T.C.

İSTANBUL KÜLTÜR UNIVERSITY

THE COMPARISON OF DATA MINING TOOLS

Data Warehouses and Data Mining
Yrd.Doç.Dr. Ayça ÇAKMAK PEHLİVANLI

Department of Computer Engineering İstanbul Kültür University
submitted by Mesut ÖZKAN

16.11.2011

Abstract

Data mining , a set of techniques used for the purpose of obtaining information from the data. Statistical analysis of data using a combination of techniques and artificial intelligence algorithms and data quality information in the disclosure of confidential information, a process of transformation. To perform data mining applications, including commercial and open source, many programs are available. This article is an open source data mining programs, RapidMiner (YALE), Weka and R are told, and these programs are included. Also WEKA in a sample application is presented. That graduate education is considered to be beneficial in all Institutes of the implementation.

İçindekiler

List of Figures

Introduction

Nowadays, databases that store this data and receive data from many sources, institutions, one of the objectives is to convert raw data to information. This process is the process of converting data into information is called data mining. In recent years, measuring devices in parallel with an increase in the number and types of data is increasing. Data collection tools and database technologies, developments, information requires large amounts of storage and analysis of information to be stored. The purpose of data mining methods in line with developments in computer technology and programs effectively and efficiently make large amounts of data. Data Mining was developed to combine the knowledge and experience in the software must be used. Rapidly growing data records (GB / hour), automatic stations, satellite and remote sensing systems, Space Telescope scans, developments in gene technology, scientific calculations, simulations, models, Data Mining has become compulsory.

Development in computer technology and the increased amount of data held in databases, new data collection methods, automatic data collection tools, database systems, increased use of computers, large data sources (Business: Web, e-commerce, shopping, stocks, ...), science world (remote sensing and monitoring, bioinformatics, simulations ..) society (news, digital cameras, YouTube, Facebook ...) responds to the question of why data mining (Kudyba, 2004).
Programs are needed to perform data mining applications. In this context, SPSS Clementine, Excel, SPSS, SAS, Angoss, KXEN, SQL Server, MATLAB, commercial and RapidMiner (YALE), Weka, R, C4.5, Orange, KNIME developed several programs, including open source.

Data Mining Open Source programs are addressed in this study, the programs were compared and a sample application. The second section, Data Mining, the third chapter of Open Source Data Mining Programs and RapidMiner (YALE), Weka and R-Programming are described. These programs were compared with the fourth chapter. Conducted with a sample application is presented in the fifth chapter of WEKA

Data Mining

Data Mining, held in data warehouses, and large amounts of a wide variety of location data reveal a previously undiscovered, use them to perform the process of decision-making and action plan. Large amounts of data, correlation, and the rules allow us to predict the scanning of the future. Data mining, data in the patterns, relationships, changes, irregularities, rules and structures those are statistically significant as the discovery of semi-automatic. The relationship between data, the computer is responsible for setting the rules and features. The goal is to identify previously unrecognized patterns of data.

Different types of data mining for the Implementation of an effective data handling, efficiency and scalability of data mining algorithm, the results of efficacy, precision, and to provide criteria for significance, discovered the rules representation in various ways, in different environments to make the data on the transaction, is required to ensure the privacy and data security features .
Alternatively, data mining is actually regarded as a part of the process of knowledge discovery. Knowledge discovery process stages are given below.

Data cleaning (to remove noise and inconsistent data)

Data Integration (combine multiple data source)

Data Selector (which is related to the performed analysis of the data set)

Data Conversion (make the transformation of the data to perform data mining technique that can be used)

Data Mining (Intelligent methods for data capture patterns of performance)

Pattern Assessment (according to some measurements that represent the information gathered to identify interesting patterns)

Information Report (which has been derived from mining to perform the presentation of information to the user), (Han M., 2001) (Delen, 2005)

Data mining studies for both commercial and open source programs have been developed to make. There are many algorithms in the programs. Algorithms that we have, using this data, meaningful information can be removed.

Open Source Programs Data Mining

Data mining applications is necessary to use a computer program to do. In this context, most software is developed. In this section, the Open Source Data Mining Programs and RapidMiner (YALE), Weka and R programs mentioned.

