"1st International Symposium on Computing in Informatics and Mathematics (ISCIM 2011)"
in Collaboration between EPOKA University and "Aleksandër Moisiu" University of Durrës
on June 2-4 2011, Tirana-Durres, ALBANIA

The influence that WEKA workbench has in processing information

Esteriana HASKASA[1], Edlira KALEMI[2], Lejdi KOCI[3]

[1]*Software Developer at Lab13 Srl, Tirana - Albania*
Email: esteriana.haskasa@gmail.com
[2]*Department of Computer Science, UAMD, Durres-ALBANIA*
Email: edlirakalemi@uamd.edu.al
[3]*CEO & Cofunder Kreatx sh.p.k, Tirana - Albania*
Email: lejdi.koci@kreatx.com

ABSTRACT

Information is very important nowadays; as a result transforming data into information has significantly increased the importance of using Data Mining. To give sense to data, data mining uses several techniques that developed within a field known *machine learning.* In this paper we will be taking a look at WEKA workbench, which is a collection of machine learning algorithms and data preprocessing tools. We will discuss how WEKA functions and what benefits does the use of this workbench give to us. Later we will reach in a concrete case of studying that implements WEKA workbench inside another application. We will illustrate in detail how the panels of the Explorer interface, the main interface of WEKA, use data mining algorithms to present the desired result of the explored data.

Keywords*: data mining, algorithm, processing, interface, data, machine learning, information.*

I. INTRODUCTION

Information is very important nowadays; as a result transforming data into information has significantly increased the importance of using Data Mining. Data mining is used for a management of a wide range of databases' types such are: relational database, huge dataware houses that are under construction, POS (Point of Sales): transactional DBs in terabytes, object-relational databases, distributet, heterogeneous and legacy databases, spatial databases(GIS), remote sensing database(EOS), scietific/engineering databases, and a huge, hyper-linked, dynamic, global information system.
It plays a key role to process information in a lot of areas as demonstrated in the following figure:

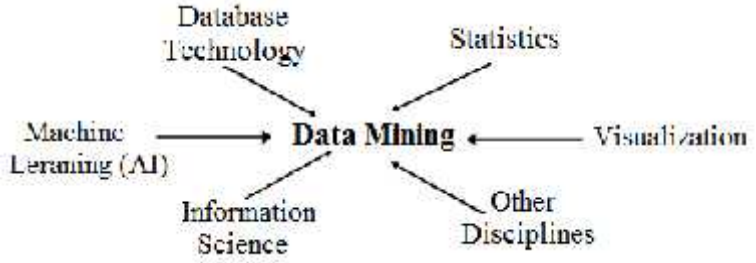


Fig 1. The areas where data mining has usage

Data mining is needed to make sense and use of data. It uses several techniques to meet this task. Most of them have developed within a field known as *machine learning.*
WEKA workbench is a collection of machine learning algorithms and data preprocessing tools. Since the beginning of the WEKA development, its developers intended to provide not only a toolbox of learning algorithms, but also a framework inside which can be easily implemented new algorithms.
Nowadays, WEKA has become a widely used tool for data mining. This workbench is offered as open source software. In this way the users has free access to the source of this application increasing its success.
In this paper we will analyze the WEKA workbench, its advantages and disadvantages and why it is used so widely today. Then we will give an example in which firstly we will generate from the database records the input file for WEKA, than we will embed WEKA workbench inside another application in order to display the results of this experimental data and finally we will interpret the data of the ARFF file through the explorer interface of WEKA workbench.

II. WEKA OVERVIEW

WEKA has a modular and extensible architecture that helps it to build up processes from the wide collection of base learning algorithms. Also it has a simple API that makes it easy to extend the toolkit and also to facilitate the integration of new learning algorithms with WEKA graphical user interface.
WEKA workbench supports the whole process of experimental data mining. This process begins with the phase of preparing input data, then it continues with evaluating learning schemes statistically, visualizing the input data and it ends with the result of learning. Inside this workbench are virtually included the top ten more used algorithms of data mining and are also included methods for the main data mining problems: regression, classification, clustering, association rule mining, and attribute selection. The most valuable resources that WEKA provides are implementations of actual learning schemes and tools for preprocessing the data, called filters.
When user fire up WEKA, then he has to choose among four different user interfaces:

Fig 2. WEKA GUI

the Explorer, the Knowledge Flow, the Experimenter and command-line interfaces. It is behind these interactive interfaces that the basic functionality of WEKA lies.

