Introduction to WEKA & Study on Data Mining tool with its Comparative Analysis

Prof. Saurabh A. Ghogare

Abstract- **Data Mining tools used deal with problems such as classification, clustering, association rule, neural networks, it is a open access tools directly communicates with each tool or called from java code to implement using this. In this paper we present introduction of various machine learning data mining tool used for different analysis, Waikato Environment for Knowledge Analysis (WEKA) is introduced by university of New Zealand it has the capacity to convert. CSV file to Flat file. In this workit shows the overall process of WEKA analysis of file which converts and select the attributes to be mined and comparison with Knowledge Extraction of Evolutionary Learning not only analysis the data mining classifications but also the genetic, evolutionary algorithms is the best efficient tool in learning and also shows the comparative analysis of different Data mining tools.**

Keywords- **GNU, Classification Techniques, YALE, Machine Learning, WEKA, KEEL,KNIME, JHepWork, ARFF.**

I. INTRODUCTION

The WEKA stand for Waikato Environment for Knowledge Analysis, it is a machine learning toolkit introduced by Waikato University, New Zealand. It is open source software written in Java language (GNU Public License. It can be run on any plate form like Windows, Linux and Mac. It consists of collection of machine learning algorithms for implementing data mining tasks which is used for assessment and solved various research problem, its GUI based tool mainly used for comprehensive set of preprocessing tools, evaluation methods and has an environment for comparing learning techniques.This article gives a comparative study of open source tools of data mining highly used and popular in the market and compare it with WEKA to solve real life scenario

II. DATA MINING TOOLS

The data mining tools are used to solve the data mining classification problems, in this article we have discuss on classification, clustering, association, regression and mining tools the process of extracting patterns from data is called data mining. It is known as an essential tool by modern business since it is able to convert data into business intelligence thus giving an informational edge.

A. Association Rule

Association rule mining is a method which aims to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. The problem of mining association rules is mainly to find the all regulations that may satisfy a user-specified minimum support and minimum confidence. The association rule mining problem can be decomposed into two sub problems [2]: Find all combinations of items, called common item-sets, whose support is greater than minimum support and time unbearable.

B. Classification

In Classification, model used to learn a model that can classify data samples into classes or known classes. The Classification process involves following basic steps:

Step-I:-Create new training data set.

Step-II:-.Identify class attributes and classes of the given data set.

Step-III:-Identify useful attributes for classification.

Step-IV:-Learn a model using training examples in Training set.

Step-V:-Use the model to classify the unknown data samples.

1. Decision Tree:

A decision tree contain tree like structure and also support tool that used tree graph or model for decision and possible consequences, including chance event outcomes, resource costs, and utility. For example, loan applicants as good or bad credit risks. As shown in the following figure.

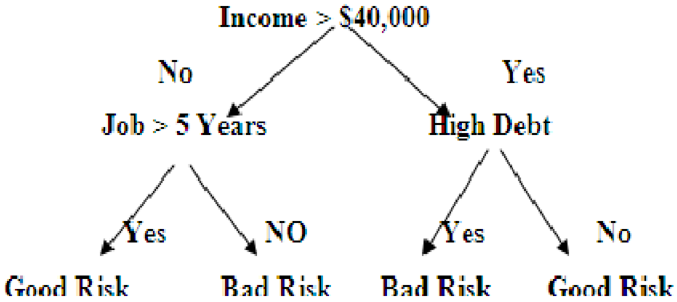


Fig: 1 Decision Tree Example

C. Clustering Technique

In Cluster technique which create group of similar objects together. It can also be defined as the organization of dataset into homogeneous and/or well separated groups with respect to distance or equivalently similarity measur. The two types of attributes associated with clustering, numerical and categorical attributes. Numerical attributes are related with ordered values such as height of a person and speed of a train. Categorical characteristics are those which unordered values such as kind of a drink and brand of car. Clustering is available in two types:

i. Hierarchical

ii. Partition (non Hierarchical)

1. K-Mean Algorithm:

K-means is the simplest unsupervised learning algorithms which solve the well known clustering problem. The process that follows a simple and easy way to classify the given data set through a certain number of clusters. The main aim or purpose is to define k centers, one for each cluster. These centers should be placed in a astute way because of different location causes different result.

C. Regression:

It is a data mining function which predicts a number. The numbers like Profit, sales, mortgage rates, house values, square footage, temperature, or distance could all be predicted using regression techniques.