RapidMiner (YALE)

By scientists from Yale University in the United States was developed using Java language. A large number of data processed at Yale, meaningful information on these removed. Aml, arff, att, bib, CLM, cms, cri, csv, dat, ioc, log, matte, mode, ObF, a bar, one pair, res, sim, thr, wgt, WLS, xrff supports files with extensions. Yale does not support the format, such as a few other programs (http://sourceforge.net).
Support Vector Machine Learning Algorithms as learning models with a large number of the classification and regression, decision trees, Bayesian, logical clusters, association rules, and many algorithms for clustering (k-means, k-medoids, dbscan), all that is WEKA, for the separation of data preprocessing, normalization, features such as filtering, genetic algorithms, neural networks, there are many features such as 3D and data analysis. Algorithm has more than 400.Oracle, Microsoft SQL Server, PostgreSQL, or MySQL databases, data can be transferred to Yale. If the database management system is not supported, can be corrected by adding the jdbc driver class path variable.
Yale data set is expressed as XML. Below are a set of sample data.

```
<Attributeset>
  <Attributeset name = "Outlook" sourcecol = "1" ValueType = "nominal" blocktype =
"single value" classes = "rain overcast sunny" />
  <Attribute name = "Temperature" sourcecol = "2" ValueType = "integer" blocktype =
"single value" />
  <Attribute name = "Humidity" sourcecol = "3" ValueType = "integer" blocktype =
"single value" />
  <Attribute name = "Wind" sourcecol = "4" ValueType = "nominal" blocktype = "single
value" classes = "true false" />
```

```
  <Label name = "Play" sourcecol = "5" ValueType = "nominal" blocktype = "single
value" classes = "yes no" />
</Attributeset>
```

Feature not contained in the hundreds of other programs, such as very superior in terms of proximity to the user. Yale's first run, a new application can be created, saying New, Open, saying the existing applications can be opened in. The program is located within the sample for each algorithm.

WEKA

WEKA project began as a lot of people in the world today were introduced by the development program, a data mining application. Weka is open source software program that was developed on the Java platform. After starting Weka, as shown in Figure 1, the Application menu lists the modes can be studied. These Simple CLIcommand mode, which lets you work, the project step by step to realize the visual environment, providing Explorer and allows to perform the project using drag and drop KnowledgeFlow options.

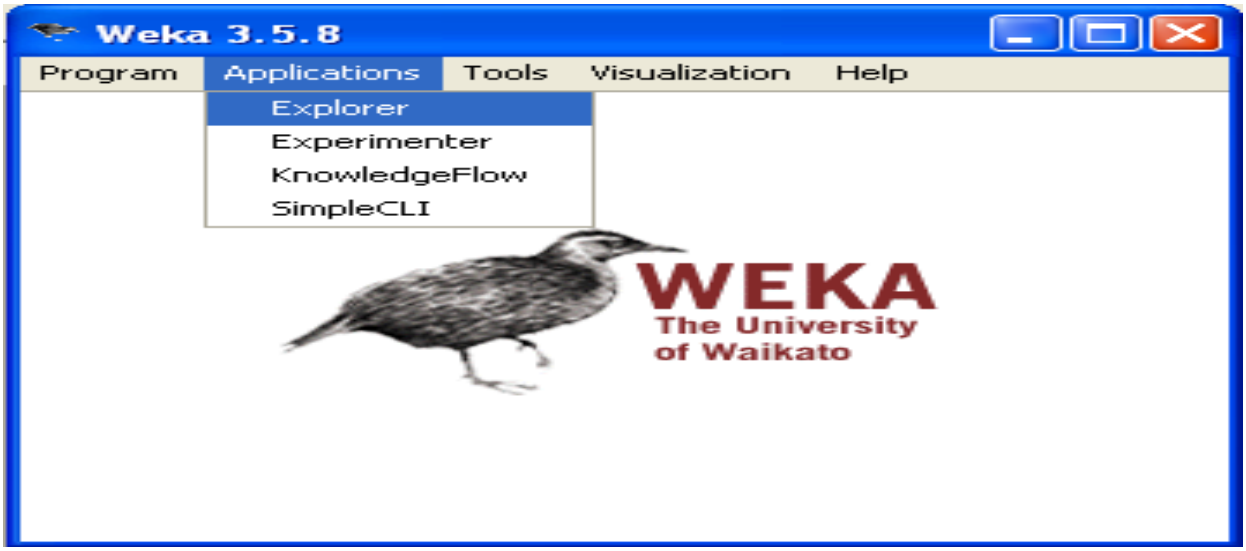


Figure 1WEKA Applications Menu

After you select the option to work on the Explorer of the
data selection, cleaning andconversion operations on the data that allows the screen to be realized are encountered.