- *The Explorer interface*

The main interface of WEKA is the *Explorer* interface. This is the intefarce that we will use for our example. For most users it is the first window into the WEKA environment, and is the easiest and most intuitive interface to use WEKA. Explorer is designed to make it easy to start exploring one's data and to give access to classification and regression algorithms.
The advantages of this interface are: a) The strength of this interface lies in its simplicity, but it does this without hiding a lot of the functionality that is possible in WEKA, b) All of the various learning algorithms filters and visualization tools are available and accessible in this interface.
The disadvantages are: a) It is designed to work with one learning model at a time and there is no way to compare different models as to their performance. However this problem is addressed in Experimenter and Knowledge Flow interfaces. b) It has a lack of data editor.
Anyway when using WEKA we should take in consideration its on line documentation that gives the only complete list of available algorithms.

III. THE ADVANTAGES AND DISADVANTAGES OF WEKA

WEKA workbench has a wide use today that exceeds over most other data mining software. This result comes because of some advantages that the use of this WEKA workbench brings. First this workbench is open source and freely available under the GNU General Public License. So it is maintainable and modifiable without depending on the commitment, health or longevity of any particular institution or company. Second it offers a comprehensive collection of data pre-processing and modelling techniques. It provides a wealth of state-of-the-art machine learning algorithms that can be deployed on any given problem Third it offers portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform-even a Personal Digital Assistant. Fourth WEKA has an easy way of use because of its

graphical user interfaces. Even though WEKA includes a few algorithms to process data incrementally, for most of the methods the available memory imposes a limit on the data size. So the applications are obligated to have small or medium-size datasets. The second disadvantage is the flip side of portability: that mean a Java implementation may be somewhat slower than an equivalent in C/C++.

IV. EXAMPLE: EMBED WEKA IN OTHER APPLICATIONS

The purpose of this example was to show a concrete use of WEKA workbench. We saved the data for this example in a database named *'test'*. After connecting to MySQL and retrieving the data from the table '*ContactLent*' of the '*test*' database, we created a class named '*convertToARFF'* in which we used the data retrieved from the '*ContactLent*' to generate the ARFF file. In a simple demonstration the actions flow will be as follow (these are not the concrete code line used inside the program; they are just a logic demonstration):

```
// creates the new file with a name that will be given as input from the console, then in the main
method we will take  the name of the file inserted in the console.
       f=new File("Arff_File_Name.arff") ;

// creates the new file
       f.createNewFile(f);

// the try catch block, inside this block we will write the content that will later be used to
// generate arff file
try {
      BufferedWriter out = new BufferedWriter(new FileWriter(f));

// after we have taken the vector with data from the table Contact length then we will write the
//records one by one inside the file Arff_File_Name.arff
        for (int i =0; i< array_of_table_ContactLent.length; i++)
      {
             out.write(array_of_table_ContactLent['field1'][i]+"," +
             array_of_table_ContactLent['field2'][i] +….. + "\n");

       }out.close();
      }catch (IOException e)
      {
         System.out.println("Exception ");

       }
```

We have created auxiliary methods to extract the number and the name of the attribute that a table has. This program is independent from the table that mean the table given as input to the program is dynamic. As experiment we have took in consideration the '*ContactLent*' table.