III. Different data Mining Tool

The best open source data mining software that are Orange, Rapid Miner, Weka, JHepWork, and KNIME.

1. Orange

The Orange is component-based data mining and machine learning software that suite that features friendly yet powerful, fast and versatile visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting. It contain cross-platform and written in C++ and Python and it also include complete set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques.

2. Rapid Miner

Previously called as YALE (Yet Another Learning Environment), it is one of the best predictive analysis system developed by the company with the same name as the Rapid Miner. It is written in JAVA programming language. It provides integrated environment for deep learning, text mining, machine learning & predictive analysis. It offers the server as both on premise & in public/private cloud arrangements.. It has a client/server model as its base.

3. WEKA

Written in Java, Weka (Waikato Environment for Knowledge Analysis) is a well-known suite of machine learning software that supports several typical data mining tasks, particularly data clustering, classification, regression, visualization, and feature selection. labeled by a fixed number of attributes. Weka which provides the access to SQL databases and utilizing Java Database Connectivity that can process the result returned to a database query. Weka is easier because a beginner can go through the process of applied machine learning using the graphical interface without having to do any programming. Its techniques are based on the theory that the data is available as a single flat file or relation, where each data point is its main user interface is the Explorer, but the same functionality which can be accessed from the command line or through the component-based Knowledge Flow interface.

4. JHepWork

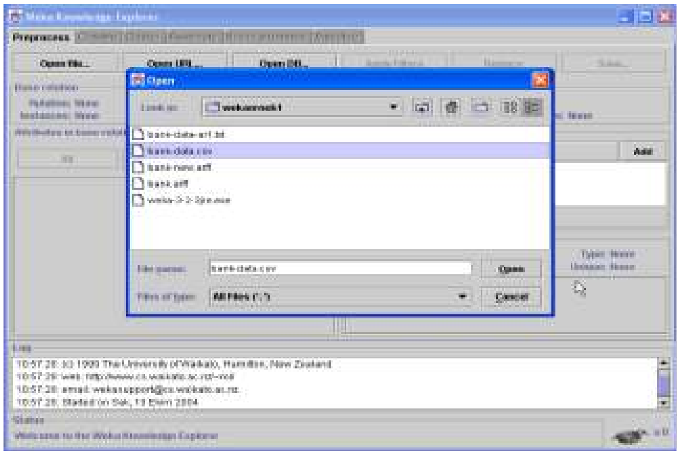
This tools is mainly designed for scientists, engineers and students because it is a free and open-source data-analysis framework that can be created as an effort to make a data-analysis environment using open-source packages with a understandable user interface and to create a tool competitive to commercial programs. It is specially prepared for interactive scientific plots in 2D and 3D, it also contains numerical scientific libraries implemented in Java for mathematical functions, random numbers, and other data mining algorithms. jHepWork is work on a high-level programming language Jython, but Java coding can also be used to call jHepWork numerical and graphical libraries.

5. KNIME

KNIME (Konstanz Information Miner) is a free and open source data integration, processing, analysis, and exploration platform. It gives users the facility to visually create data flows or pipelines, selectively execute some or all analysis steps, and later studies the results, models, and interactive views. It is written in Java, and it is based on Eclipse and use of its extension method to support plug-in thus providing additional functionality.

IV. WEKA IMPLEMENTATION

WEKA has the facility to read in ".csv" format files. As many databases or spreadsheet applications can save or export data into flat files in this format can be seen in the sample data file, the first row that contains the attribute with the names separated by commas and followed by each data row with attribute values listed in the same order (also separated by commas). The loaded file in WEKA, the data set can be saved into ARFF format. Involved in converting a ".csv" file into WEKA's native ARFF, then the recommended approach is to use the following from the command line shown in Figure 1.



bank-data-arff.txt
bank-data.csv
bank-new.arff
bank.arff
weka-3-2-3jre.exe

File name: bank-data.csv
Files of type: All Files (*.*)

