Vol. 11(8), pp. 114-120, November 2018
DOI: 10.5897/AJMCSR2018.0669
Article Number: 9EF53DD59307
ISSN: 2006-9731

http://www.academicjournals.org/AJMCSR

African Journal of Mathematics and Computer Science Research

Full Length Research Paper

Comparison of open source data mining softwares on a data set

Abdullah BAYKAL[1*] and Cengiz COŞKUN[2]

[1]Department of Mathematics, Faculty of Science, Dicle University, Diyarbakır, Turkey.
[2]Department of Economics Faculty of Economics and Administrative Sciences, Dicle University, Diyarbakır, Turkey.

Received 22 September, 2016; Accepted 1 November, 2018

Data mining is the process of extracting informative and useful rules or relations, that can be used to make predictions about the values of new instances, from existing data. A wide range of commercial and open source software programs are used for data mining. In this study, a comparison of several classification algorithms included in some open source softwares such as WEKA, Tanagra and Scikit-learn using SEER (Survillance Epidemiology and End Results) data set which consists of 60948 instances is performed.

Key words: Data mining, classification analysis, open source data mining tools.

INTRODUCTION

A wide range of algorithms can be used to extract information from data for data mining purposes. This information then can be used to make future predictions about new instances. Application areas of data mining continually grows with the development of technology and needs. One of the application areas of data mining is health systems among very many others. There are various data sets related to health systems (Banu Rahaman et al., 2010). Despite the availabilty of abundant data in health systems, usually data is not sufficiently clean or information on these data is mostly insufficient.

There are many different algorithms used for data mining. Because of its application on a wide range of different areas, studies in data mining technology are going on continously, new methods are being developed and enhancements to the existing ones are taking place continously. Also, the progresses on different disciplines such as Mathematics, Statistics, İnformatics and Computer Science help improve the methods used in data mining. Thus, data mining, having a very vast application area and having relations to other disciplines, is open to new developments and attracts interest not just by academic world but also by the business world.

Process known as Knowledge Discovery from Databases (KDD), has become a popular analysis method among the health researchers in order to discover and define the patterns and relations between vast amount of variables and paves the way to predict any future value for the output variable.

*Corresponding author. E-mail: baykal.abdullah@gmail.com.

In this study, a comparison of several classification algorithms in data mining on health data has been conducted. Evaluation of the experiment results was realised by the accuracy and error rate occured with the constructed model. This study provides an insight in determining the best classification method according to the accuracy and the error of the realised experiments.

There are many free and commercial data mining tools available. In data mining applications, generally these tools are used. In this study, classification algorithms provided by open source tools WEKA, Tanagra, and Scikit-learn were compared.

LİTERATURE SURVEY

"Knowledge Discovery from Databases" (KDD) and "Data Mining" terms are frequently used interchangeably. Hand et al. (2001) define data mining as "the process of discovering interesting knowledge from large databases". Witten and Frank (2000) define it as "the extraction of implicit, previously unknown, and potentially useful information from data".

Previous studies on the comparison of data mining tools, examine a subset of the available tools mostly limited to a small number of available softwares either free or commercial. In these studies, mostly the user friendliness, interface design, algorithms provided, and platform compliances are considered. An example that can be given is Elder et al. (1998) who present a comparison of several free and commercial data mining softwares.

In 1994, a team of twelve researchers, six of which affiliated to academics and the other six being industrial researchers have conducted a research on classification algorithms under European Stat Logs Project. The results of this study was published with the name "Machine Learning, Neural and Statistical Classification" (Michie and Spiegelhalter, 1994).

Abdullah et al. (2011) designed a template for characterization based on several dynamic sample databases of tools and other supporting attributes like services provided by system, business goals, and other features of processing data and user interface. Around 40 data mining tools were evaluated and a general schema was proposed for tool selection to achieve business goals.

Lin et al. (2004) used various discriminant techniques, including Fischer's discriminant analysis and kernel-based discriminat analysis in order to place students into three categories according to their level properly. They concluded that kernel-based approach with bandwith selected by cross validation performs reasonably well for categorizing the students according to their level.

It can be seen from the previous studies that the success of the algorithm used substantially depends on the data set used. Thus, most of the similar studies conclude to achieve contradicting results.

Use of classıfıcatıon algorıthms on health data

Data mining is being used in medical area in order to improve decision making. One of the most challenging difficulties in medical systems is extracting comprehensible information from diagnosis data. Here, classification methods used for data mining in health related diagnosis data is reviewed.