To illustrate the format of the generated file ARFF we will use the data of Contact Lent table. The format is as below:

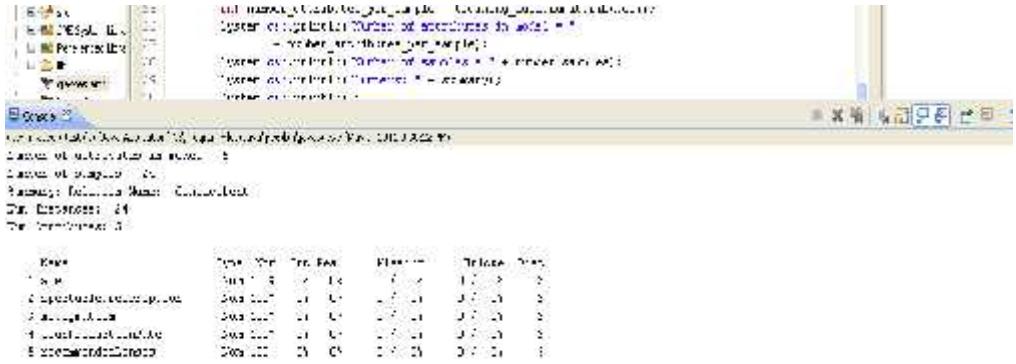
```
@RELATION ContactLent
@ATTRIBUTE age  {young,pre-presbyopic,presbyopic}
@ATTRIBUTE spectaclePrescription {myope,hypermetrope}
@ATTRIBUTE Astigmatism {yes,no}
@ATTRIBUTE tearProductionRate {reduced,normal}
@ATTRIBUTE recommendedLenses {none,soft,hard}


@DATA
young,myope,no,reduced,none
young,myope,no,normal,soft
young,myope,yes,reduced,none
young,myope,yes,normal,hard
young,hypermetrope,no,reduced,none
young,hypermetrope,no,normal,soft
……
```

The header of the ARFF file contains the relation that represent the name of the table and the attribute that are the columns of the table.
The relation has the format `@relation <relation-name>` , the attribute has the format `@attribute<attribute-name> <datatype>`

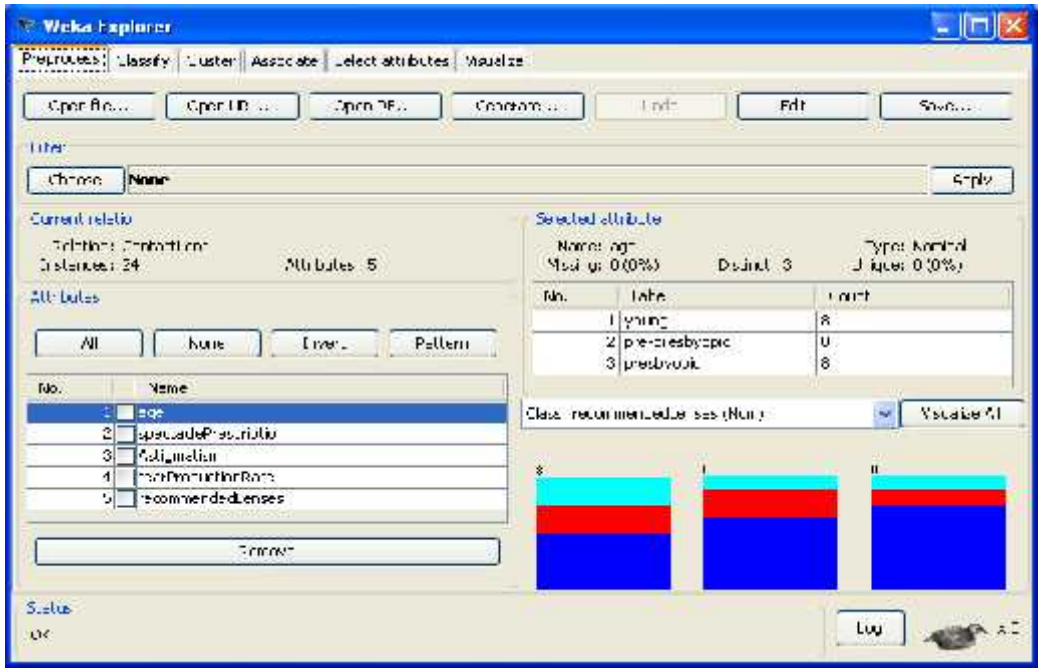
The datatype of the attribute supported by WEKA are numeric, string, date [<date-format>] that has the default "yyyy-MM-dd'T'HH:mm:ss" and <nominal-specification> for example `{reduced,normal}`