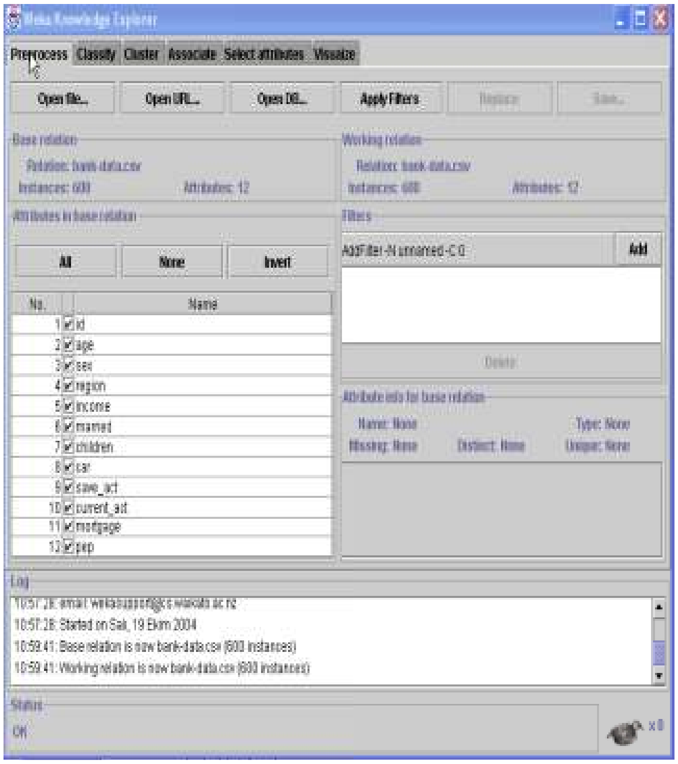
```
10:57:28: (c) 1999 The University of Waikato, Hamilton, New Zealand
10:57:28: web: http://www.cs.waikato.ac.nz/~ml
10:57:28: email: wekasupport@cs.waikato.ac.nz
10:57:28: Started on Salı, 19 Ekim 2004
```

Status: Welcome to the Weka Knowledge Explorer

Fig.1. Loading the Data into WEKA

A. Choosing The Data From File

After data is loaded, the next step is to recognize the attributes. The process scan of the data will compute some basic statistics on each attribute. The left panel side in given Figure, which shows the list of known attributes, and while the top panels indicate the names of the base relation (table) and the current working relation. By Clicking on attribute in the left side panel will display the basic statistics on that attribute. Like min, max, mean, standard deviation, etc.



Base relation
Relation: bank-data.csv
Instances: 600 Attributes: 12

Working relation
Relation: bank-data.csv
Instances: 600 Attributes: 12

Filters
AddFilter -N unnamed -C 0

Attribute info for base relation
Name: None Type: None
Missing: None Distinct: None Unique: None

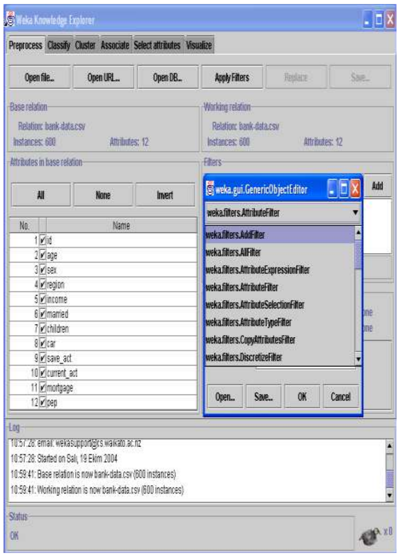
No.		Name
1	☑	id
2	☑	age
3	☑	sex
4	☑	region
5	☑	income
6	☑	married
7	☑	children
8	☑	car
9	☑	save_act
10	☑	current_act
11	☑	mortgage
12	☑	pep

```
10:57:28: email: wekasupport@cs.waikato.ac.nz
10:57:28: Started on Salı, 19 Ekim 2004
10:59:41: Base relation is now bank-data.csv (600 instances)
10:59:41: Working relation is now bank-data.csv (600 instances)
```

Status: OK

Fig 2.Choosing the Data into File

B. Prepare the Data to Be Mined
1. Selecting Attributes
In sample data file, each record is uniquely identified by a id; we have to to remove this attribute before the data mining step and using the Attribute filter. In the "Filters" section, click on the filter button (to the left of the "Add" button). The message show as popup window with list available filters.