Gandhi et al. (2010) proposed constructing classification rules using the Particle Swarm Optimization Algorithm for breast cancer data set. The model generated with the constructed rules was reported to achieve accuracy rate defining the underlying attributes effectively.

Manaswini and Ranjit (2011) used an artificial neural network based classification method to classify the diabetes patients. They used Pima Indians Diabetes Database for their study. They adopted 10-fold cross validation on both training and test data.

Kahramanli and Allahverdi (2009) used an adaptive activation function in training the artificial neural network and then used artificial immune algorithm to extract rules for liver disorder data set. Their proposal takes all the input attributes into consideration, and extracts rules from the trained neural network efficiently.

DATA SET USED

This study was conducted using Surveillance Epidemiology and End Results (SEER). SEER, being a unique, well documented, reliable data source containing patient data on several kinds of cancer, has an important place in scientific researches. It provides important statistical information containing the instances from several states of USA, affiliated to 26% of the population, is provided by National Cancer Institute (NCI). SEER, annually being upgraded, has been a fruitful source for thousands of scientific researches.

Data set, since 1973, the time it started, contains different types of cancers in text format. It is a big source consisting of 118 attributes. Since the start of the data collection, attribute set have changed in time. Some attributes related to some cancer types are not applicable to some other cancer types. Some attributes found to have lost its meaning or importance by time. These kinds of attributes had been removed from the data set definition on later versions of the set. On the other hand, new attributes had been introduced in newer versions of the set as new parameters arouse because of the development of technology and medical science.

In this study, despite the data being well formed and well documented, it was a necessity to do some preprocessing before the application of data mining algorithms. The set used in this study, contains 118 attributes.

Preprocessing of data

Attributes in the set have been analyzed thoroughly and some of the attributes thought to have an ignorable effect on the output classes were elected as part of the attribute selection process.

Also, some non-informative or meaningless attribute values that may cause problems in the analysis of the data have been replaced with proper values as the the following.

Age at diagnosis

This attribute indicates the age of the patient at diagnosis and is 3 characters long with values ranging from 000 to 130. A value of “999” indicates that the age of the patient is unkown. This value “999”, has been replaced with a “?” character to reflect the meaning that it is missing value.

Regional nodes positive

Being a numeric attribute, the maximum value for this attribute is 90. The values above 90 have the meanings, that can be put in the same category, such as “Positive aspiration or core biopsy of lymph node(s)”, “Positive nodes-Number unspecified”, “No nodes examined”, “Unknown-Not applicable, not documented in patient record”. Thus, such values above 90 is replaced by “?” character to represent the attribute as missing value.

CS tumor size

This attribute had been included in the set after 2004, so is absent for the data gathered before 2004. But yet, data set gathered before 2004 has an attribute named “EOD-Tumor size”. Thus, EOD-Tumor size values for pre-2004 data were taken as the values for CS Tumor Size. Values “989” and “999” meaning “no information” were replaced by “?” character. Values 991, 992, …, 997 meaning “maximum 10 mm”, “maximum 20 mm”, …, “maximum 70 mm” were reaplced by “010”, “020”, …, “070”, respectively.

Tumor marker

This attribute has no value after 2004. Such values have been replaced by “9” meaning “Not known, or not entered”.

RX Summ-Surgery of primary site

Empty values for this attribute were replaced with “99” meaning “Unknown if primary surgery performed”. Codes having values in the range of 10 to 19 with similar meanings were replaced with “10” and codes in the range of 20 to 80 were replaced with value “20”.

RX Summ-Scope of regional lymph node surgery

Empty values for this attribute were replaced with “9” meaning “Unknown/Not applicable”.

RX Summ-Surgical procedure of other site

Empty values for this attribute were replaced with “9” meaning “Unknown”.

Assigning the classes

Despite there is no indicator of the survival of the patient in the data set, there are fields “Survival Time Recode, STR”, “Casue of Death to Seer Site Recode, COD”, and “Vital Status Recode, VSR” that we can deduce the survival status of the patient from, and then assign a class value for each instance. Delen et al. (2004) used STR field for this purpose. In addition to the STR field, Bellaachia et al. (2006) also uses VSR and COD attibutes to assign the class value. In this study, we employed the following algorithm to deduce the class, that is, the survival status of the patients due to cancer.

```
if (VSR ?= '1')          /* patient alive? */
      { if ( STR > 60)
          Class= '1';          /* survived cancer */
      }
else                     /* not alive */
      { if (COD=cancer)                /* cause cancer? */
          Class='2';          /* not survived cancer */
      }
```

The instances not classified as neither “1” or “2” according to the algorithm were ignored. The experiments were performed with 60948 instances remaining after the preprocessing step.

