3.4.

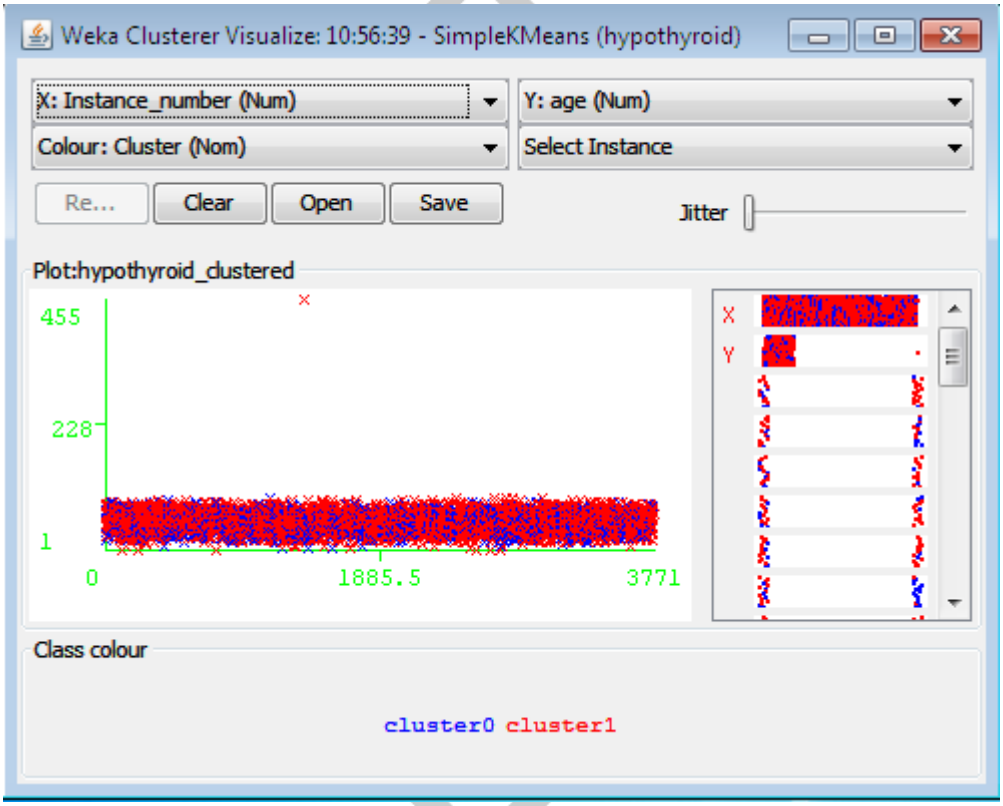

Figure 3.2Visualizing Cluster assignments using Simple K-Means

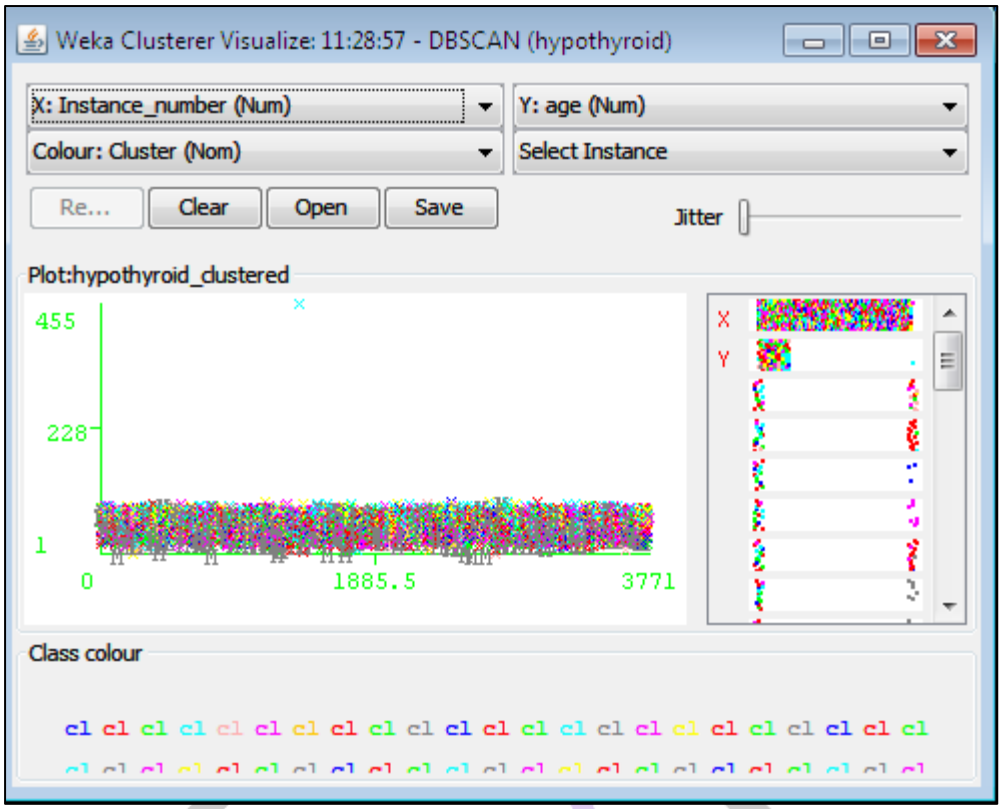

Figure 3.3: Visualizing Cluster assignments using DBSCAN

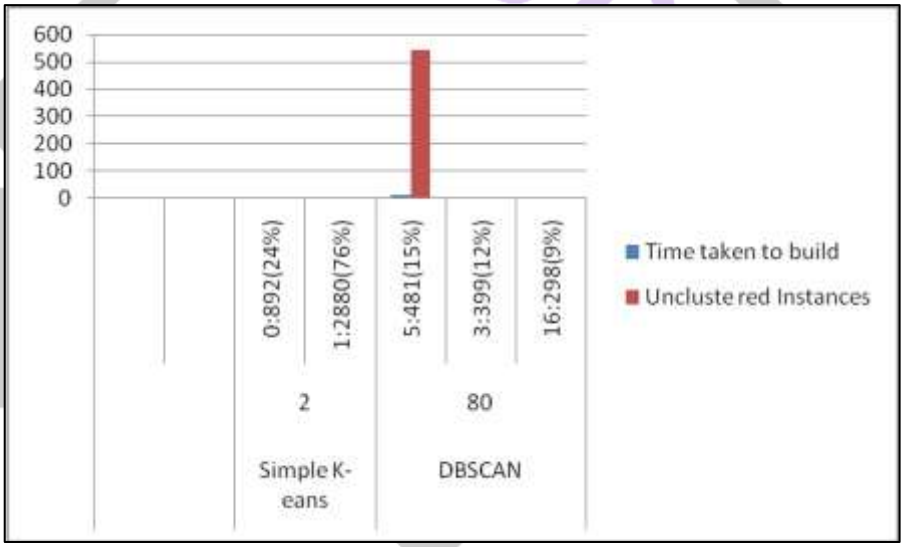

Figure 3.4: Performance Graph of Simple K-Means and DBSCAN

IV. Conclusion

In this paper we have compared two clustering algorithms i.e. Simple K-Means and DBSCAN We have analyzed the no. of clusters and cluster instances and also time taken to build the model. The performance of DBSCAN is better than the K-Means.

V. Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

[1] Klosgen W and Zytkow J M (eds.), Handbook of data mining and knowledge discovery, OUP, Oxford, 2002.

[2] Provost, F., & Fawcett, T., Robust Classification for Imprecise Environments. Machine Learning, Vol. 42, No.3, pp.203-231, 2001.

[3] Larose D T, Discovering knowledge in data: an introduction to data mining, John Wiley, New York, 2005.

[4] Kantardzic M, Data mining: concepts, models, methods, and algorithms, John Wiley, New Jersey, 2003.

[5] Goldschmidt P S, Compliance monitoring for anomaly detection, Patent no. US 6983266 B1, issue date January 3, 2006, Available at: www.freepatentsonline.com/6983266.html

[6] Bace R, Intrusion Detection, Macmillan Technical Publishing, 2000.

[7] Smyth P, Breaking out of the Black-Box: research challenges in data mining, Paper presented at the Sixth Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD-2001), held on May 20 (2001), Santra Barbara, California, USA.

[8] MacQueen J. B., "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. 1967, pp. 281–297.

[9] Lloyd, S. P. "Least square quantization in PCM". IEEE Transactions on Information Theory 28, 1982,pp. 129–137.

[10] Timonthy C. Havens. “Clustering in relational data and ontologies” July 2010.