Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 85 (2016) 662 – 668

Procedia Computer Science

International Conference on Computational Modeling and Security (CMS 2016)

Correlation review of classification algorithm using data mining tool: WEKA , Rapidminer , Tanagra ,Orange and Knime

Amrita Naik[a]* , Lilavati Samant[b]

[a]Assistant Professor ,Computer Engg. Dept.,Don Bosco College of Engineering,Margao-Goa,403604.India
[b]Assistant Professor ,Computer Engg. Dept.,Don Bosco College of Engineering,Margao-Goa,403604.India

Abstract

This paper conducts a correlation review of classification algorithm using some free available data mining and knowledge discovery tools such as WEKA, Rapid miner, Tanagra, Orange and Knime. The accuracy of classification algorithm like Decision tree, Decision Stump, K-Nearest Neighbor and Naïve Bayes algorithm have been compared using all five tools. Indian Liver Patient DataSet is used for testing the Classification algorithm in order to classify the people with and without Liver disorder.

Peer-review under responsibility of the Organizing Committee of CMS 2016

Keywords: Classification; Data Ming

1. Introduction

Since the data is tremendously increasing, it becomes difficult for an individual to manually analyze the data for strategic decision making. Hence humans need help of data mining to mine interesting information from the available data. Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories .One of the important problem in data mining is the Classification which involves finding rules that partition given data into predefined classes. In the data mining

* Corresponding author: 7798677144
*E-mail address:*amritametri@gmail.com

Peer-review under responsibility of the Organizing Committee of CMS 2016
doi:10.1016/j.procs.2016.05.251

domain where trillions of data is used, the execution time of existing algorithms can become time consuming. Hence there is a need for automated tools that can assist us in transforming those huge data into Information.

Now-a-days, many open-source data mining tools and software are available for use such as the Rapidminer [1], Waikato Environment for Knowledge Analysis (WEKA) [2],KNIME, R-Programming, Orange, NLTK etc. These tools and software provide a set of methods and algorithms that help in better analysis of data. These tools help in cluster analysis, data visualization, regression analysis, Decision trees, Predictive analytics, Text mining, etc.

We have conducted a comparison study between classification algorithm such as Decision tree, Decision Stump, K-Nearest Neighbor and NaiveBayes algorithm using WEKA, Rapidminer, Tanagra, Orange and Knime tool.. The accuracy measure; which represents the percentage of correctly classified instances, is used for judging the performance of the classification algorithm

2. Classification Algorithms used in the Experiment

Data Classification Algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observations. Data Classification process includes two steps:
i) Building the Classifier Model: Here the classifier is built by learning the training set and their associated class labels.
ii) Using Classifier for Classification: In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules.

We have studied the following Classification Algorithm in our paper:

- Decision Tree
- Naïve Bayes
- K-Nearest Neighbor

Decision Tree: A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.
Decision Stump: A decision stump is a machine learning model consisting of a one-level decision tree.
K-Nearest Neighbor: K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).
Naive Bayes: Is a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features.

3. Tools Description

3.1 RapidMiner : is an user interactive environment for machine learning and data mining processes. It is open-source, free project implemented in Java. It represents a modular approach to design even very complex problems - a modular operator concept which allows the design of complex nested operator chains for a huge number of learning problems. RM uses XML to describe the operator trees modeling knowledge discovery (KD) processes. RM has flexible operators for data input and output in different file formats. It contains more than 100 learning schemes for classification, regression and clustering tasks.

3.2 WEKA : is a widely used toolkit for machine learning and data mining that was originally developed at the University of Waikato in New Zealand. It contains a large collection of state-of-the-art machine learning and data mining algorithms written in Java. WEKA contains tools for regression, classification, clustering, association rules, visualization, and data pre-processing. WEKA has become very popular with the academic and industrial researchers, and is also widely used for teaching purposes.

3.3 Tanagra **:** is a free suite of machine learning software for research and academic purposes developed by Ricco Rakotomalala at the Lumière University Lyon 2, France. Tanagra supports several standard data mining tasks such as: Visualization, Descriptive statistics, Instance selection, feature selection, feature construction, regression, factor analysis, clustering, classification and association rule learning. Tanagra makes a good compromise between the statistical approaches (e.g. parametric and nonparametric statistical tests), the multivariate analysis methods (e.g. factor analysis, correspondence analysis, cluster analysis, regression) and the machine learning techniques (e.g. neural network, support vector machine, decision trees, random forest).

3.4 Orange: is an open source machine learning and data mining software (written in Python). It has a visual programming front-end for explorative data analysis and visualization, and can also be used as a Python library. The program is maintained and developed by the Bioinformatics Laboratory of the Faculty of Computer and Information Science at University of Ljubljana.Orange is a component-based visual programming software for data mining, machine learning and data analysis.Components are called widgets and they range from simple data visualization, subset selection and preprocessing, to empirical evaluation of learning algorithms and predictive modelling.

