INTERNATIONAL ENGINEERING, SCIENCE AND EDUCATION GROUP

Middle East Journal of Science
(2018) 4(2): 104 - 112
Published online December 26, 2018 (http://dergipark.gov.tr/mejs)
doi: 10.23884/mejs.2018.4.2.06
e-ISSN 2618-6136
Received: December 14, 2018 Accepted: December 24, 2018
Submission Type: Research Article

PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHMS OF SEVERAL DATA MINING SOFTWARES

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

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

[*] Corresponding author; baykal.abdullah@gmail.com

Abstract: *Data mining is to find correlations and rules which ensure meaningful and potentially useful estimations to be carried out for future among vast amount of existing data through computer programs. Today, many commercial or open source software tools are used regarding this matter. In this study, Classfication Analysis comparisons were carried out over the car evaluation data set consisting of 1728 registers especially on Weka, Orange, KNIME and Tanagra as open source software tools.*

Keywords: Data Mining,Classfication Analysis, Open Source Data Mining Tools

1. Introduction

It is estimated that the total size of the data produced by humanity in 2020 will reach 44 zettabytes (44 trillion GB)[3]. As predicted that the amount of information doubles every 20 months, opportunities of data gathering and storing this gathered data are increasing. Today, even the most basic actions such as using credit cards, medical test results, telephone conversations, purchasing products on supermarket at a time are being registered on computer environment. Businesses and goverment agencies are having more investment on data base system and storing more data on this system day by day. To take advantage of this large format data, it is needed to discovering valuable information by applying methods and rules over these data[4].

Data mining is discovering connections and regulations which are meaningful in a great available data, potentially useful and ensuring predictions about future by using computer programs. One of the data mining application areas which are becoming widespread in many sectors is the market basket analysis in which connections and rules are obtained by benefiting from customer, product and sale informations at supermarkets. Obtaining sale connections of products on market basket analysis and establishing association rules which is one of data mining matters are the profit

growning factors of companies. Association rules supply producing prudential predictions by discovering objects which act together inside sale action data and correlations between objects. Since the beginning of 90s, many algorithms have been developed to obtain these rules. It is available that these algorithms have superiorities on each other in different conditions and they have different working methods. Data base searching, applying defragmentation and pruning methods and discovering association connections between objects with minimum support value help constitute formal logic of algorithms[5].

To study on Data Mining, it is necessary to use programs developed for this matter. Lots of commercial and open source programs[1] are developed in this context. The main commercial programs are SAP Kxen, SPSS Clementine, SAS, Angoss, SQL Server, MATLAB, the top five of open source software are Orange, RapidMiner[6], Weka ,JHepWork, KNIME , Tanagra [7].

2. Data Mining Processes

Data Mining Processing consists of many stages. The main Data Mining Processing stages are below. These are;

1) Understanding problem area
2) Data selection
3) Preprocessing and data cleaning
4) Model Setup
5) Interpretation and Validation of the model

2.1. Understanmding Problem Area

The stage of understanding problem area requires gathering knowledge about the problem besides defining the problem and the objective of the study. Use of data mining techniuques without a proper understanding of the problem domain and sufficient knowledge mostly results with discovering irrelevant or meaningless information.

2.2. Data Selection

The data selection step requires user to aim at a data base and to select attributes and data for model creation. Having understood the problem area helps to select beneficial data on this step. Sometimes sufficient data is not available on a company structure. In this circumstance, data is obtained from external source.

2.3. Data Cleaning and Preprocessing

This step is the most time consuming step of all data mining processing. Raw data is usually neither neat nor suitable for data mining. The followings are the situations to be taken into consideration during the preprocessing and data cleaning step in order to prepare the data for further processing:

2.3.1 Data Cleaning

- Repetition: This kind of data conflict occurs when a sample exists several times in the data. This is the most common data conflict issue seen on the databases of companies like the credit card firms that personally deals with customers.
- Deficient Data Fields: There can be deficient fields on a data base for a variety of reasons. For example the customer filling a registry form may be bored with filling out required information or there can be deficient value because of inappropriate data entry in the field.
- Outliers: Outlier value in a field is the value that varies from the other values in the same field. As an example of this, think about monthly energy consumptions of customers in a public organization. If the values in this field are typically in between 0-1000KW and if we have 10,000KW entry for some customers, then these extreme values are described as outliers.

2.3.2 Preprocessing

On this step data selection is made depending on model to be set up. For example; for an estimator model, this step has the meaning for variable selection which will be used on model for dependent and independent variables selection.

