International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-8)

Research Article

August 2015

Study of Various Clustering Algorithms Used by WEKA Tool

Pooja Bhandari
Uttarakhand Technical University,
Dun, Uttarakhand, India

Abstract—

With the increasing trends and scenario we have to deal with an immense data in our day to day life. Data mining is an interdisciplinary field in computer science that deals with the huge amount of data. It analyses the data in order to extract useful information from the database that can be further used for various purposes like business, banking and education etc. Data mining performs clustering , classification ,preprocessing, visualization and etc to discover hidden valuable information . Weka tool is one of the solution for all these processes.It is a machine learning tool.In this paper we are mainly dealing with the clustering. Clustering is a way to categorize objects of data set into different sets called as clusters based on the fact that objects of same cluster possess same quality or attributes and objects belonging to different clusters have different attributes.Objects within the same clusters are alike and objects of different clusters are unlike to each other. Here in this paper we are discussing and analyzing the various clustering algorithms.

Kewyord—Data mining, WEKA tool, k-mean algorithm, Agglomerative method,Divisive method,Density Based Method.

I. INTRODUCTION

Data mining is basically an extraction of previously unknown, and potentially useful information from databases/data warehouses. It uses various machine learning algorithm, statistical and visualization techniques to uncoverand present knowledge in a form, which is easily comprehensive to humans [1].Data mining helps to find out useful information that is hidden in a large volumes of data. It helps user to focus on the valuable information of their databases and neglect the unnecessary data. It comprises of lots of techniques and procedures that extract useful information to discover knowledge used for decision making. Data mining is also known by a name KDD which means Knowledge discovery in database. It includes following steps:

1. Data cleaning- This process removes noise and inconsistent data.
2. Data integration-This process deals with combination of multiple sources of data.
3. Data selection - The data required for analysis is retrieved from the database.
4. Data transformation - It consolidates and transforms data into forms appropriate for mining
5. Data mining - It uses intelligent techniques to extract patterns from data.
6. Pattern evaluation –It Identifies patterns that are interesting.
7. Knowledge presentation - Visualization and knowledge representation techniques are used to present the extracted or mined knowledge to the end user [3].

II. WEKA TOOL

WEKA is an acronym used for Waikato Environment for Knowledge Analysis and it is a popular suite of machine learning software written in Java.Weka is a product of University of Waikato, New Zealand. Weka is a free and open source software (available for public use)that uses GNU General Public License(GPL)[4]. The word weka has been derived from the word woodhen which is an endemic bird of a New Zealand. TheWeka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, and contains a graphical user .interfaces for interacting with data files and produces desirable result. Weka tool is used for data mining techniques like regression,clustering,classification,datapreprocessing,visualization and pattern recognition[2]. Its portability feature makes it platform independent as it is fully implemented in java programming language .It comprises of different machine learning algorithms for data mining.

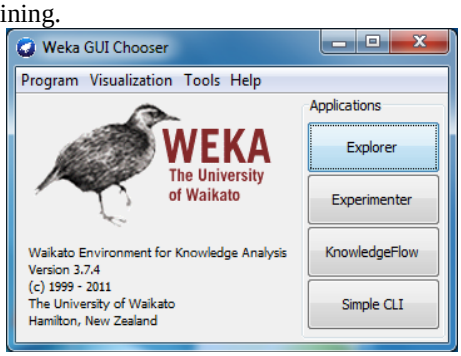

Fig 1: Weka tool screen

The GUI Chooser consists of four buttons:
- Explorer: An environment for exploring data with WEKA.
- Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes.
- 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.
- Simple CLI: Provides a simple command-line interface that allows direct executionof WEKA commands for operating systems that do not provide their own command line interface. [3]

III. WHAT IS CLUSTERING?

It is a procedure to split the data into groups to determine pattern from the data .These groups are called as clusters.A cluster contains similar objects but these objects are dissimilar to other cluster's objects. Objects within the same cluster possess similar properties whereas objects of different clusters have different properties. Clustering is useful in various fields like industries, education , banking, business and agriculture etc[1].