Arff, csv, files WEKA in C4.5 format can be imported. WEKA data is impossible to handle with any text in soy. In addition, the operations can be done here to connect to the database using JDBC. WEKA in the Data Processing, Data Classification, Data Clustering, Data Association amenities are available.
After this step, which will be opened on the page according to the purpose of the project appropriate tabdaki (classification, clustering, association) by selecting the appropriate algorithm or algorithms that are applied to data that is accurate and that the algorithm can be selected.

R-Programming

Graphics, statistical calculations, a program developed for data analysis. Similar to the language of S is a GNU project. The University of Auckland in New Zealand Department of Statistics, Robert Gentleman and Ross Ihaka was developed by scientists. R & R is known as. R, with different applications is superior to the S language. Linear and nonlinear modeling, classical statistical tests, time series analysis, classification, clustering, such as the properties are consisting of. R, Windows, MacOS X and Linux systems can run on (http://project.org).

R is widely used for window systems. R on the X Window system is recommended. One of the most important features offered by the user of open systems with the X Window, Linux started to support from the moment of birth. Distributing free Linux distribution on the Internet has established itself as a standard under. X-Windows, based on client-server model works. Home on machine running the X server, and graphics hardware has on the entire input-output powers. An X client, the server connects to the server will request any transactions. The client's duty to give orders, to make the server appear in the order given in (Hania Gajewska, 1990). R to run on Windows or MacOS there is a need for specialist help. Users can employ the R mainly on UNIX machines.

R the following steps are suggested to run on UNIX machines.

To accommodate the need for the solution of the problem of data files-directory is created.
$ Mkdir work
$ Cd work

-R program is written to execute the following command.
$ R

-R is written the following command to quit the program.
> Q ()

- To learn the properties of functions written in the following commands.
> Help (solve)
>? Solve
in the shape of the data given below.

> Incomes <- c (60, 49, 40, 61, 64, 60, 59, 54, 62, 69, 70, 42, 56, 61, 61, 61, 58, 51, 48, 65, 49, 49, 41, 48, 52, 46, 59, 46, 58, 43)

Comparison of Programs

YALE, including Weka and R is the leader in open source data mining programs. In both ease of use as well as to host hundreds of feature makes it superior to Yale WEKA. 3D visuals are very helpful to the user at Yale excess. According to Yale, but the number of supported algorithms are easy to use WEKA less. YALE close to 22 file formats supported, WEKA supported file format is limited to the number 4. However, most of the WEKA data mining application development is adequate. Therefore, most

users prefer WEKA. R is the ease of use and supported by both algorithms and Yale and is located under WEKA. R is widely used on UNIX machines. R on the Windows system wants to use the expert help. Therefore, R, YALE, and more than WEKA not preferred.

In 2007 a survey was obtained as a result of information given in Figure 2. This survey is a site where the site visits data mining experts. The first bar, the options selected, only the votes of one of the second bar represents the votes represent a few selected option.