Relation: bank-data.csv
Instances: 600 Attributes: 12

Relation: bank-data.csv
Instances: 600 Attributes: 12

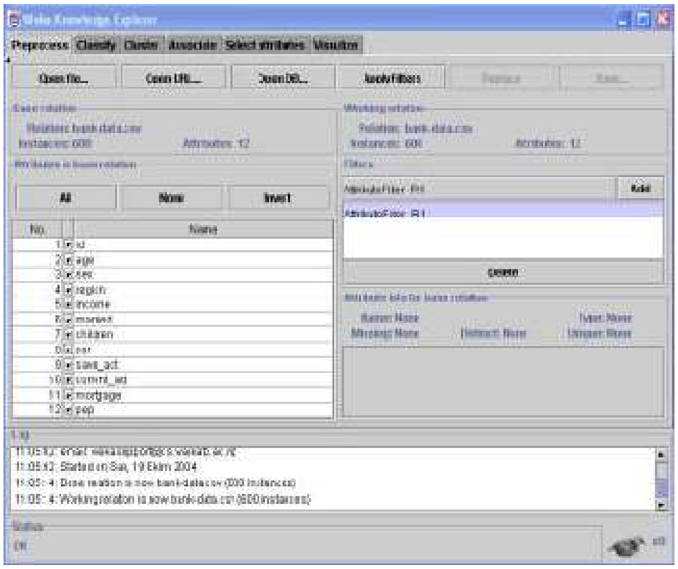
weka.filters.AttributeFilter
weka.filters.AddFilter
weka.filters.AllFilter
weka.filters.AttributeExpressionFilter
weka.filters.AttributeFilter
weka.filters.AttributeSelectionFilter
weka.filters.AttributeTypeFilter
weka.filters.CopyAttributesFilter
weka.filters.DiscretizeFilter

No.		Name
1	☑	id
2	☑	age
3	☑	sex
4	☑	region
5	☑	income
6	☑	married
7	☑	children
8	☑	car
9	☑	save_act
10	☑	current_act
11	☑	mortgage
12	☑	pep

```
10:57:28: email: wekasupport@cs.waikato.ac.nz
10:57:28: Started on Salı, 19 Ekim 2004
10:59:41: Base relation is now bank-data.csv (600 instances)
10:59:41: Working relation is now bank-data.csv (600 instances)
```

OK

Fig 3. Mining Process Of Selecting Attributes.



Relation: bank-data.csv
Instances: 600 Attributes: 12

Relation: bank-data.csv
Instances: 600 Attributes: 12

AttributeFilter -R1

AttributeFilter -R1

Name: None Type: None
Missing: None Distinct: None Unique: None

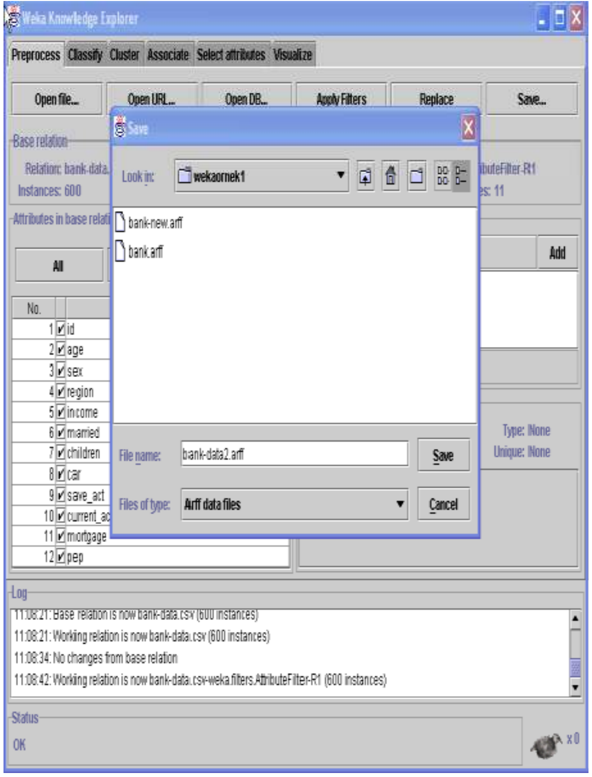
No.		Name
1	☑	id
2	☑	age
3	☑	sex
4	☑	region
5	☑	income
6	☑	married
7	☑	children
8	☑	car
9	☑	save_act
10	☑	current_act
11	☑	mortgage
12	☑	pep

```
11:05:42: email: wekasupport@cs.waikato.ac.nz
11:05:42: Started on Salı, 19 Ekim 2004
11:05:44: Base relation is now bank-data.csv (600 instances)
11:05:44: Working relation is now bank-data.csv (600 instances)
```

OK

Fig 4. Represents Attributes is to be filtered

And it finally, click the button "Apply Filters" on the top panel to apply the filter to the current working relation. And possible to select several filters and apply all of them at once.