The body of the file contains the data or in other words the records of the tables as illustrated above.
In the class that embed the WEKA inside a java application we have change ADTree into JADETree because ADTree is not a two-class problem, as it is our problem, in contrast JADETree is a classifier tree for a two-class problem.
The result of the running is as follow:

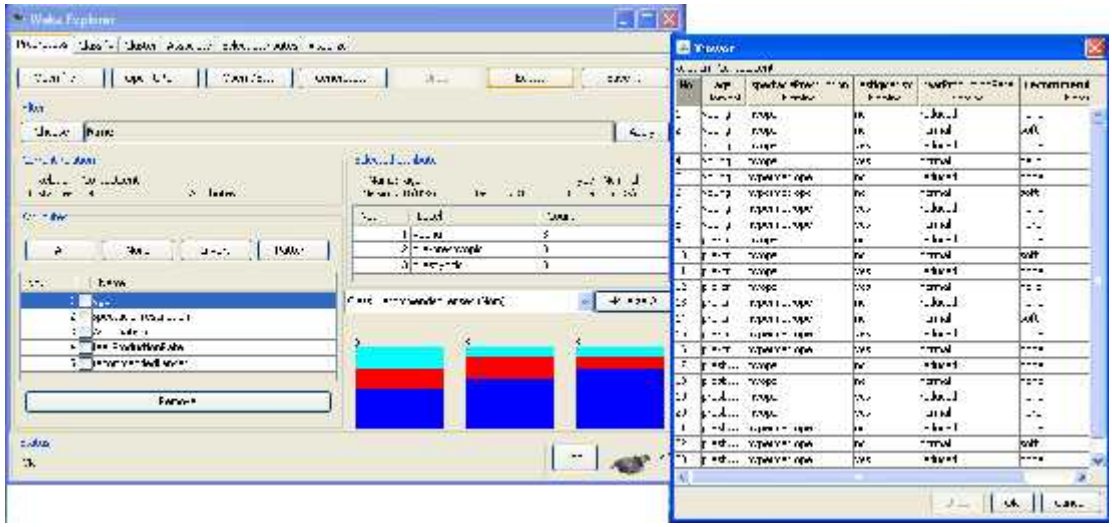


This output format may change according to the need of the person who wants to use it to accomplish a given task.
The file ARFF that takes as input this program that generate the above output is interpreted in WEKA workbench. The interface used for this example is *the Explorer*. The panels that this interface offers give access to the main components of the workbench.

- *The preprocess*
  The first panel of this explorer is the pre-process panel. This panel pre-process the data using filtering algorithm. These filters can be used to tranform data and also to make it possible to delete instances and attributes according to specific criteria. If we analyse the data of our arff file with this panel then we will have the below view.