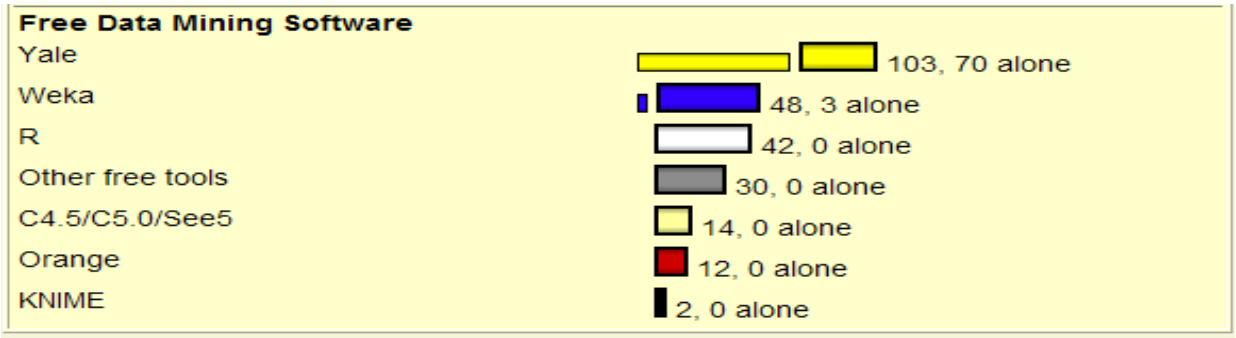


Figure 2 Open Source Data Mining Programs (kdnuggets.com)

In 2008, RapidMiner (YALE) (Figure 3), and Weka (Figure 4) are given below for the download charts. As is evident from the WEKA graphics, RapidMiner (YALE) to have been more than the download. Reason for the difference between the statistics and the statistics above, WEKA of the more popular among professionals despite RapidMiner (YALE) is afforded the more popular.

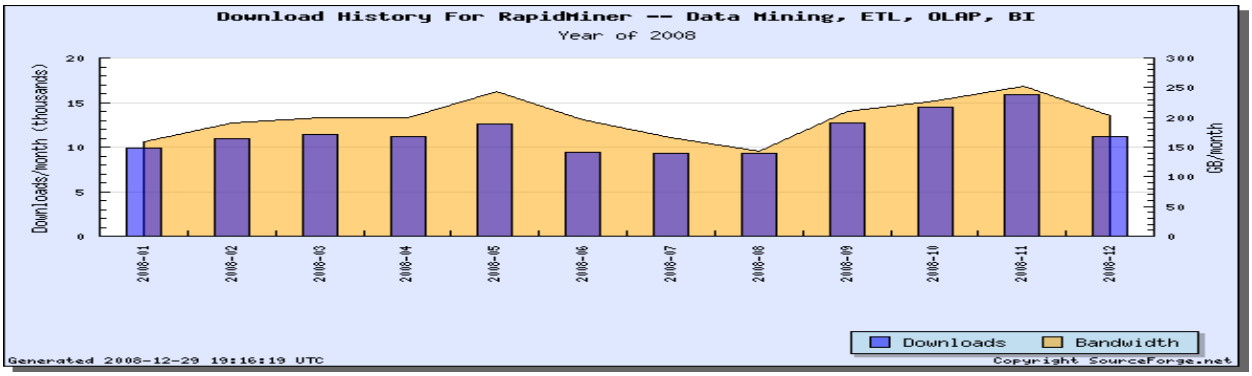


Figure 3 RapidMiner (YALE) numbers for Download (sourceforge.net)

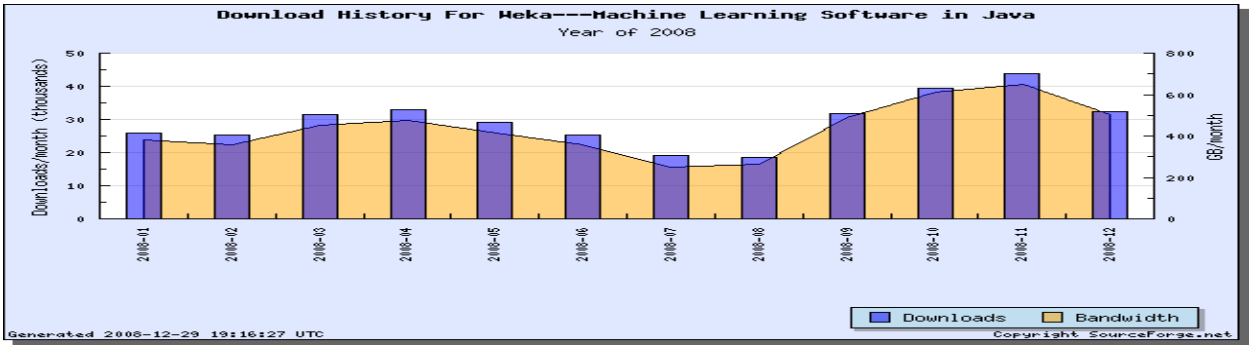


Figure 4 WEKA numbers for Download (sourceforge.net)