3.5 Knime : is a widely used open source data mining, visualisation and reporting graphical workbench used by over 3000 organisations. Knime desktop is the entry open source version of Knime .It is based on the well regarded and widely used Eclipse IDE platform, making it as much a development platform (for bespoke extensions) as a data mining platform.

4. Experiment

*4.1 Dataset***:** We have downloaded Indian Liver Patient Dataset (ILPD) from the UCI repository [3]. The numbers of instances are 583. This data set contains 416 liver patient records and 167 non liver patient records. The data set was collected from north east of Andhra Pradesh, India. Liver Patient or Not is a class label used to divide into groups(liver patient or not). This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age "90". The table has the following attribute:

- Age: Age of the patient .
- Gender: Gender of the patient.
- TB: Total Bilirubin.
- DB: Direct Bilirubin
- Alkphos: Alkaline Phosphotase
- Sgpt: Alamine Aminotransferase
- Sgot: Aspartate Aminotransferase
- TP: Total Protiens
- ALB: Albumin
- A/G Ratio: Albumin and Globulin Ratio
- Liver Patient or Not field used to split the data into two sets.
  1 -Indicates Patient with Liver problem and 2- Indicates Patient with not a Liver problem .

4.2. Evaluation of Classification Algorithm using Rapid miner:

We evaluate the performance of the classification algorithm using Confusion Matrix, a table that reveals true versus predicted values. Table (1-4) depicts the confusion matrix for Decision tree, Naïve Bayes, Decision Stump and K-Nearest Neighbor.

4.2.1.Decision Tree:

The decision tree for the experimental data is as follows:

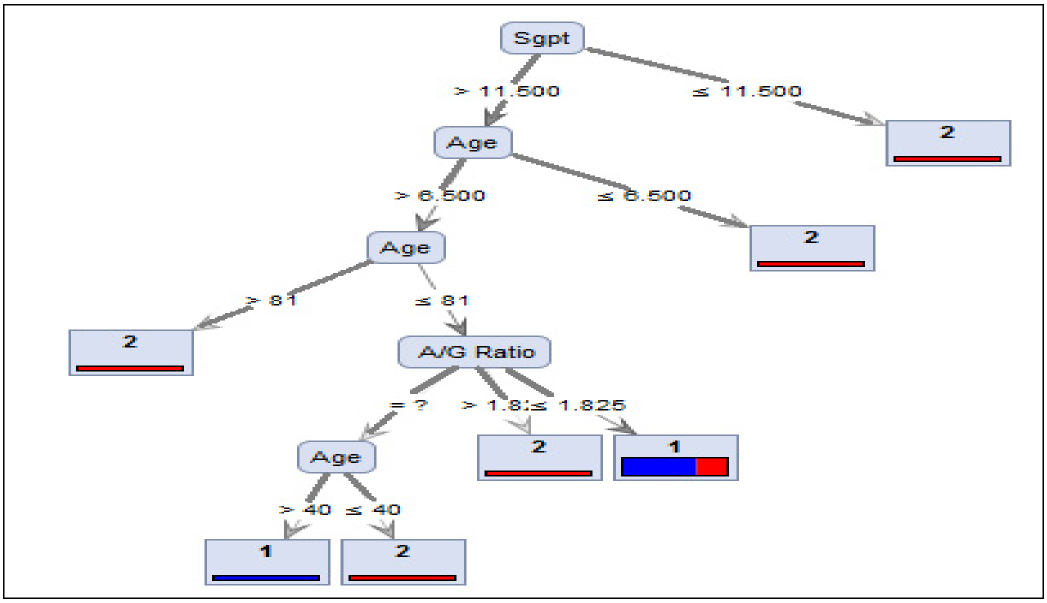

Fig 1: Decision Tree for classifying patients with Liver Disorder

Table 1: Performance of Decision Tree classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1(Liver Patient)	412	158	72.29%
Pred2 (Not a Liver Patient)	4	9	69.23%
Class Recall	99.04%	5.39%	

Table 2: Performance of Naïve Bayes classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	169	7	96.02%
Pred2 (Not a Liver Patient)	247	150	37.78%
Class Recall	40.62%	95.81%	

Table 3: Performance of K-NN classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	304	112	73.07%
Pred2 (Not a Liver patient)	4	9	69.23%
Class Recall	98.7%	7.32%	

4.3. Evaluation of Classification Algorithm using WEKA:

We evaluate the performance of the classification algorithm using Confusion Matrix, a table that reveals true versus predicted values.