Meaningless variables such as sequence number, ID number shouldn't go into model. Because these kind of variables can cause reduction in contribution of other variables on model and extending time to reach data. Some data mining softwares automatically eliminate these kind of irrelevant variables, however, it will be more rational not to leave it to the software in practice.

For example; demonstration of actual birthdate of each customer in practice can have adverse effect. Instead, seperating and grouping customers into different ages can be better.

2.4. Model Setup

We can say that model setup is the centre for all data mining application. It is the stage in which secret patterns and tendencies in this data come to light. There are lots of approaches about model setup stage. These are assembly, classification, clustering, alignment analysis and monitoring. Each approach can be put into practice by using one of the competitive methods such as statistical data analysis, machine learning and neurotic operations. The reason why data mining is mostly thought as an interdisciplinary area is that a large variety of methods from different disciplines are being used.

2.5. Interpretation and Validation

Interpretation and validation step of data mining is used for evaluating qualifications and values of the resulting model to determine whether turning back on previous steps by user is necessary or not. It is very important to understand the problem to appraise the result of the resulting model in this step.

3. Data Set Example

As specified on Table-1 below, Vehicle evaluation data set[2] consists of 1728 recordings and areas such as purchasing car, maintenance cost, number of doors, passenger capacity, luggage width and vehicle safety. In the classification algorithms used in this study, 66% of the samples in the set were used for learning and 34% were used for testing of the models.

Vehicle Evaluation Data Set		Area:7	Recording:1728
Area	**Type**	**Content**	**Information**
Buying	Discrete	vhigh,high,med,low	Purchase Price
Maint	Discrete	vhigh,high,med,low	Maintenance Price
Doors	Discrete	2,3,4,5more	Number of doors
Persons	Discrete	2, 4, more	Carrying Capacity
Luggage	Discrete	Small, med, big	Luggage Capacity
Safety	Discrete	Low, med, high	Safety Level
Class	Discrete	unacc, acc, good, vgood	Vehicle acceptance status

Table 1. Vehicle Evaluation Data Design Table

4. Processes and Data Mining Softwares Used

4.1. Weka

Weka software which has been developed open source on java by Waikato University and is still developing owns data classification, clustering, association and monitoring features. The name of Weka, is a software consisting of first letters of "Waikato Environment for Knowledge Analysis". Weka has totally a modular design and can make operations like visualization on data clustering, data analysis, business mind applications, data mining. Weka software comes idiocratically with support of .arff extention. But in Weka software, there are also vehicles for conversion CSV files into ARFF format. [8] It is not possible to process data on any text file with Weka, Arff, Csv, C4.5 formatted files. Also by using Jdbc and connecting data base, operations can be taken here.

Basically 3 Data Mining operation can be taken with Weka:

- (Classification)
- (Clustering)
- (Association)

Also, in addition to operations above, pre-processing and after-processing can be taken on data clusters.

- (Data Pre-Processing)
- (Visualization)

Lastly, there are plenty of built-in function working on files including data clusters on Weka Library. For Weka application, the data set given in Table 1 was converted to arff format as given below and the classification algorithms on this arff data were tested as given in Figure 1. Test results are given in Table 2

ARFF formatted vehicle evaluation data set recording design:

@RELATION CarTable
@ATTRIBUTE Buying {vhigh,high,med,low}
@ATTRIBUTE Maint {vhigh,high,med,low}
@ATTRIBUTE Doors {2,3,4,5more}
@ATTRIBUTE Persons {2,4,more}
@ATTRIBUTE Luggage {small,med,big}
@ATTRIBUTE Safety {low,med,high}
@ATTRIBUTE Class {unacc,acc,good,vgood}