Application Example

Data Mining with WEKA

WEKA is java based open source data mining tool which has collection of data mining algorithms such as lazy, rules, decision trees and so on. WEKA opens with 4 options (Explorer, Experimenter, KnowledgeFlow and Simple CLI). Mainly, Explorer and Experimenter are used for data mining. For multiple algorithms comparison, Experimenter is used but for specific results of data mining, Explorer is used. Explorer opens with a screen of data preprocessing. The difficulty on WEKA is opening file because most of data sets are in excel and excel can turn into CSV format but excel file is semicolon but CSV must be for comma, so it needs to convert in text file format but it takes time. However; information on algorithms (capabilities and descriptions) and user options are the best features about WEKA, especially any user can use without any training as well as user can implement its own algorithm.

As an example, WEKA 3.6.5 is used in Windows XP Home Edition. Iris dataset with Unsupervised Normalization and Principal Components is used for Naïve Bayes algorithm. It can be seen in figure 5 and figure 6.

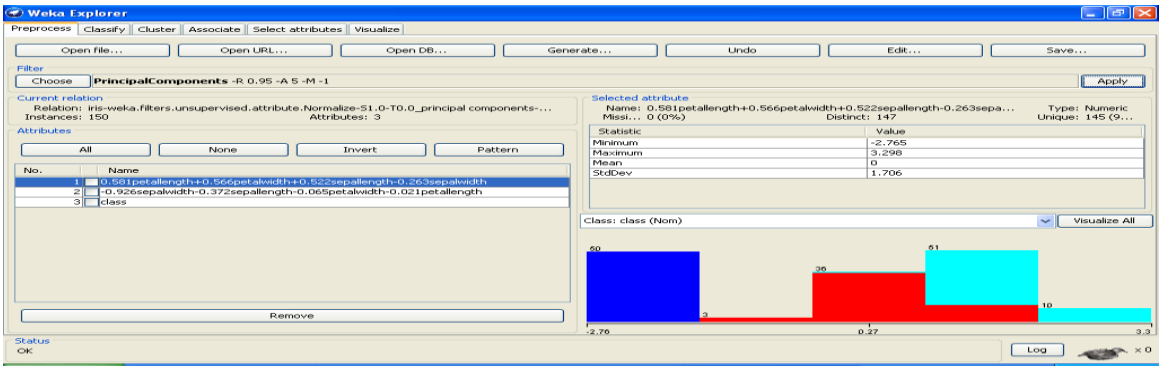


Figure 5 Data preprocessing screen in WEKA

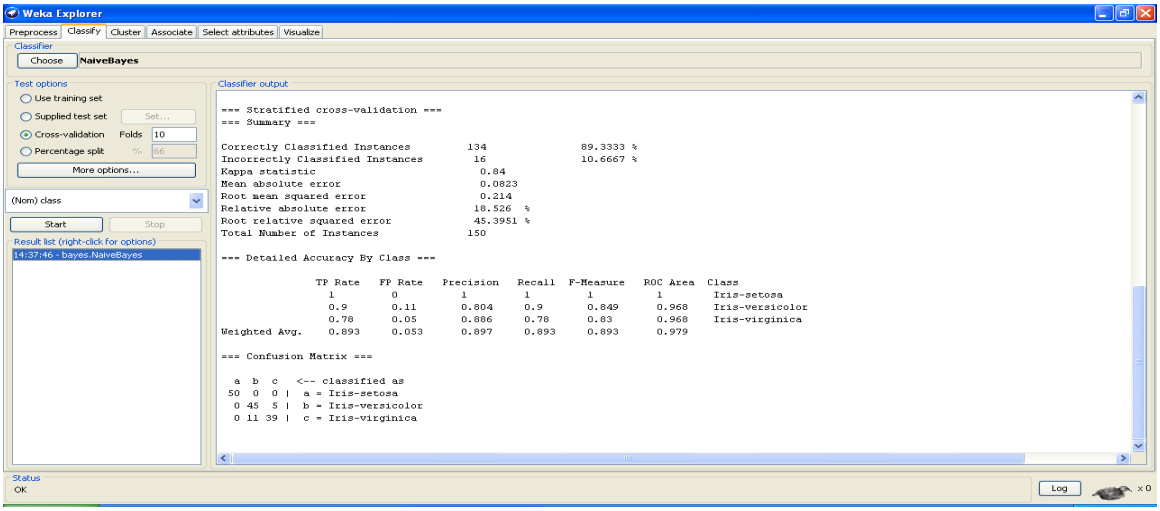
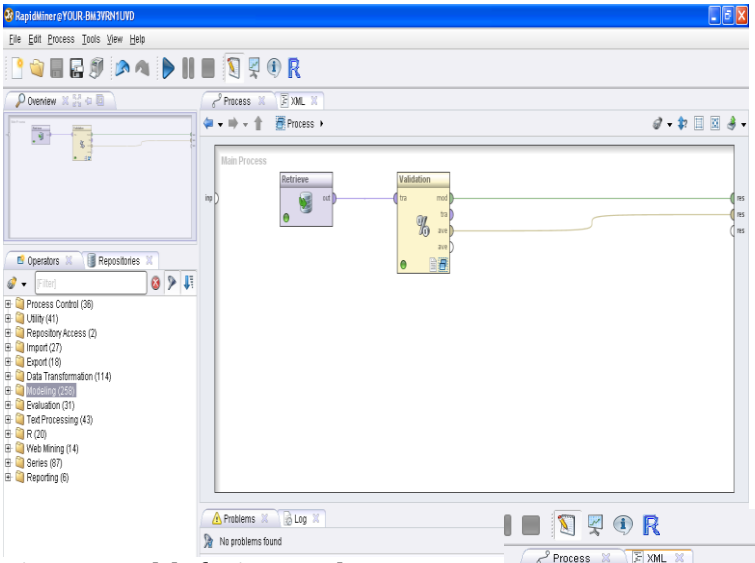


Figure 6 Result Screen for Naïve Bayes Classification in WEKA

Data Mining with RapidMiner



By “**Edit**”,

Choose “**New Building Block**”

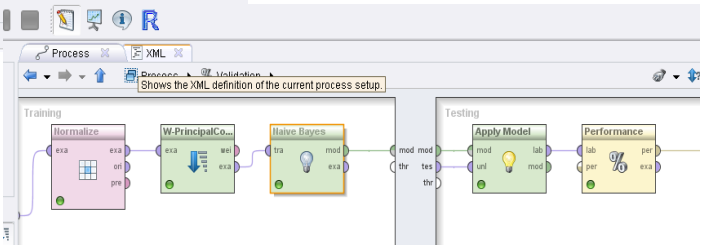


Figure 7 Model of Iris example on RapidMiner with normalization and Principal Component for Naïve Bayes Algorithm.

RapidMiner is another data mining tool which has ARENA simulation like environment as it is seen in figure 7 because every process is described in a similar manner. RapidMiner can use every algorithm in WEKA as well as its own algorithm and R programming can be open in RapidMiner with a connection like any other algorithm. Maybe, its difficulty is that’s not easy to use and the result is only based on confusion matrix. Same example seen in WEKA is applied with normalization and principal component for Naïve Bayes algorithm. The result of example can be seen in figure 8.

