The International Journal Of Engineering And Science (IJES)
|| Volume || 5 || Issue || 10 || Pages || PP 35-39 || 2016 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Applying K-Means Clustering Algorithm to Discover Knowledge from Insurance Dataset Using WEKA Tool

Dr. Abdelrahman Elsharif Karrar[1], Marwa Abdelhameed Abdalrahman[2], Moez Mutasim Ali[3]

[1]*College of Computer Science and Engineering, Taibah University, Saudi Arabia*
[2]*College of Computer Science and Information Technology, University of Science and Technology, Sudan*
[3]*College of Computer Science and Information Technology, University of Science and Technology, Sudan*

----------ABSTRACT----------

Data mining works to extract information known in advance from the enormous quantities of data which can lead to knowledge. It provides information that helps to make good decisions. The effectiveness of data mining in access to knowledge to achieve the goal of which is the discovery of the hidden facts contained in databases and through the use of multiple technologies. Clustering is organizing data into clusters or groups such that they have high intra-cluster similarity and low inter cluster similarity. This paper deals with K-means clustering algorithm which collect a number of data based on the characteristics and attributes of this data, and process the Clustering by reducing the distances between the data center. This algorithm is applied using open source tool called WEKA, with the Insurance dataset as its input.

Keywords: *Clustering, Centroid, Data Mining, knowledge discovery, K-means, WEKA.*

Date of Submission: 17 May 2016 Date of Accepted: 05 November 2016

I. INTRODUCTION

The present age is characterized by the use of advanced data technology to save and retrieve data and enormous quantities and which is described data warehousing. These data provide open the door to a range of specialized in the management of those data subjects was the most prominent topic of data mining, which is one of the important methods to get useful information from the data. Scientists have known the term data mining as "part of the process of knowledge discovery in databases, which are made using multiple methods its goal configure models of data". [1]

Data mining techniques are designed to extract hidden information in the database, this modern technology has imposed itself firmly in the information age and in the light of the great technological development and widespread use of databases, they offer institutions in all areas the ability to explore and focus on the most important information in the databases. Data mining techniques focus on predictions future, explore behavior and trends allowing to take decisions correct and taken at the right time, also as the data mining techniques to answer Many questions in record time, especially those questions that are difficult to answer them by using methods statistical traditional. [2]

II. DATA MINING

Data mining technology is simply extract the important information from huge amount of information to follow certain mechanisms analysis this information, or it's a technique used in the process of extracting data from data warehouses. Data Mining passes a number of stages starting from data cleansing, and standardization of data, and the relevant test data, then transfer them, classify then evaluated and extract data. [3]

There are many definitions of the concept of data mining:

- Is a computerized search for the knowledge of data without prior assumptions about what can be this knowledge?
- Is the process used by companies to convert the raw data into useful information?
- Analysis of large sets of data and summarize the data in the new forms be understandable and useful to its users and is done in banks telecommunications companies commercial transactions and scientific data (Biology - Astronomy).
- Process data analysis by linking them with artificial intelligence techniques and statistical process, is simply a process of exploration and search for specific and useful information in a huge amount of data.

- Search the relevant information together collected by common characteristics and linked unit subject or specialization is then to find this information between a very large amounts of information that has no relationship was presented to the decision maker. [3]

III. TECHNIQUES OR METHODS OF DATA MINING

Data mining process uses several techniques through which you can discover the hidden trends and models in large amounts of data and can be used one or more of these techniques are as follows:

- **Prediction:** Use of available data and the application of certain techniques to give them values successful future.
- **Description:** Process description available data to see their ratings by the presence of the relationships between them.
- **Classification:** It is a set of data analysis to create a set of database assembled that can be used to classify any future data to find information that relates to the common characteristics and classification many tools such as decision tree, nearest neighbor and regression.
- **Association**: Is the database that includes fixed coupling relationships among a group of objects in the database any association between the occurrence of an event and another event occurs, which is often called the market basket analysis.
- **Sequential Analysis**: Which is similar to association and placed in the name the link, but analysis linked in time in the Search for models occurs in any deal with the succession of data that occur in separate cases.
- **Clustering**: The idea of collecting data is a simple idea in nature and very close to the human way of thinking where we are whenever we deal with a large amount of data tends to summarize the vast amount of data into a small number of groups or categories in order to facilitate the process of analysis. Algorithms assembly used widely not only for the organization and classification of data but also the data compression and build a model arrangement.The process of clustering is the process of collecting objects or items that possess the qualities and attributes are similar in groups called clusters. The process of clustering one of the main roads in the process of data mining and can be used as a standalone tool to gain insight into how the distribution of data, control characteristics of each group and focus on a specific set of groups so for further analysis and can be as a preliminary step to the work of an elementary or other techniques such as characterization and classification. [4]

