 ISBN: 978-81-927230-0-6

Evaluation of Data Classifiers Using WEKA

Mrs. Vaishali Prashant Bhosale
Bharati Vidyapeeth Deemed University
Institute of Management, Kolhapur.
Vaishali.p.bhosale@gmail.com

Abstract — ***In this paper, I tried to evaluate classification algorithms in Data Mining. I analyzed important characteristics of the applications when executed in well known tool WEKA. My current work is focusing on evaluating the applications on different data sets to allow the retailers to increase customer understanding and make knowledge- driven decisions in order to provide personalized and efficient customer service. I tried to evaluate the performance of various classifiers on two test mode 10 fold cross validation and percentage split with different data sets at WEKA 3-6-10. The results after evaluation discussed here.***

Keywords— ***Data Mining, WEKA, Classification, UCI datasets.***

I. INTRODUCTION

It is generally know that different classification methods are suited to different types of data. There are various approaches to determine the performance of classifiers. The performance measure in this paper is taken to be classification accuracy. The WEKA (Waitko Environment for Knowledge Analysis) toolkit is used for the classification.

II. CLASSIFICATION OF DIFFERENT TYPES OF DATA

A. The datasets

The author tested several datasets from the UCI dataset that can be found on the Weka website [WEKA]. Detailed descriptions of the datasets are also to be found there. The reason for the choice of these datasets is that they contained the highest numbers of instances among the data sets listed at the Weka website and also come from a variety of fields which means that they can be considered to provide rather generic view of the data that data mining methods work on. The datasets where according to the Weka site ready for use and where therefore not pre-processed using e.g. filter or wrapper functions.

B. Use of the Weka packet

The primary focus was on utilising the Weka packet for the task of classification using different. The Weka packet provides three kinds of interfaces: an command line one, Explorer and the Experimenter. It was decided after some testing to use the Explorer for the evaluations. It should be added that the Weka packet algorithms are Open Source so that it is possible to change them and add new ones.

III. THE SETUP OF THE CLASSIFIERS

Tenfold cross validation was used for each model creation to get a reliable estimation of the models capabilities in classifying the dataset.

The classification methods used as described below:

(1) Lazy-IBK

(2) Lazy-K Star

(3) Function- Linear regression, Logistic

(4) Rules- Zero R

(5) Tree-REP

(6) Tree- Decision stump

IV. THE SETUP OF THE TEST ENVIRONMENT

The performance of different classifiers on each dataset was evaluated with the Explorer component of the Weka Environment. The data used in this experiment is either real world data obtained from UCI data repository and widely accepted dataset available in Weka toolkit, during evaluation multiple data sizes were used, each dataset is described by the data type being used, the types of attributes, the number of instances stored within the dataset, also the table demonstrates that all the selected data sets are used for the classification and clustering task. These datasets were chosen because they have different characteristics and have addressed different areas.

To evaluate the selected tool using the given datasets, several experiments are conducted. For evaluation purpose full training set used with the k-fold (k=10) cross-validation mode. It is common to choose k=10 or any other size depending mainly on the size of the original dataset.

V. THE RESULTS

The results of each classifier for each dataset are presented as its correlation coefficient, mean absolute error, root mean squared error, relative absolute error and root relative squared error and represented in table below.

ISBN: 978-81-927230-0-6

Table 1. cpu.with.vendor data results

Classifier	Correlatin coefficient	Mean absolute error	Root mean squared error	Relative absolute Error %	Root relative squared Error %
Lazy-IBK	0.9467	20.82	78.635	53.63	34.66%
Lazy-KStar	0.9593	12.53	44.675	14.29	28.87
Function-Linear regression	0.9257	36.97	58.452	42.17	37.77
Rules-ZeroR	-0.1442	87.66	154.76	100	100
Tree-REP	0.8803	28.46	74.492	32.46	48.13
Tree-Decision stump	0.6925	67.70	111.69	77.24	72.17

Table 2. supermarket data results

Classifier	Correctly Classified %	Mean absolute error	Root mean squared error	Relative absolute Error %	Root relative squared Error %
Lazy-IBK	82.46	0.128	0.3411	53.635	34.66%
Lazy-KStar	89.47	0.095	0.2742	20.718	57.432
Function-Logistic	92.98	0.064	0.2438	14.007	51.065
Rules-ZeroR	64.91	0.457	0.4775	100	100
Tree-REP	78.94	0.289	0.4433	63.206	92.852
Tree-Decision stump	80.70	0.210	0.3358	45.959	70.335

Table 3. iris data results

Classifier	Correctly Classified %	Mean absolute error	Root mean squared error	Relative absolute Error %	Root relative squared Error %
Lazy-IBK	95.33	0.04	0.1747	9.343	37.07%
Lazy-KStar	94.67	0.043	0.1555	9.658	32.982
Function-Logistic	96	0.029	0.1424	6.456	30.214
Rules-ZeroR	33.33	0.444	0.4714	100	100
Tree-REP	94	0.056	0.1936	12.675	41.059 9
Tree-Decision stump	66.67	0.222	0.3333	50	70.711

VI. CONCLUSION AND FURTHER WORK

Judging from the result it can be concluded that further tests are required to get a reliable estimation of the true performance of classifier models when given different types of datasets. Some trends are though visible which further research could test for validation. The same work can be extended for clustering methods for data mining on different datasets.

VII. REFERENCES

1. [WEKA]: www.cs.waitko.ac.nz/ml/weka
2. [WEKAB]: Witten, Ian, H. and Eibe, Frank 2000: Data Mining, Practical machine learning tools and techniques with java implementations. Morgan Kaufmann, San Diego, CA.
3. Jiawei Han, Micheline Kamber, Jian PeiData Mining, Second Edition: Concepts and Techniques