Table 4: Performance of Decision Tree

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	388	28	93.2%
Pred2 (Not a Liver Patient)	46	121	72.45%
Class Recall	89.4%	81.21%	

Table 5: Performance of Naïve Bayes classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	165	251	39.66%
Pred2 (Not a Liver Patient)	7	140	95.23%
Class Recall	95.9%	35.8%	

Table 6:Performance of K-NN classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	416	0	100%
Pred2 (Not a Liver Patient)	2	165	98.8%
Classs Recall	99.5%	100%	

4.4. Evaluation of Classification Algorithm using Tanagra:

Table 7: Performance of Decision Tree classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	388	26	93.71%
Pred2 (Not a Liver Patient)	49	116	70.31%
Classs Recall	88.78%	81.69%	

Table 8: Performance of Naïve Bayes classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	335	79	80.91%
Pred2 (Not a Liver Patient)	95	70	42.42 %
Class Recall	77.9 %	46.97%	

Table 9: Performance of K-Nearest Neighbor classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	369	45	89.13%

Pred2 (Not a Liver Patient)	77	88	53.33 %
Class Recall	82.73%	66.16%	

4.5. Evaluation of Classification Algorithm using Orange:

Table 10: Performance of Decision Tree classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	323	93	77.64%
Pred2 (Not a Liver Patient)	105	62	37.16%
Classs Recall	75.46%	40.30%	

Table 11: Performance of Naive Bayes classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	279	137	67.06%
Pred2 (Not a Liver Patient)	53	114	68.26%
Classs Recall	84.03%	45.41 %	

Table 12: Performance of K-Nearest Neighbor classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	328	88	78.84%
Pred2 (Not a Liver Patient)	115	52	31.13%
Classs Recall	74.0 %	37.14 %	

4.6 Evalaution of Classification Algorithm using knime:

Table 13: Performance of Decision Tree classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	404	12	97.11%
Pred2 (Not a Liver Patient)	15	152	91.10%
	96.46%	92.68%	

Table 14: Performance of Naive Bayes classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	393	23	94.47%
Pred2 (Not a Liver Patient)	137	30	17.96%
Classs Recall	74.15%	56.60 %	

Table 15: Performance of K-Nearest Neighbor classification

	True 1 (Liver Patient)	True 2 (Not a Liver Patient)	Class Precision
Pred1 (Liver Patient)	377	37	91.06%
Pred2 (Not a Liver Patient)	58	107	64.84%
Classs Recall	86.6 %	74.30 %	

5 Conclusions:

In this paper we used liver patient data sets from ILPD (Indian Liver Patient) Data Set. It has 583 samples with 10 independent variables and one class variable. The performance of the Classification models on the basis of Accuracy was compared, which defined as follows in Table 16.
Accuracy = (TP+TN) / (TP+FP+TN+FN)

Table 16: Accuracy measure of Classification Algorithm

Algorithm	**Rapid miner**	**WEKA**	**Tanagra**	**Orange**	**Knime**
1.Decision Tree	72.21%	87.76%	87.05%	66.04%	95.37%
2.Naive Bayes	56.67%	54.17%	69.95%	67.41%	72.56%
3.K-Nearest Neighbor	72.96%	99.66%	78.93%	65.18%	86.58%

From above table it is clear that WEKA tool estimates a lowest accuracy for Naive Bayes , however for the same algorithm Knime tool estimates better accuracy when compared to WEKA. In case of Decision tree evaluation, Orange tool showed lower accuracy where as Knime tool estimated better accuracy when compared to prior. Overall KNIME tool estimates higher accuracy for all three classification algorithm.
From the above table, it is also clear that the accuracy of Decision Tree and K-Nearest Neighbour is better when compared to Naïve Bayes .

References

1. https://rapidminer.com/
2. http://www.cs.waikato.ac.nz/ml/weka/
3. http://eric.univ-lyon2.fr/~ricco/tanagra/
4. https://www.knime.org/knime
5. http://orange.biolab.si/
6. Christos Tjortjis, John Keane "A classification algorithm for data mining" Intelligent Data Engineering and Automated Learning — IDEAL 2002
7. Abdullah H Wahbeh,Qasem A. Al-Radaideh, Mohammed N Al-Kabi, Emad M Al Shawakfa, "Comparitive study of data mining tools over some classification methods", (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence.
8. Magdalena Graczyk , Tadeusz Lasota , Bogdan Trawiński1 , "Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA".
9. Bendi Venkata Ramana , Prof. M. Surendra Prasad Babu , Prof. N. B. Venkateswarlu, "A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 2, May 2012.
10. Magdalena Graczyk1 , Tadeusz Lasota2 , Bogdan Trawiński1 , "Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA".