Red Sea University Journal of Basic and Applied Science ISSN: 1858-7658 (Online)
Vol (2) Special Issue (3) May-2017 ISSN: 1858-7690 (Print)

A Comparative Study of Clustering Algorithms for Detect SQL Injection Attack

Salma Babker Mohammed[1, a *], Mohamed Ahmed Elmobark[2,b] and Mohammed kabashi Abd Elrhman[3,c]
[1,2,3]computer science, sudan

[a] Sulmab5@gmail.com, [b]m.a.almobark@hotmail.com, [c] mohkabashi@hotmail.com

Abstract
data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Waikato environment for knowledge analysis (WEKA) is a data mining tool. It contains many machines leaning algorithms. It provides the facility to classify our data through various algorithms. In this paper we studied the various clustering algorithms to detect SQL injection attack. Our main aim is to show the comparison between the different clustering algorithms of WEKA and to find out which algorithm will be most suitable to detect SQL injection attack. k-means and Make density algorisms most suitable to detect SQL injection attack according to our result.

Keywords: Data mining algorithms, WEKA tools, Clustering methods, SQL injection attack.

Introduction

Data mining [1] is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering. In this research paper we work only with the clustering because it is suitable for our aim to detect SQL injection attack and we have a very large dataset. We use WEKA tools for clustering. Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. We use WEKA data mining tools because provides a better interface to the user than compare the other data mining tools and we can work in WEKA easily without having the deep

knowledge of data mining techniques.

Web Application Architecture

Web application is simply recognized as a program running on a web browser, generally has a three-tier. A presentation tier is sent request to the web browser. Web application consists of presentation tier, graphical user interface tier and database tier [2].

SQL Injection Attack

An SQL injection attack takes place when a hacker changes the semantic or syntax logic of an SQL text string by inserting SQL keywords or special symbols within the original SQL command, executed at the database layer of an application. Different attack techniques exist which include the use of SQL tautologies, logic errors/Illegal queries, union queries and Piggy-back queries. Other more advanced techniques use injections based on interference and alternative codification [3].

WEKA

In [4] WEKA is the product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997. The figure of WEKA is shown in the figure1. The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results. We are using the some clustering algorithms of WEKA for comparison of algorithms to detect the SQL injection, for complete to this purpose we used CSIC 2010 HTTP Dataset.

ISSN: 1858-7658 (Online)
ISSN: 1858-7690 (Print)

Fig 1: Front view of WEKA tools

Dataset The CSIC dataset was created in 2010; it was designed with the aim to overcome the described drawbacks of existing datasets. A public dataset, usable by the whole scientific community, allows the comparison of different detection systems. In total, the CSIC dataset contains 36000 normal requests and more than 25000 anomalous requests. The requests are labeled either as normal or anomalous. Regarding the generation of attacks, both static and dynamic attacks were generated, including modern web attacks such as SQL injection, buffer overflow, information gathering, cross-site scripting, server side include and parameter tampering [5]. Figure2 show load CSIC 2010 HTTP dataset in WEKA.

Methodology

Our methodology is very simple. We used CSIC 2010 HTTP dataset and apply it on the WEKA. In the WEKA we applied different clustering algorithms and predicted useful results for detect SQL injection attack, that are very helpful for the new users and new researchers.

Performing Clustering in WEKA

For performing cluster analysis in WEKA, We have loaded the data set in WEKA that is shown in the figure2. For the WEKA the data set should have in the format of CSV or .ARFF file format. If

ISSN: 1858-7658 (Online)
ISSN: 1858-7690 (Print)

the data set is not in arff format we need to convert it into arff.

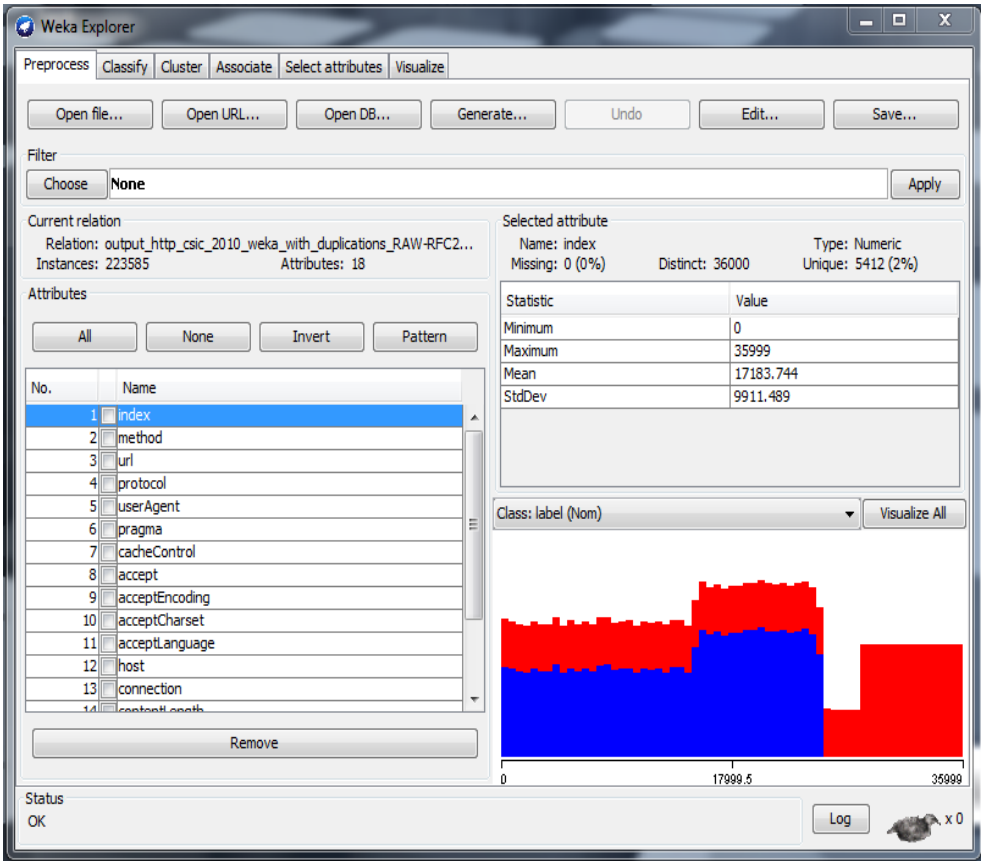

Fig 2: Load data set in to the WEKA

After that we have many options shown in the figure 2. We perform clustering so we click on the cluster button. After that we need to choose which algorithm is applied on the data. It is shown in the figure 3. And then click ok button.

ISSN: 1858-7658 (Online)
ISSN: 1858-7690 (Print)

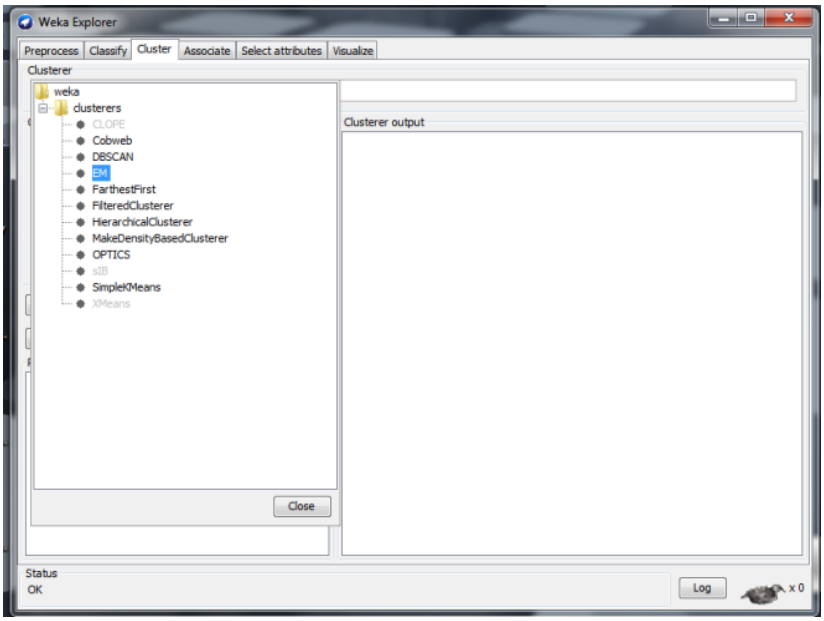

Fig 3: Various clustering algorithms in WEKA

Farthest First Algorithm

Farthest first [6] is a variant of K-means that places each cluster centre in turn at the point furthest from the existing cluster centers. This point must lie within the data area. This greatly sped up the clustering in most cases since less reassignment and adjustment is needed.

ISSN: 1858-7658 (Online)
ISSN: 1858-7690 (Print)

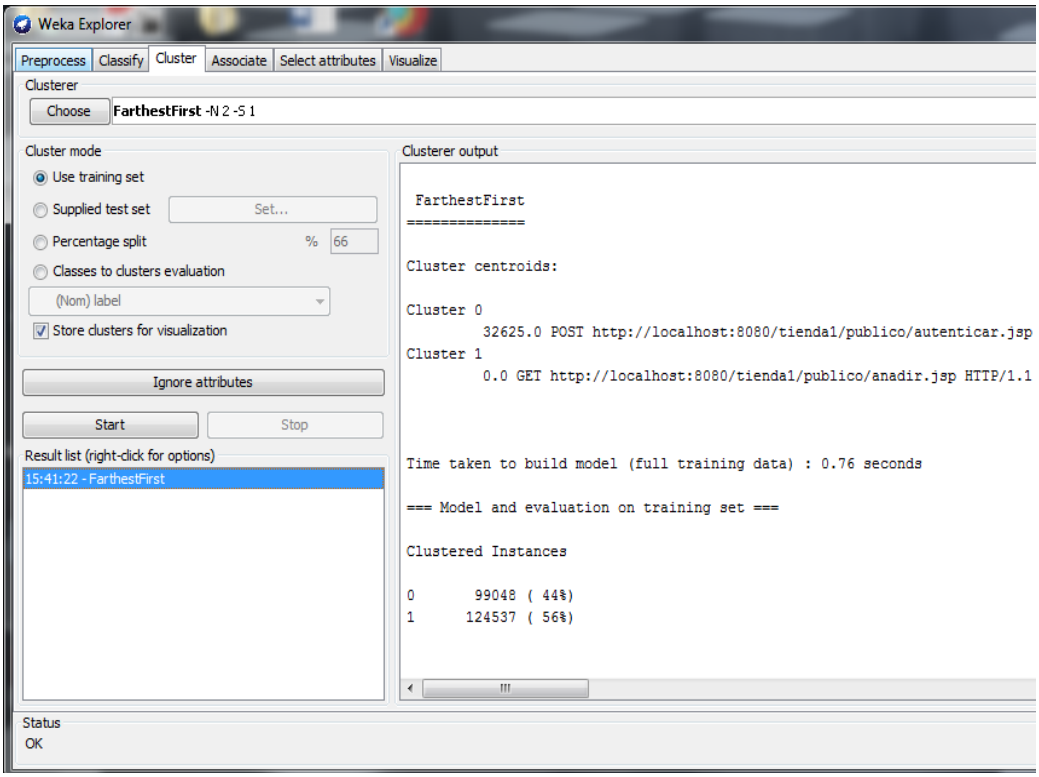

Fig 4: Farthest first algorithm

Result of farthest first algorithm is shown in the figure 4. It is divide the whole data set in two clusters. Cluster(0) for SQL attack and cluster(1) for normal SQL.

Advantages Farthest-point heuristic based method has time complexity O (nk), where n is number of objects in the dataset and k is number of desired clusters. Farthest-point heuristic based method is fast and suitable for large scale data mining applications.

K-means Clustering Algorithm

In data mining, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

K-means [7] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed

a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to recalculate new centroids as bar centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized and calculated.

ISSN: 1858-7658 (Online)
ISSN: 1858-7690 (Print)

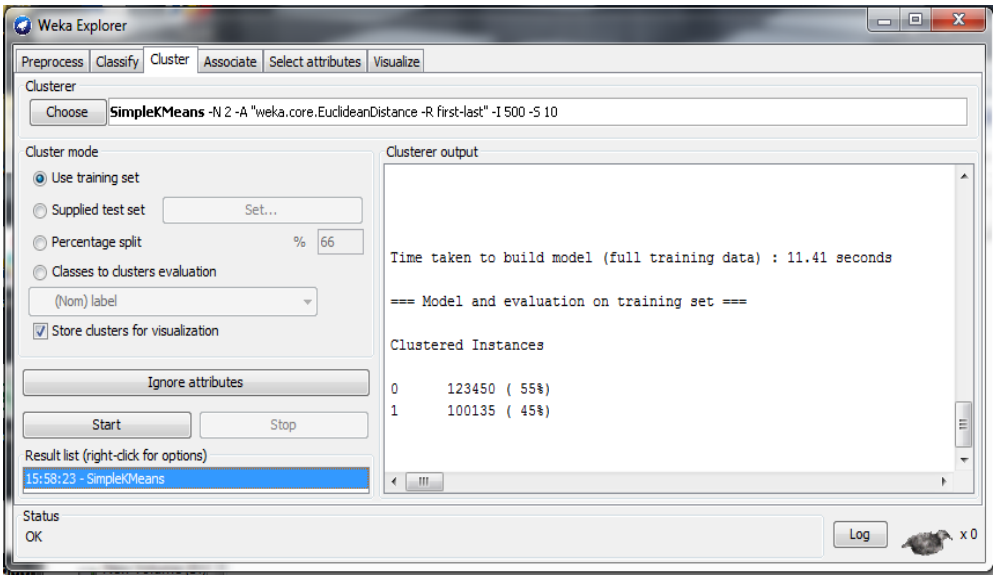

Fig 5: k- means clustering algorithm

Figure 5 show that the result of k-means clustering methods. Cluster(0) for SQL attack and cluster (1) for normal SQL.

Advantages of Using this Technique

1- With a large number of variables, K-means may be computationally faster than hierarchical clustering [4] (if K is small).

2- K-means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

Disadvantages of Using this Technique

1- Difficulty in comparing quality of the clusters produced (e.g. for different initial partitions or values of K affect outcome).

2- Fixed number of clusters can make it difficult to predict what K should be.

3- Does not work well with non globular clusters. Different initial partitions can result in different final clusters. It is helpful to rerun the program using the same as well as different K values, to compare the results achieved.

Make Density Based Clustering Algorithm

A cluster [8] is a dense region of points that is separated by low density regions from the tightly dense regions. This clustering

ISSN: 1858-7658 (Online)
ISSN: 1858-7690 (Print)

algorithm can be used when the clusters are irregular. The make density based clustering algorithm can also be used in noise and when outliers are encountered. The points with same density and present within the same area will be connected to form clusters.

The result show on in figure 6.