Data mining tools

The data mining tools that were compared in this study are shown subsequently.

Weka

Weka is an open source data mining tool developed at Waikato University. It is implemented in Java and is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. It also enables to develop and add new machine learning schemes. Weka is an abbreviation for “Waikato environment for knowledge analysis”. It has a modular design and provides functions to preprocess the data, visualisation of data, and a large scale of algorithms for data mining and is well suited for business intelligence. Weka has its own ASCII text file format definition named “Attribute-Relation File Format”, ARFF. It provides tool to convert any csv file to arff format. Weka can process files of type arff, csv, and C4.5. Also it lets to connect to a database with Jdbc connection and work on data in the database.

Tanagra

Tanagra is a free version of data mining software that was developed by Ricco Rakotomola at Lumière University Lyon for research and academic purposes. It provides facilities for visualization, statistical value calculations, sample selection, attribute selection, attribute creation, and tools for regression, clustering, factor analysis, classification and association rules. User can design a data mining process visually by the help of diagrams. Every node in the diagram is either a statistical or machine learning algorithm and the connection between two nodes represents data flow. Tanagra lets the use of files with extensions txt, xls, arff and dat. Results can be viewed in html format, therefore it is easy to represent the results on any web explorer.

Scikit-learn

Scikit-learn is one of the most used python library on machine learning. It comes with Anaconda and WinPython and includes many methods such as linear regression, logistic regression,

Table 1. Data set definition.

Attributes	Number of Instances	Number of Classes
7	60948	2

Table 2. Results from Weka**.**

Algorithm	Accuracy rate	Error rate
Naive Bayes	81.70	18.30
J48	84.23	15.77
Random Tree	81.64	18.36
KStar	85.44	14.56
Logistic	81.57	18.43
SMO	81.02	18.98
IBk	71.64	28.36

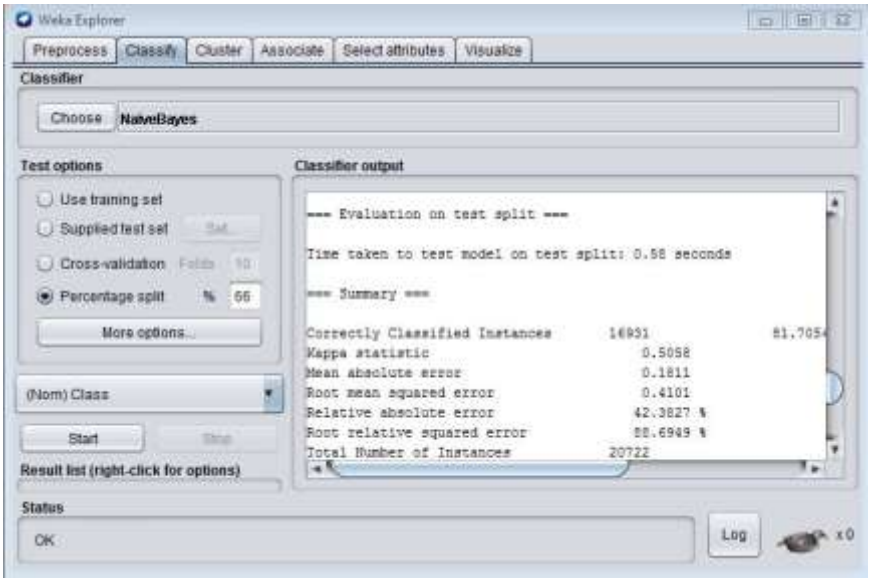

Figure 1. Result screen of Weka.

decision trees, and random forests. It is possible to download the python package from its web site.

Scikit-learn includes most of the data mining algorithms, and provides tools for data analysis. It prohibits the need for extra tools by providing modules for data cleaning, attribute selection, cross validation, and result interpretation.