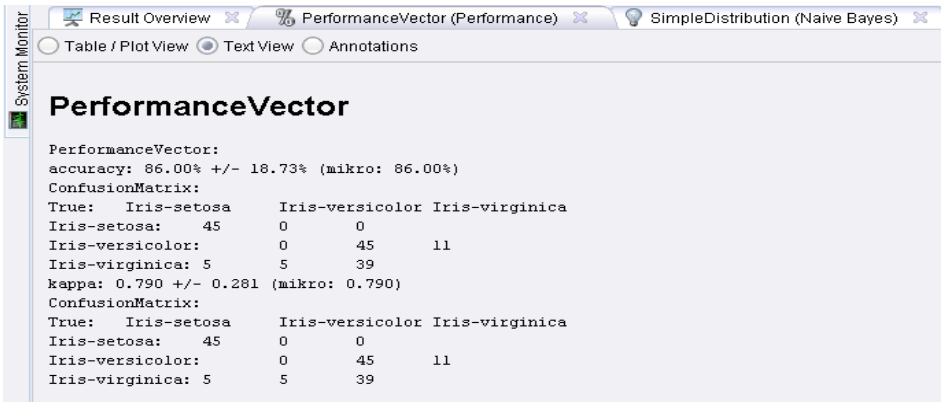


Figure 8 the result of the example in RapidMiner

Data Mining with R-Programming

R-project is very similar to Matlab and it uses same logic and user can easily load CSV format file and work with functions. However; first of all, packages must be downloaded like in figure 5. However; WEKA preprocessing function is used because there is no available preprocessing function and R project is only used for Naïve Bayes algorithm, confusion matrix and summary of the data. These results with used functions can be seen in figure 6.