IV. K-MEANS ALGORITHM

Is one of the clustering algorithms, it collect a number of data based on the characteristics and attributes of this data and the process of the Clustering by reducing the distances between the data center. The steps of this algorithm are:

- Determine the number of clusters K which is a step Initialize Preliminary.
- Determine the coordinates of the centers of clusters (Centroid) randomly for the first time and calculate the average of the points that belong to the center for the rest of the times.
- Calculate the distance between each example and among all centers and is used Euclidean dimension. Given the Euclidean distance (d_{ij}) between the two examples(i,, j) the following relationship:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

- Data collection (examples) with its nearest center.
- Repeat steps 2 through 4 until you get stability (and the absence of moving objects within the clusters), or even repeating a certain number of times. [5]

Before you start in the presentation of the results of the application must describe the data which were used in the search: this paper apply the algorithm (k-means) to the insurance company data to clarify the best payment method that can be available to the customers, as the company suffered from the challenges and difficulties limit their ability to deal with payment methods. This was done using a program WEKA.

V. WEKA

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-ofthe-art facility 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. [6]

WEKA implements algorithms for data preprocessing, classification, regression, clustering and association rules. It also includes visualization tools. [6]

To perform cluster analysis in weka the dataset is needed to be loaded to weka and it should be in the format of CSV or ARFF file format. If the dataset is not in arff format we need to be converting it.

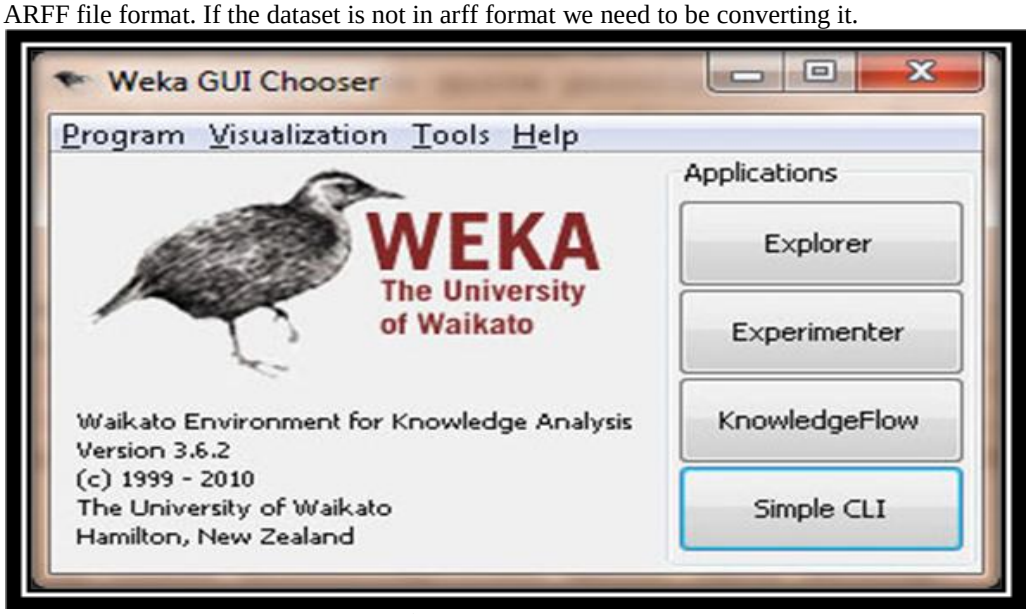

Figure 1: WEKA GUI

	A	B	C	D	E	F	G	H	I	J
1	Id	DocumentNo	PayMoney	PayDate	IsalNo	PayType	DocType			
2	93	2	0	17/04/2013	55455	check+cash	Third part			
3	96	1	1000	18/04/2013	443	check	Third part			
4	97	1	1000	18/04/2013	144	check	Third part			
5	98	1	900	18/04/2013	200	check	Third part			
6	99	1	1240	18/04/2013	112	check	Third part			
7	100	1	1140	18/04/2013	124	check	Third part			
8	101	1	1140	18/04/2013	1000	check	Third part			
9	102	1	2180	18/04/2013	25	check	Third part			