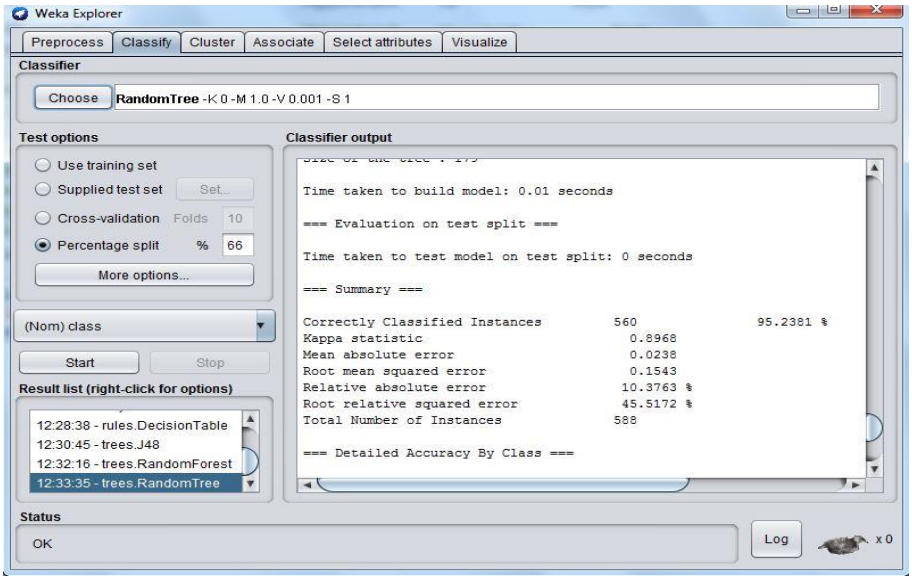

Figure 1. Result image in Weka software

Algorithm	Accuracy Rate
Tree J48	90.98
Decition Table	86.90
Lazy IBk	90.64
NaiveBayes	87.58
Random Forest	92.51
Lazy Kstar	86.39
Random Tree	84.01

Table 2. Weka application algorithm results

4.2 Orange

ORANGE software is developed by artificial intelligence research team on Slovenia Ljubljana University Computer and Information Science department.Orange is a data mining and machine learning application which is written by using C++ and Python and which uses Qt framework cross-platform for grafical interface. It consists of wide ranging component set such as user-friendly strong and flexible, data pre-processing, scoring feature and filtering, modelling, model evaluation and

discovery techniques. Orange can read data on *.tab, *.txt, *.basket, *.names, *.csv, *.tsv, *.arff, *.xml, *.svm file types . As shown in Figure 2 in Orange application, the results given in Table 3 were found.

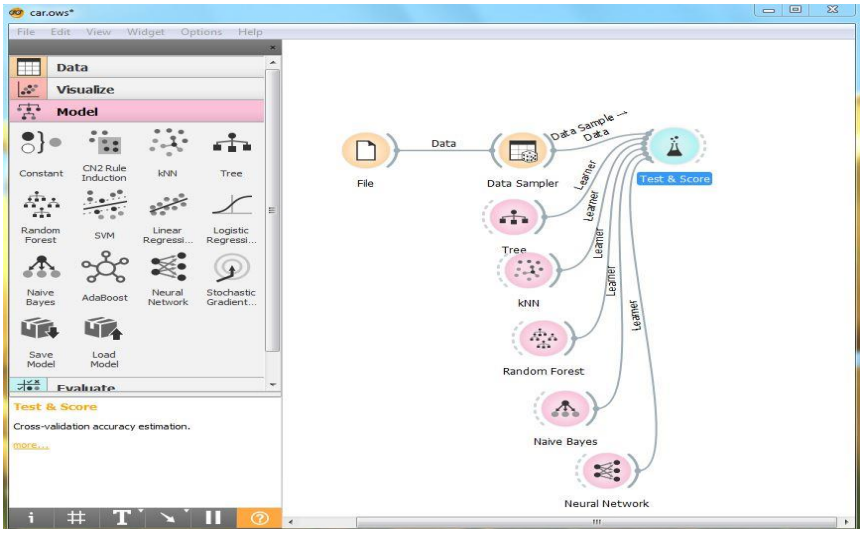

Figure 2. Result image in Orange software

Algorithm	Accuracy Rate
k Nearest Neighbor	87.70
Tree	95.00
Random Forest	91.71
Neural Network	97.12
Naive Bayes	85.75
AdaBoost	94.10
Costant	70.15

Table 3. Orange application algorithm results

4.3. Knime

It is developed by Konstanz University visual data mining research team on Eclipse Rich Client Platform.[9] Almost all data mining methods used frequently are available on this software. Among them, methods like support vector machines, Bayes and Multi dimensional Scaling (MDS) are also available. Knime can read data from files of *.txt., *.csv ,*.arff and also supports data read operation based on XML named PMML (Predictive Model Markup Language) which submits an opportunity to transfer data between data mining and statistical applications and access data with SQL querries from database servers. KNIME also has Data Write component which is useful for writing the data read on a different format which doesn't exist in the similar programs. As shown in Figure 3 in Knime application, the results given in Table 4 were found.