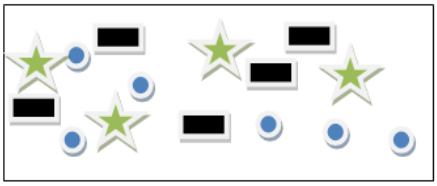

Fig 2: Data in Database before clustering

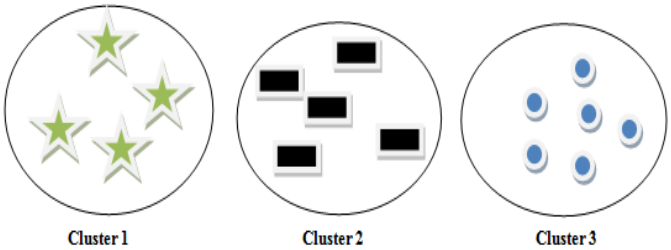

Fig 3 :Clusters Formation

IV. CLUSTERING ALGORITHMS AND TECHNIQUES

Various Clustering algorithms and methods have been developed. There are mainly divided into three categories which are then further categorized.

1. PARTITIONING METHOD

1.1 K-Mean Algorithm

In 1967 MacQueen was the first person who used k mean algorithm. IN 1957 Stuart Lloyd proposed the first standard algorithm for k mean. It is one of the partitioning method widely used in the industries. It is widely used due to easy implementation and efficient execution .It is a simpler clustering algorithms as it divides the dataset into fixed no. of clusters. For each cluster a centroid is defined .We are given with a data set D and fixed no. of clusters as k.

K-mean Algorithm for clustering is as follows:[7]

(i)Randomly choose k no. of objects equal to no. of clusters and marked them as cluster centroid which is a mean value of objects in the cluster.

(ii)Now place each object to the cluster to which the object has closest mean value.

(iii)Keep updating the means for each cluster.

(iv)Repeat steps (ii) and (iii) until the centroid doesnot changes.

(v)Stop

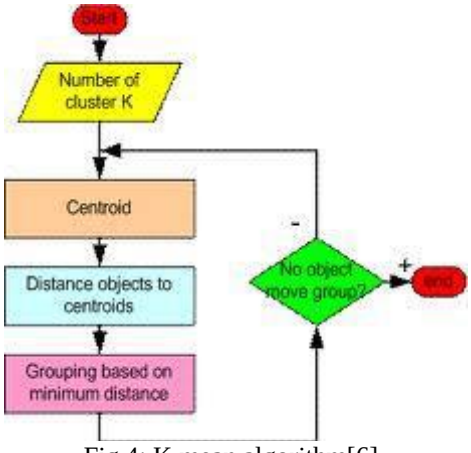

Fig 4: K mean algorithm[6]

2. Hierarchical Method

Hierarchical method attempts to build a hierarchy of clusters in a tree form. It can be constructed either in a top down orBottom Up manner. We can further categorize it into two ways:[1]

2.1 Agglomerative method

This generally begins from bottom with an individual cluster and moves towards upward by combining pairs of clusters together. It is also named as Bottom Up Approach. It comprises of following steps:

(i)Begin with single cluster.

(ii)Recursively combines two or more clusters

(iii)Repeat step ii until left with fixed no. of k clusters.

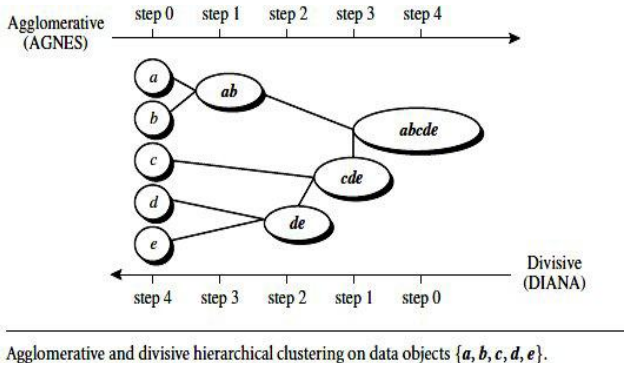

Fig 5: Hierarchical clustering[7]

2.2Divisive method

It is also known as Top Down Approach. It starts from Top with single cluster and moves downward by splitting them into more clusters. It includes following steps:

(i)Begin with a main cluster

(ii)Recursively divide the cluster into smaller clusters.

(iii)Repeat step ii until k no. of clusters are left.