Relation: bank-data.
Instances: 600

ibuteFilter-R1
es: 11

Look in: wekaornek1

bank-new.arff
bank.arff

File name: bank-data2.arff

Files of type: Arff data files

Type: None
Unique: None

No.		
1	☑	id
2	☑	age
3	☑	sex
4	☑	region
5	☑	income
6	☑	married
7	☑	children
8	☑	car
9	☑	save_act
10	☑	current_ac
11	☑	mortgage
12	☑	pep

```
11:08:21: Base relation is now bank-data.csv (600 instances)
11:08:21: Working relation is now bank-data.csv (600 instances)
11:08:34: No changes from base relation
11:08:42: Working relation is now bank-data.csv-weka.filters.AttributeFilter-R1 (600 instances)
```

OK

Fig 6 : ARFF Dataset Result save Process

V. Comparative Analysis

All the above discuses data mining tools are open source software to assess evolutionary algorithms for problems in mining include data classification techniques. Knowledge extraction based on evolutionary learning contains collection of knowledge mining algorithms preprocessing technique such as training selection feature discretization imputation methods for missing values.

Computational intelligence based learning algorithms including evolutionary rule learning based on different approaches and hybrid models such as genetic fuzzy systems evolutionary neural networks major feature of KEEL developed to ensemble different data mining models of evolutionary learning algorithms with open source code in java which includes data preprocessing in specialized discretization training feature selection imputation methods for missing values and noisy data. KEEL provides an user friendly interface oriented to the analysis of algorithms to create in online that education supports to learn the operation of the algorithm and different evolutionary rule learning models have been implemented fuzzy rule with good trade-off between accuracy and interpretability, genetic programming algorithms that use tree representations for extracting knowledge and based on patterns subgroups discovery have been integrated to reduce the data evolutionary algorithms have included. Compare Knowledge extraction based evolutionary learning WEKA is collection of machine learning directly applied to dataset or called from java code. WEKA used to find the goal of data mining by predicting and validating values which are indeed correct. The Algorithms can be used API to build custom tools, applications and algorithms as well. WEKA system has been able to put into operation and evaluate a number of different Algorithms for different steps in the machine learning process. The Output information provided by the package is the sufficient for an expert in machine learning and also results as displayed by the system show a detailed description of the flow and the steps involved in the entire machine learning process. The comparative outputs provided by different algorithms are easy to understand and analyzing

ARFF is a file format is one of the most widely used data storage formats for research databases, making this system easier for use in research oriented projects. This package provides and number of application program interfaces which help Data miners build their systems using the"core WEKA system". As the system is entirely based on Command Line parameters and switches, it is difficult for an amateur to use the system efficiently. The, main disadvantage is that the system is a Java based system and requires Java Virtual Machine installed for its execution.

VI. Conclusion

In this article a general data mining tools for classification, clustering, association rule, neural networks datasets and an explanation mechanism to explain the novel results was described and explained. The approaches of mining tools learning are characterized, we developed the WEKA method is based on choosing the file and selecting attributes to convert .csv file to flat file and discussed features of KEEL & WEKA performance. Our work extends to utilize the implementation of dataset for each data mining tools present in the section III to achieve a high rate of accuracy in the case of unfamiliar attacks can also improve the efficiency when analyzing the complex dataset.

References

1. W. Lee, S. J. Stolfo Data Mining Approaches for Intrusion Detection.

2. An Implementation of ID3: Decision Tree Learning Algorithm Wei Peng, Juhua Chen and Haiping Zhou Project of Comp 9417: Machine Learning University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, and Australia.

3. D.P.Greene and S.F. Smith, Competition-based induction of decision models from examples, Machine Learning 3 (1993) 229-257.

4. J. Bacardit and J.M. Garrell, "Bloat control and generalization pressure using the minimum description length principle for a Pittsburgh approach learning classifier system", in Advances at the frontier of Learning Classifier Systems, LNCS Vol. 4399 (2006) 61-80.

5. www.junauza.com/2010/11/free-data-mining software.html