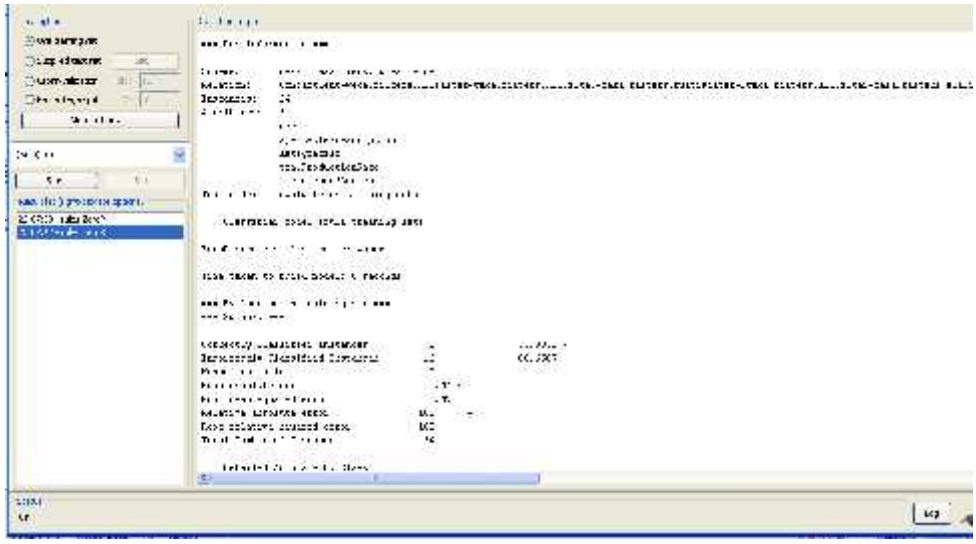


If we change the attribute then the value of the graphics will change too. We have the possibility to open a file, an url, a db etc sot all the type of the input described above for this interface. The edit button of this interface allow to change all the data entered in the table as below:



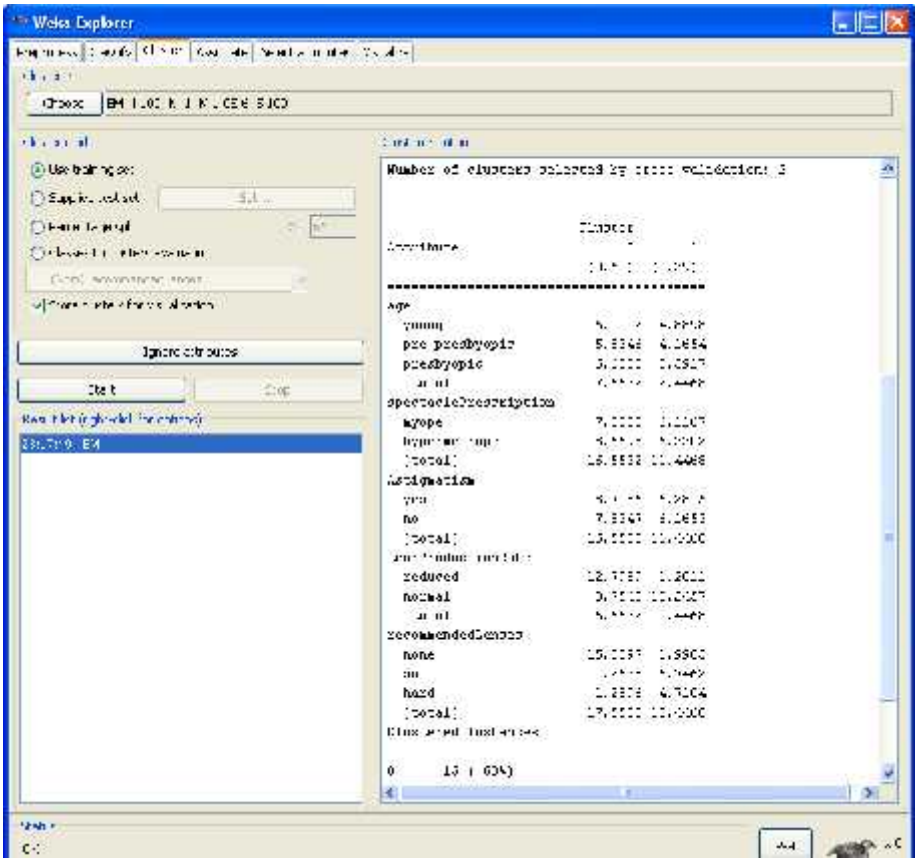
We also have the possibility to select a filter in this panel of the explorer.

- *The Classify* offer the possibility to apply to the resulting dataset the classification and regression algorithms. So this panel estimates the accuracy of the resulting predictive model and to visualise the decision tree for example. An ilustration of this is:



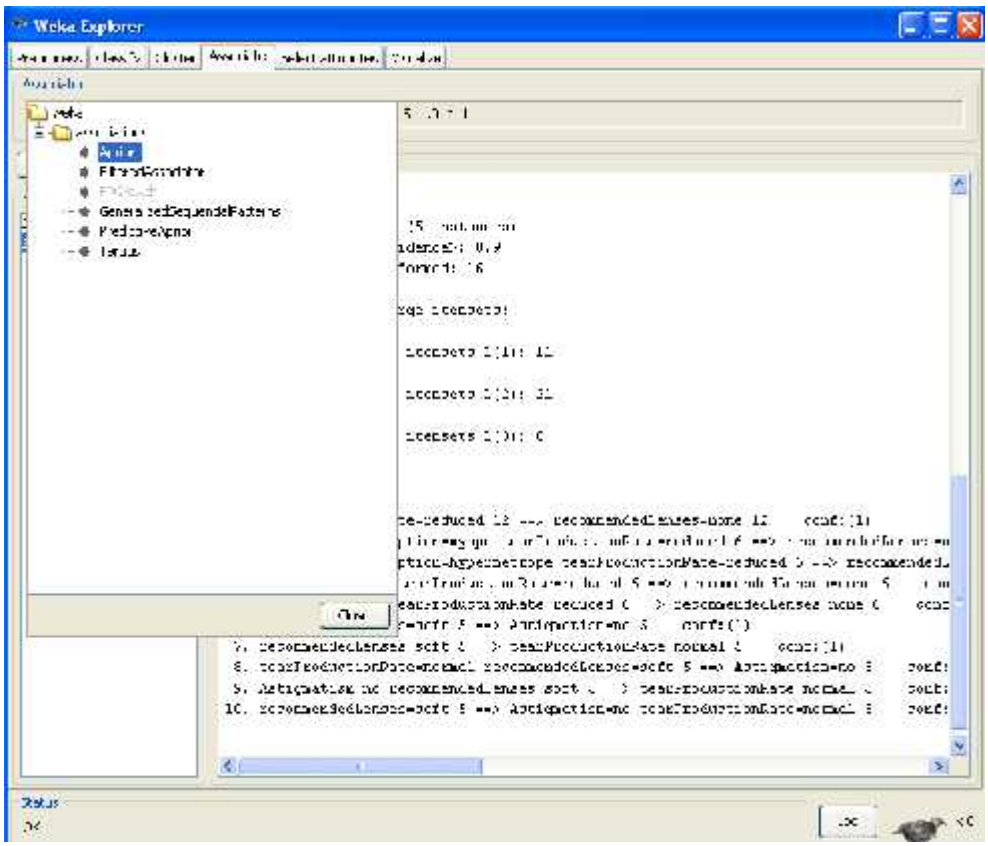
This panel continuos with java coding line.

- The *Cluster* panel give access to the clustering techniques in WEKA, for example in k-mean algorithm. It also implements the expectation maximization algorithm to learn a mixture of normal distributions. For our example the results will be as below.