Algorithm	Accuracy Rate
Naive Bayes	87.76
Decision Tree	92.90
Gradient Boosted Tree	84.70
Random Forest	95.90
Tree Ensemble	95.60
Logistic Regression	92.50
k Nearest Neighbor	95.43

Table 4. Knime application algorithm results

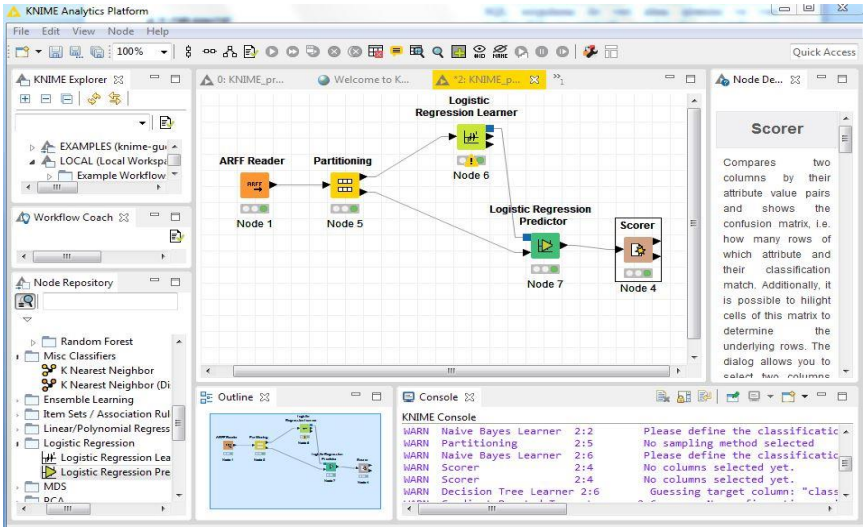

Figure 3. Result image in Knime software

4.4. Tanagra

Tanagra provides lots of data mining methods like Data analysis, statistical and machine learning. Tanagra consists of controlled learnings such as clustering, factorial analysis, parametric and nonparametric statistic, association rule, feature selection and structure algorithm and also other paradigms. [10]. Tanagra can read data from *.txt, *.ls, *.arff ve *.dat extention files. In the Tanagra application, the values given in Table 5 were found in the classification tests shown in Figure 4.

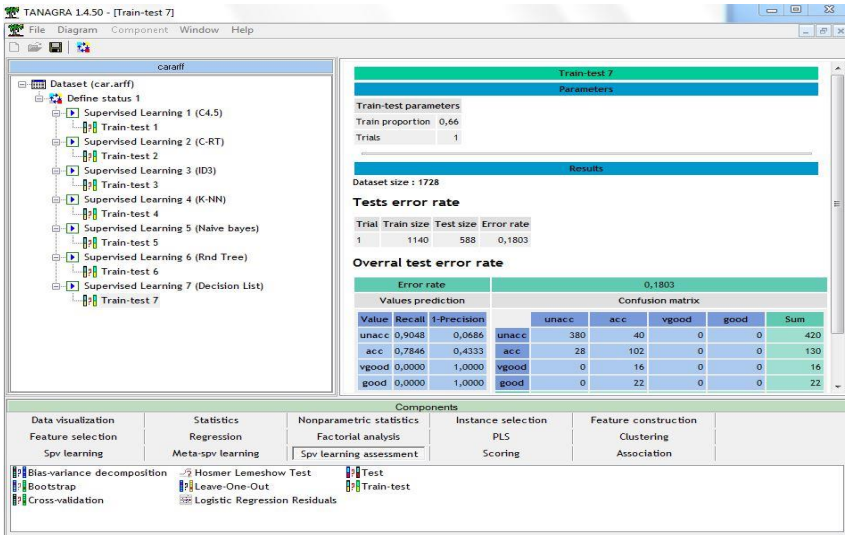

Figure 4. Result image in Tanagra software

Algorithm	Accuracy Rate
ID3	74.49
k Nearest Neighbor	81.80
C-RT	88.44
C4.5	86.90
Naive bayes	84.86
Random Tree	81.12
Decision List	81.97

Table 5. Tanagra application algorithm results

5. Result

The best accuracy rates of the classification algorithms achieved with the tests on WEKA, Knime, Orange and Tanagra using Vehicle Evaluation Data Set is shown in Chart 1.