Experimental analysis

In order to achieve the goalsof this study, data mining tools including Weka, Tanagra and Scikit-learn were used. Accuracy and error rate of the test results were used to interpret the results. A high value for accuracy and a low value for error rate in a test of a classification of a data set indicates the success of the model.

$$Accuracy = \frac{Number\ of\ correct\ Predictions}{Total\ number\ of\ Predictions}$$

For the experiments, the data was split into two parts for training and testing. Training set was used to construct the classification model. Training set was selected to include 66% of the instances and the rest of the instances, that is the 34% constitute the test data. A description of data set used in this study is given in Table 1.

RESULTS OF EXPERIMENTS

Results achieved with Weka is given in Table 2 and the result screen of Weka is given in Figure 1.The best result

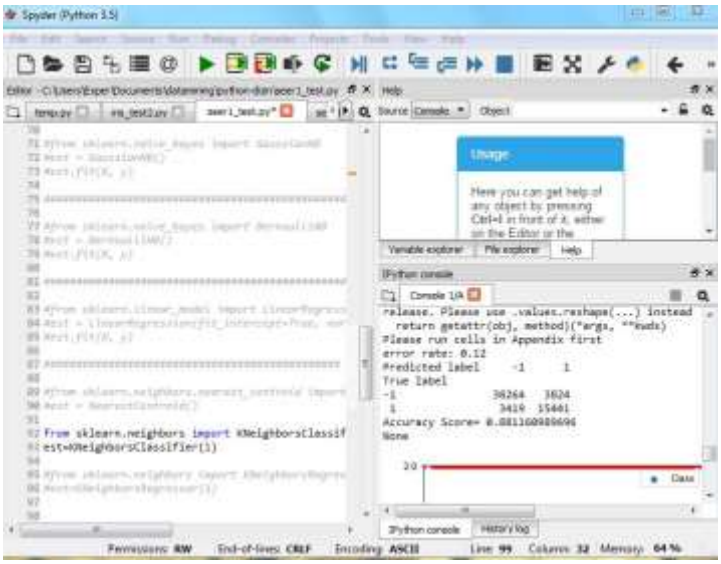

Figure 2. Result screen of Scikit-learn.

Table 3. Results from Scikit-learn.

Algorithm	Accuracy rate	Error rate
Decision Tree Classifer	91.84	8.16
Naive Bayes GaussianNB	81.06	18.94
Naive Bayes BernoulliNB	81.94	18.06
Linear Regression	80.46	19.14
Nearest Centroid	68.15	31.85
KNeighbors Classifer(k=1)	88.11	11.89
KNeighbors Classifer(k=2)	86.87	13.13

Table 4. Results from Tanagra.

Algorithm	Accuracy rate	Error rate
C4.5	87.26	12.74
ID3	83.52	16.48
Naive Bayes	80.51	19.49
kNN	86.62	13.38
Rnd Tree	87.75	12.25
Logistic Reg.	82.05	17.95
Linear dis analysis	81.45	18.55

achieved with Weka is obtained with KStar with an accuracy of 85.44%, followed by J48 algorithm with an accuracy of 84.23%. These results were then followed by 1Bk algorithm with an accuracy of 71.64%.

To construct the models with Scikit-learn, a python program was implemented (Figure 2), and the models were created by using this program.

As can be seen in Table 3, the best result achieved with Scikit-learn was obtained with Decision Tree algorithm with an accuracy of 91.84% and the worst result was with the Nearest Centroid algorithm with an accuracy of 68.15%. It was also observed that the accuracy with KNeighbors classification algorithm decreases with the increasing values of k.

As can be seen in Table 4, the best result with Tanagra was achieved with the Rnd algorithm with an accuracy of