3. Density Based Method

DBSCAN stands for Density Based Spatial Clustering Algorithm is developed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and XiaoweiXu in 1996. It creates clusters which is a maximal set of density connected points. Each object is associated with density and density represents maximum number of points that are within specified radius Eps[5]. It forms cluster of arbitrary shape and size and handles noise. It classified objects into three major groups:

- Core points: These objects have more minimum number of points termed as Minpts within Eps region. These objects form cluster. They lie inside cluster.These points are density reachable points.
- Border point: These points have fewer number of points Minpts.They lie close to core points but not inside the cluster.
- Noise point: These are neither core points nor border points. They are exterior to clusters.

Algorithm for DBSCAN is as follows:

Step 1: Randomly select a point p.

Step 2: Identify those points that are density reachable from p with respect to Eps and Minpts

Step 3 : If a point is core it forms a cluster.

Step 4: Repeat steps 2 and 3 until all points are checked for density reachability.

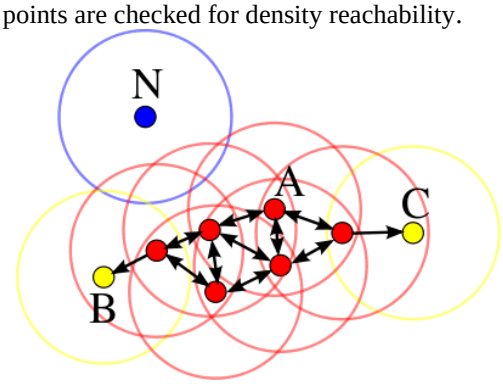

Fig 6: DBSCAN[5]

In this diagram, minPts = 3. Point A and the other red points are core points, because at least three points surround it in an ε radius. Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor density-reachable.

V. CONCLUSIONS

In the recent few years data mining techniques covers every area in our life. We use data mining techniques in mainly in the medical, banking, insurances, education etc. before start working in the with the data mining models, it isvery necessary to knowledge of available algorithms. The main aim of this paper to provide a detailed introduction of weka clustering algorithms. Weka is the data mining tools.It is the simplest tool to classify the data various types. It isthe first model for provide the graphical user interface of the user. For performing the clustering we used the promise data repository. It provides the past project data for analysis we are showingadvantages and disadvantages of each algorithm. Every algorithm has their own importance and we use them on thebehavior of the data, but on the basis of this research wefound that k-means clustering algorithm is simplest, fasterand effective algorithm which can be easily implemented. DBSCAN does not perform well on small datasets. In contrast to k-means DBSCAN does not requires us to know no. of clusters in advance.We didn't require deep knowledge of algorithms for working in weka.

VI. FUTURE WORK

In this paper, firstly we briefly discuss the concepts of Data Mining and KDD then data clustering and steps in Weka and techniques such as clustering by Weka steps. We describe aboutWeka tool. There is no doubt that the data mining is the useful term in the present and future. This paper shows only the clustering operations in the weka, tool we will try to make a complete reference paper of weka comprising many other processes that can be done using weka tool.

ACKNOWLEDGMENT

I hereby take this opportunity to express my heartfelt gratitude towards the people whose help is very useful to complete my dissertation work on the topic of "**Study of Various Clustering Algorithms Used by WEKA Tool**". My deepest gratitude goes to my *Project Guide*, **Mr. Kamal kant Verma**, *Asst. professor, Department of Computer Science and Engineering*, for his guidance, support, motivation and encouragement throughout the period this work was carried out. I extend my heartfelt thanks to my parents and friends for their moral and technical support.

REFERENCES

[1] Manish Verma, Mauly Srivastava, NehaChack, Atul Kumar Diswar, Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, pp. 1379-1384, 2012.

[2] Han J and kamberM,"data Mining Concepts and Techniques,"Morgankaufmann Publishers, San Francisco,2000.

[3] Bharat Chaudhari1, MananParik"A Comparative Study of clustering algorithms Usingwekatools"International Journal of Application or Innovation in Engineering & Management (IJAIEM).

[4] Pallavi, SunilaGodara "A Comparative Performance Analysis of Clustering Algorithms"International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 1, Issue 3, pp. 441-445.

[5] SlavaKisilevich, Florian Mansmann, Daniel Keim ―P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos, University of Konstanz.

[6] A.K. Jain, "Data Clustering: 50 Years Beyond K-Means" , Pattern Recognition Letters, Vol 31 Issue 8 : pp.651-666 , June, 2010.

[7] Sapna Jain, M AfsarAalam and M N Doja , "K-means clustering using weka interface",Proceedings of the 4th NationalConference;INDIA-Com2010.