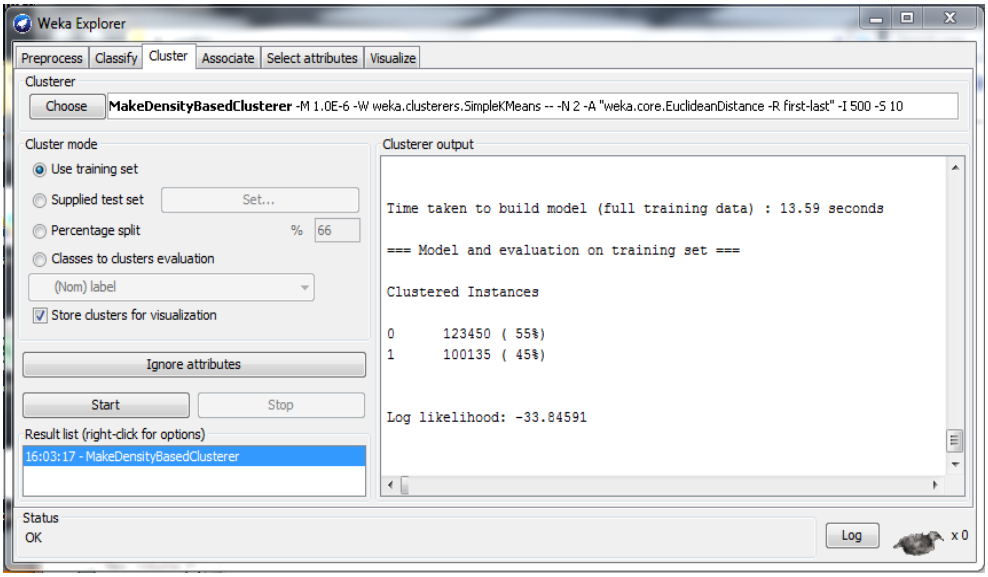

Fig 6: Make density based clustering algorithm.

Results

In the recent few years data mining techniques covers every area in our life. We use data mining techniques in mainly in the education, banking, medical. Before start working with the data mining models, it is very necessary to know the available algorithms. The main aim of this paper is to compare WEKA clustering algorithms to cluster SQL injection attack data. WEKA is the data mining tools. It is the simplest tool for classifying the various types of data. It is the first model for providing the user with the graphical user interface of. We show advantages and

disadvantages of each algorithm. Every algorithm has its own importance and we use them on the behavior of the data, but on the basis of this research. We are not required to have a deep knowledge of algorithms for working in WEKA. After the analysis of dataset by clustering algorithm, table1 shows comparison of results.

We use Eq. 1 to calculate false alarms rate.
False alarms rate = wrong attack identifications / total normal connections. →(1)

According to the below table1 we found that k-means and Make density algorisms give us the same result but farthest first algorithm result far from real data.

Table1: Results

Algorithm	Attack	normal	False alarms rate	true alarms rate
Real data	53.5%	46.5%		
Farthest first	44%	56%	19.7%	80.3%
k-means	55%	45%	3.7%	96.3%
Make density	55%	45%	3.7%	96.3%

Discussion

In this paper we found that k-means and Make density

algorisms give us the same result but farthest first algorithm result far from real data to detect attacks.

k-means and Make density algorisms give us 3.7% false alarms rate.

According to result come from this paper use of k-means and Make density algorisms to detect attacks in security issues better than farthest first algorithm.

Acknowledgement

References

[1] Sahu, Hemlata, S. Shrma, and S. Gondhalakar. "A Brief Overview on Data Mining Survey." International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1 (2011).

[2] Jarosch, D. I. Dennis. Effects and opportunities of native code extensions for computationally demanding web applications. Diss. Humboldt-Universitätzu Berlin, 2011.

[3] Pinzón, Cristian I., et al. "idMAS-SQL: intrusion detection based on MAS to detect and block SQL injection through datamining." Information Sciences 231 (2013): 15-31.

[4] S. yana, A. shwin. "Software tools for teaching undergraduate data mining course." Proceedings of the ASEE-2013 Mid-Atlantic Fall Conference, University of the District of Columbia.(2013).

[5] T. Giménez, Carmen. "Study of stochastic and machine learning techniques for anomaly-based Web attack detection." (2015).

[6] Zafar, M. Husnain, and M. Ilyas. "A Clustering Based Study of Classification Algorithms." International Journal of Database Theory and Application 8.1 (2015): 11-22.

[7] Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666.

[8] Bressert, E., et al. "The spatial distribution of star formation in the solar neighborhood: do all stars form in dense clusters?" Monthly Notices of the Royal Astronomical Society: Letters 409.1 (2010): L54-L58.