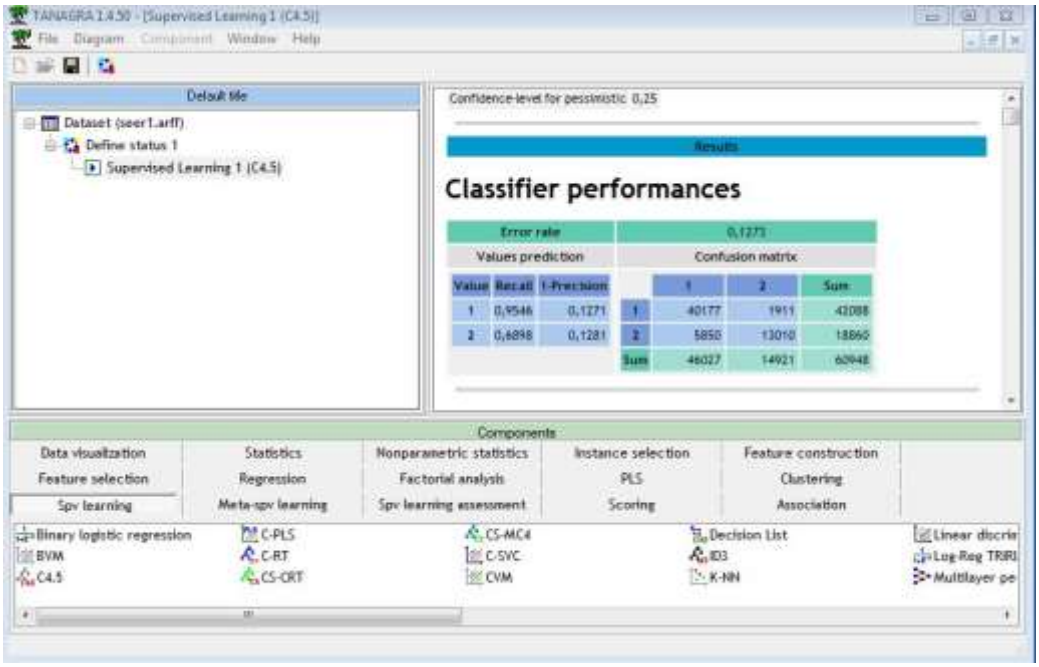

Figure 3. Result screen of Tanagra.

Table 5. Best results obtained from the experiments.

Program	Algorithm	Accuracy rate
Scikit-learn	Decision Tree	91.84
Tanagra	Rnd Tree	87.75
Weka	KStar	85.44

87.75 and the worst result was with the Naive Bayes Algorithm with an accuracy of 80.51%.
The result screen of Tanagra is given in Figure 3.

Conclusion

The experiments in this study were performed on 60948 instances taken from the SEER data set. Seven classification algorithms were run on Weka, Tanagra and Scikit-learn on the same computer.

As can be seen from Table 5 the best results obtained were the result of the experiment with the Scikit-learn tool with the Decision Tree Algorithm. This was followed by Tanagra and then by Weka.

Of the three tools compared in this study, Scikit-learn gives the best results. However, since it requires Python programming skills to construct models, usage of this tool is a bit more bothersome compared to the other two that provides visual interfaces.

It can be said, as a result of this study, that Scikit-learn software is a good alternative among the three tools compared.

CONFLICT OF INTERESTS

The authors have not declared any conflict of interests

REFERENCES

Abdullah HW, Qasem AA, Mohammed NA, Emad MA (2011). “A Comparison study between data mining tools with some classification methods”, International Journal of Advanced Computer Science and Applications 8(2):18-26.

Banu RS, Shashi M (2010). Sequential mining equips e-health with knowledge for managing diabetes. In New Trends in Information Science and Service Science (NISS), 2010 4th International Conference on IEEE. pp. 65-71.

Bellaachia A, Guven E (2006). Predicting breast cancer survivability: a comparison of three data mining method; Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining (SDM 2006).

Delen D, Walker G, Kadam A (2004). Predicting breast cancer survivability:a comparison of three data mining methods; Artificial Intelligence in Medicine 34(2):113-127.

Elder JF, Abbot DW (1998). A Comparison of Leading Data Mining

Tools; Fourth International Conference on Knowledge Discovery and Data Mining, New York.
Gandhi RK, Karnan M, Kannan S (2010). “Classification rule construction using particle swarm optimization algorithm for breast cancer datasets,” Signal Acquisition and Processing. ICSAP, International Conference pp. 233-237.
Hand DJ, Mannila H, Smyth P (2001). Principles of data mining, MIT Press, Boston, MA.,USA.
Kahramanli H, Allahverdi N (2009). Mining Classification rules for liver disorders. International Journal of Mathematics and Computers in Simulation 3(1):9-19.
Lin M, Huang S, Chang Y (2004). Kernel-based discriminant technique for educational placement; Journal of Educational and Behavioral Statistics 29:219-240.
Manaswini P, Ranjit KS (2011). “Predict the onset of diabetes disease using Artificial Neural Network (ANN)”, International Journal of Computer Science and Emerging Technologies (E-ISSN: 2044-6004) 2(2):303-311.
Michie D, Spiegelhalter DJ (1994). Machine Learning, Neural and Statistical Classification; Taylor CC; Prentice Hall.