Figure 2: Customer Data in the Insurance Company

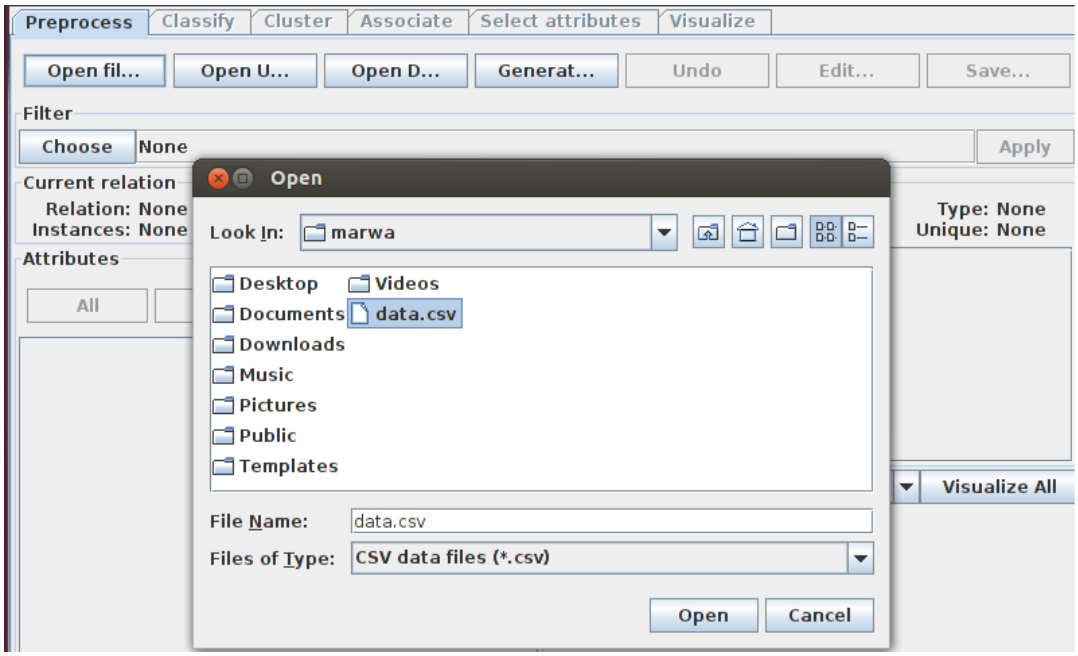

Figure 3: Choose the Data File

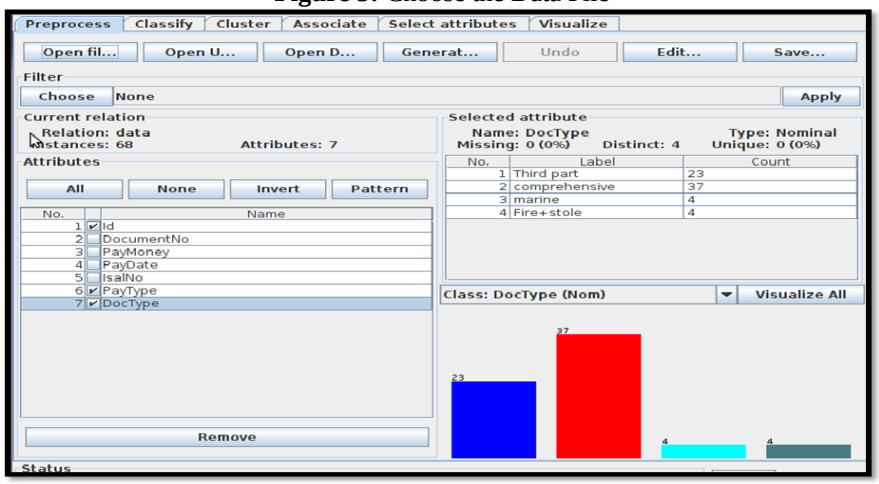

Figure 4: Choosing the Attributes

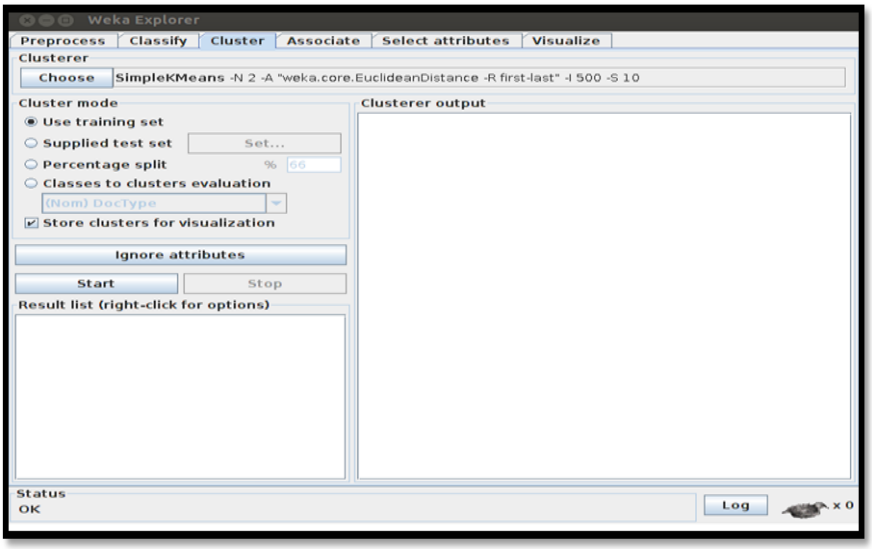

Figure 5: Cluster mode

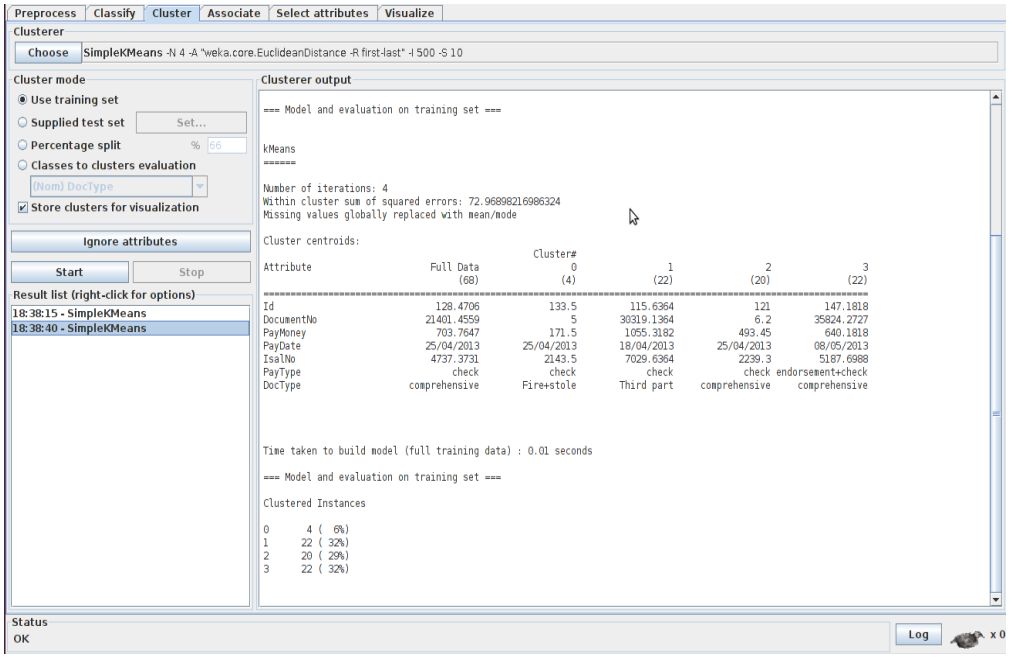

Figure 6: Clustered output

VI. RESULT DISCUSSION

When we study the output report we can see four clusters (cluster 0, cluster1, cluster2, cluster3). In the report appear as:

Number of clusters selected by cross validation: 4

The first column gives you the overall population centroid. The second and third and fourth and fifth columns give you the centroids for cluster 0, 1, 2 and 3 respectively. Each row gives the centroid coordinate for the specific dimension.

Here we must notice that finding the centroids is an essential part of the algorithm. The centroids are a result of a specific run of the algorithm and are not unique, a different run may generate a different centroid set.

For each cluster there is "clusters prior" probability. The estimators consist of a number for each possible attribute value, and the attribute values are treated in order.

- Cluster0 has total 4 objects, out of which majority of objects (6) data sets.
- Cluster1 has total of 22 objects, out of which majority of objects (32) data sets.
- Cluster2 has total 20 objects, out of which majority of objects (29) data sets.
- Cluster3 has total 22 objects, out of which majority of objects (32) data sets.

VII. CONCLUSION

Clustering is organizing data into clusters or groups such that they have high intra-cluster similarity and low inter cluster similarity. K-means algorithm is one of the clustering algorithms, it collects a number of data based on their characteristics and attributes, and run the process of Clustering by reducing the distances between the data center. WEKA an open source tool is used to apply K-means algorithm on insurance dataset.

REFERENCE

[1]. Daniel T. Larose, "Data Mining Methods and Models", 2011.
[2]. G. K. Gupta, "Introduction to Data Mining with Case Studies", 2006.
[3]. Gregory Piatetsky, "From Data Mining to Knowledge Discovery:An Introduction", 2012.
[4]. Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques (Second Edition)", 2014.
[5]. Subhash Sharma, Ajith Kumar, "Cluster Analysis and Factor Analysis", 2014.
[6]. Yizhou Sun, "an Introduction to WEKA", 2008.