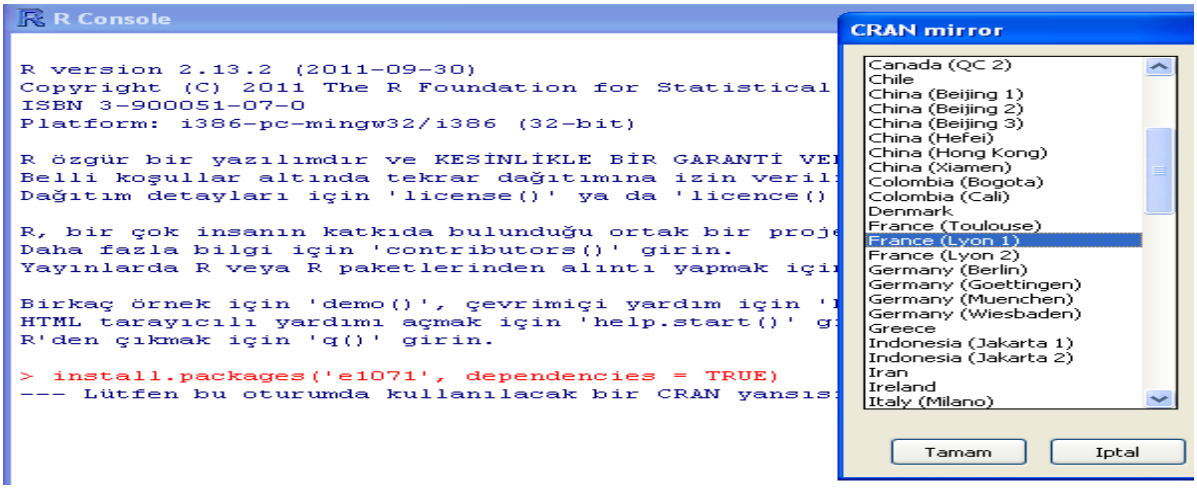


Figure 9 Loading Naïve Bayes package

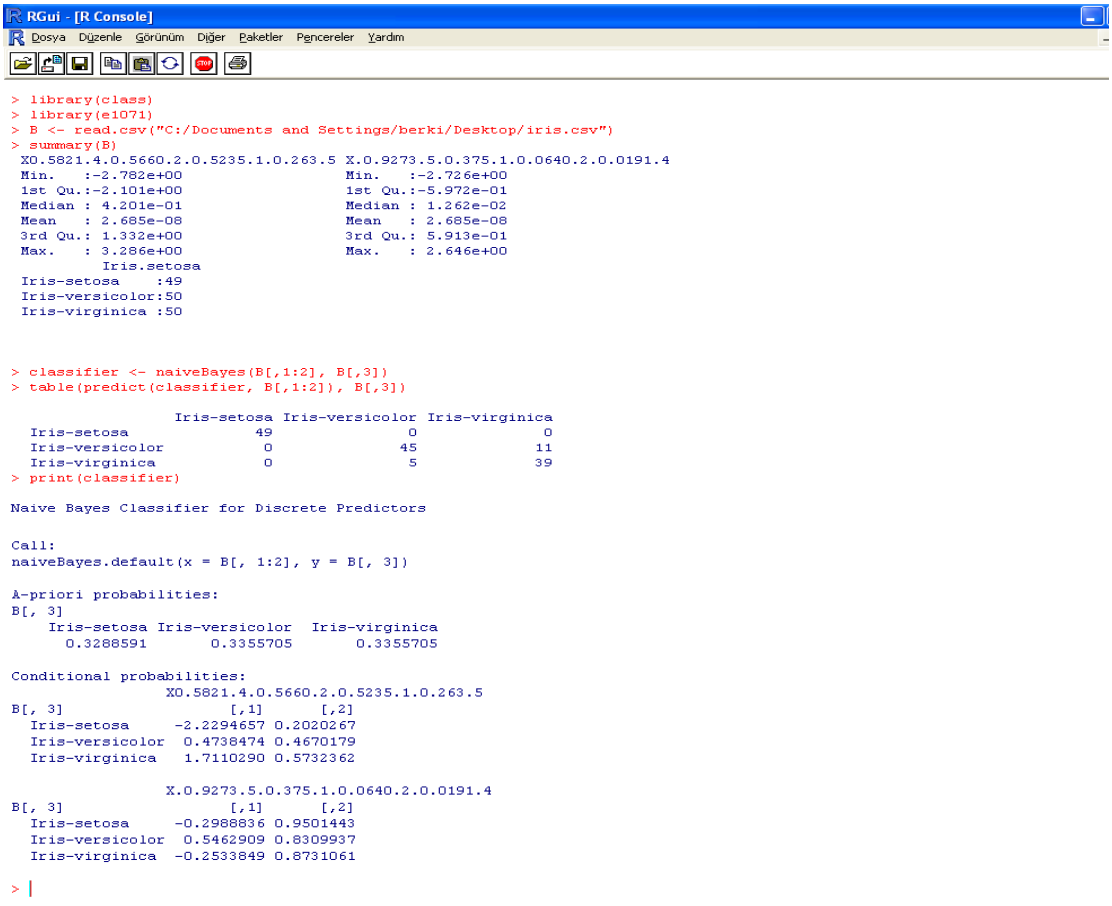


Figure 10 Code of Naïve Bayes algorithm is used with the result

Discussion and Conclusion

The growing problem of inaccessibility of information due to the amount of data derived from the area is known as data mining. Computer programs are needed to make data mining applications. These programs are in the data clustering, decision trees and Bayesian classifiers, such as the a priori method is available in many algorithms. Thanks to the algorithms processed the data, information extraction can be performed. In this study, Open Source Data Mining programs, RapidMiner (YALE), Weka, R is explained and the differences are emphasized. WEKA' were the most commonly used data mining program. WEKA an example application is presented.

References

(tarih yok). 11 15, 2011 tarihinde http://sourceforge.net: http://surfnet.dl.sourceforge.net/sourceforge/YALE/rapidminer-4.2-tutorial.pdf adresinden alındı

(tarih yok). 11 15, 2011 tarihinde http://project.org: http://cran.r-project.org/doc/manuals/R-intro.pdf adresinden alındı

(tarih yok). 11 15, 2011 tarihinde kdnuggets.com: http://www.kdnuggets.com/polls/2007/data_mining_software_tools.htm adresinden alındı

(tarih yok). 11 15, 2011 tarihinde sourceforge.net: http://sourceforge.net/project/stats/detail.php?group_id=5091&ugn=yale&type=prdownload&mode=year&package_id=0&release_id=0&file_id=0 adresinden alındı

(tarih yok). 11 15, 2011 tarihinde sourceforge.net: http://sourceforge.net/project/stats/detail.php?group_id=5091&ugn=weka&type=prdownload&mode=year&package_id=0&release_id=0&file_id=0 adresinden alındı

Delen, D. W. (2005). *Artificial Intelligence in Medicine.* sdfdsf.

Han M., J. K. (2001). *Data Mining Concepts and Techniques.* Morgan Kaufmann Publishers.

Hania Gajewska, M. S. (1990). *Why X Is Not Our Ideal Window System , Software — Practice & Experience vol.*

Kudyba, S. (2004). *Managing Data Mining.* CyberTech Publishing.