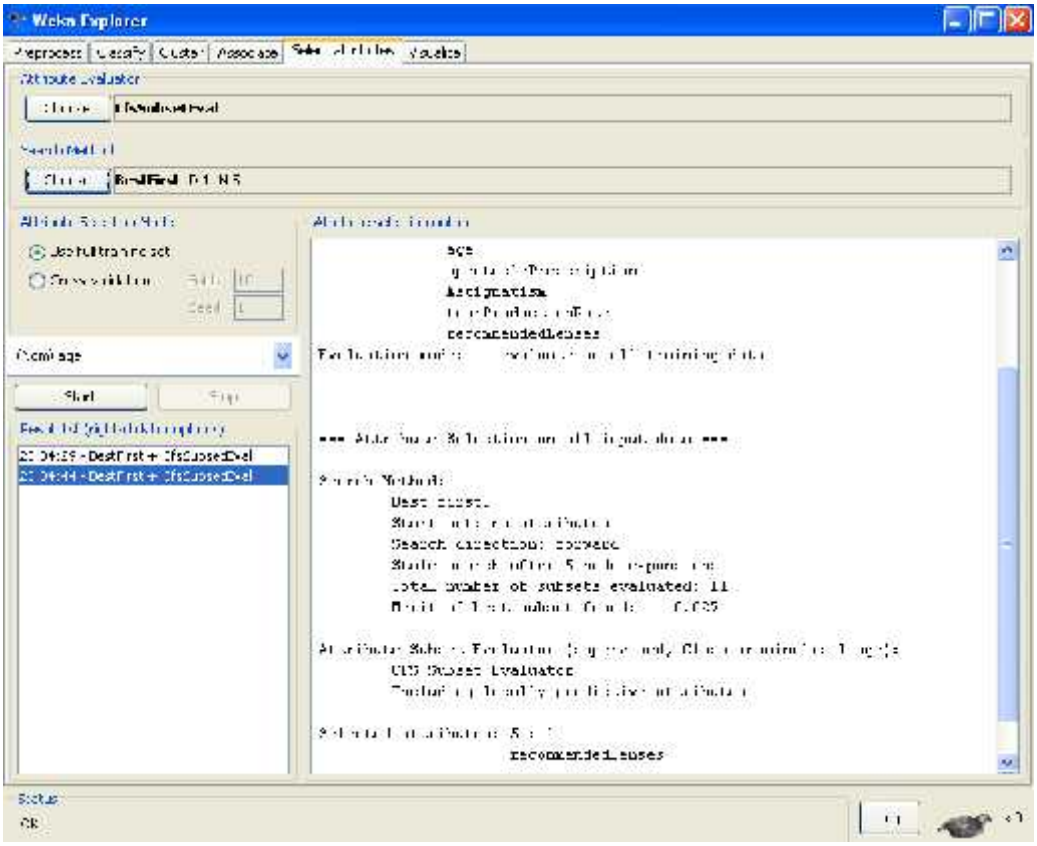


The above view will change if another type of clusters is chosen.

- The *Associate* panel attempts to identify all the important interrelationships between attributes in the data. It does this by accessing the association rule learners.



- The *Select* attribute panel provides algorithms in order to identify the most predictive attributes in a dataset. This panel offers the possibility to select the Attribute evaluator and the search method in order to have then the most predictive attributes dataset.



- The last panel, the *Visualize* one shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

WEKA workbench is completely written in java so it is easy to support the java packages and is more convenient to use java code. This is the reason why we have used java to develop our example.

CONCLUSION

In this paper we have presented WEKA workbench, a data mining tool, continuing with its main functionalities. We have argued why WEKA is a very useful workbench nowadays and what benefits brings the use of this workbench. We have discussed the advantages and disadvantages of WEKA. By a concrete case of studying developed by us, we have concretized these benefits and importance of using WEKA.

REFERENCES

[11] Ian H.Witten, Eibe Frank, Mark A. Hall: (2011) Data Mining, *Practical Machine Learning Tools and Techniques*, Third Edition. Part I Chapter 1,2, 3, 4; Part III Chapter 10;

[12] Chow, V. T. (1959) Open Channel Hydraulics. *McGraw-Hill Book Co*., New York, NY, USA.

[13] Ardiclioglu, M. Genç, O. Girayhan, A. and Kırkgöz, M.S. (2008) ADV Measurements of velocity distributions in natural rivers. *International Conference on Fluvial Hydraulics*, Çeşme, İzmir.

[14] Quang Nhat Nguye (2008-2009), Machine Learning: Algorithms and Applications, Faculty of Computer Scinece Free University f Bozen-Bolzano

[15] XindongWu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg : *Top 10 algorithms in data mining.*

[16] WekaDoc – http://wekadocs.come/node

[17] Laboratory Module 1 Description of WEKA (Java-implemented machine learning tool) http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab2-descriptionOfWEKA.pdf

[18] http://www.inf.ed.ac.uk/teaching/courses/dme/html/software2.html

[19] Mark Hall, Eibe Frank, Geoffrey Holmes, , Bernhard Pfahringer, Peter Reutemann, Ian H.Witten; '*The WEKA Data Mining Software: An Update'*

[20] http://www.cs.waikato.ac.nz/~ml/weka/arff.html

[21] 'Chapter 1, WEKA A Machine Learning Workbench for Data Mining'. Eibe Frank, Mark Hall, Geoffrey Holmes, Richard Kirkby, Bernhard Pfahringer, Ian H. Witten.

[22] http://blog.irodata.com/2011/02/embedding-weka-in-java-application.html

[23] http://en.wikipedia.org/wiki/Weka_(machine_learning)

[24] http://www.cs.waikato.ac.nz/~ml/weka/arff.html