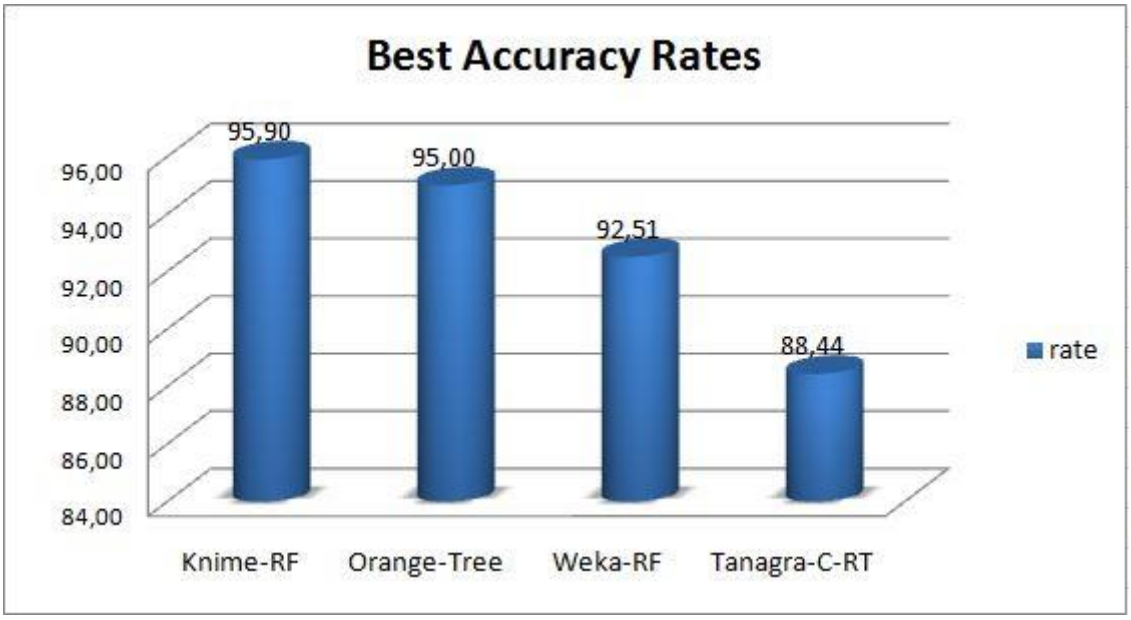

Chart 1. Best accuracy rates in the study

As can be seen from Chart 1 the best accuracy rate achieved was 95.90% with that of KNIME software's Random Forest Algorithm which was then followed by Orange's Tree algorithm, WEKA's Random Forest Algorithm and Tanagra's C-RT algorithm. It can be seen that when the algorithms' results were compared irrespective of the softwares used, the best result for the accuracy rate is achieved with Random Forest algorithm. All the softwares used in this study except Tanagra yielded results with accuracy rates above 90%. However, all the algorithms run on Tanagra software yielded results with accuracy below 90%. Comparing the results achieved by the softwares, we can say that the best results achieved were produced by KNIME software and the worst results were produced by

the Tanagra software.

6.References

[1] K. Fogel, *Producing Open Source Software: How to Run a Successful Free Software Project*, O'Reilly Media Inc., Boston, 2005

[2] [Last cited on 2018 Sep 10], Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/

[3] [Last cited on 2018 Sep 20] , https://www.emc.com/leadership/digital-universe /2014iview/executive-summary.htm, Dell EMC World Conference,Las Vegas, 2014

[4] Goebel, M., Gruenwald, L., "A survey of data mining and knowledge discovery software tools", *ACM SIGKDD Explorations Newsletter*, 1, 20-33, 1999

[5] J. Alcala-Fdez, A. Fernandez, J. Luengo, J. Derrac, and S. Garc , "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework", *Multiple-Valued Logic and Soft Computing*, 17, 255-287, 2011

[6] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler**, "**YALE: rapid prototyping for complex data mining tasks**",** *KDD '06 Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ,New York, USA* , 2006, pp. 935–940

[7] [Last cited on 2018 Oct 29] , TechSource, http://www.junauza.com/2010/11/free-data-mining-software.html

[8] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann P., Witten, I., H.," The WEKA data mining software: an update", *ACM SIGKDD Explorations Newsletter*, 11 , 37-57, 2009

[9] [Last cited on 2018 Sep 25] , Knime Tutorial, http://didawiki.cli.di.unipi.it/lib/exe/fetch.php/dm/knime_slides_mains.pdf

[10] [Last cited on 2018 Oct 29],Tanagra, http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html