2A1-IOS-3b-5

Comparative Study & Performance Evaluation of Various Classifiers using a data set

Amresh Kumar

Department of Computer Science and Engineering
Christ University, Bangalore, INDIA

ABSTRACT

Real-world knowledge discovery processes typically consist of complex data pre-processing, machine learning, evaluation, and visualization steps. Hence a data mining platform should allow complex nested operator chains or trees, provide transparent data handling, comfortable parameter handling and optimization, be flexible, extendable and easy-to-use. Modern machine learning techniques have encouraged interest in the development of various systems that ensure secure, reliable and many more operations in the different fields and applications. In an earlier study, many other approaches/methods were investigated to develop various applications using modern machine learning techniques and more specific classification algorithms.

The Weka machine learning workbench provides a general-purpose environment for automatic classification, clustering and feature selection, and common data mining problems in bioinformatics research.

Here in this Project Report paper we have used various classifiers with filters to perform classification and we have done analysis of data with different classifiers and then we have done feature selection process and during all these activities we have observed and record the various performance change and different graphs which are briefed inside this paper.

Keywords: Machine Learning, WEKA, Data mining, KDD, Classification, Filters, Feature Selection

I. INTRODUCTION

In an earlier study, many other approaches/methods were investigated to develop various applications using modern machine learning techniques and more specific classification algorithms. Modern machine learning techniques have encouraged interest in the development of various systems that ensure secure, reliable and many more operations in the different fields and applications.
The Weka machine learning workbench provides a general-purpose environment for automatic classification, clustering and feature selection, and common data mining problems in bioinformatics research. Therefore Weka also contains an extensive data pre-processing methods and the experimental comparison of different machine learning techniques on the same problem.

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), is a field at the intersection of computer science and statistics is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

The term ***Knowledge Discovery in Databases*** or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.
In feature selection operation we are going to find out which are the most important instances to carry out the classification to get accurate result by improving their performance.

Here in this Project Report paper we have used various classifiers with filters to perform classification and we have done analysis of data with different classifiers and then we have done feature selection process and during all these activities we have observed and record the various performance change and different graphs which are briefed inside this paper.

Contact: amreshkr.19@gmail.com

To improve the performance we have experimented with Artificial Intelligence based (AI Classifier), a rule-based learning method using statistical analysis and also Decision tree and Support Vector Machine (SVM) based classification schemes were used to analysis and inspection of data. This selected algorithm efficiency and overall performance for the given data set (Temp.csv) are observed and calculated. This experiment/study has been conducted using six classifiers, namely SMO, REPTree, IBK, Logistic and Multilayer perceptron, with Temp.csv datasets having 41 instances. The Waikato Environment for Knowledge Analysis (WEKA) learning tool has been used in this Experiment/study.

II. METHODOLOGY

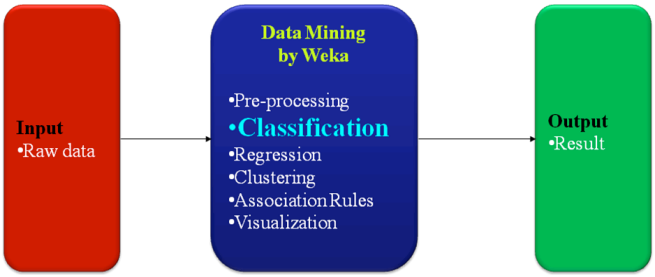

Fig. 1. Flow of Methodology involved.

Using Weka tool we have executed six various classifiers algorithm on our dataset and compared the various classifiers based on the ROC Area (Weighted Average) value.

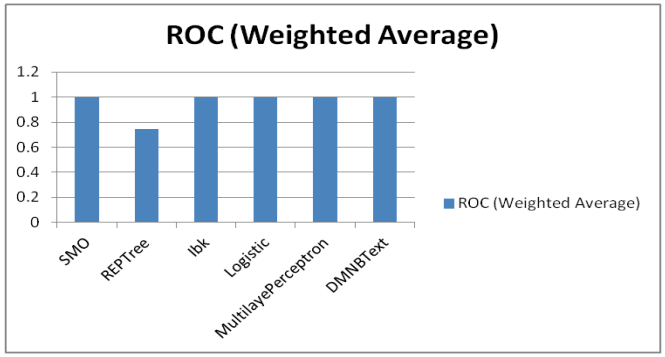

Fig. 2. ROC (Weighted Average)

And also we found that out 6 classifiers, 4 Classifiers are showing 100% correctly classified instances and 2 Classifiers REPTree and DMNBText are showing incorrectly classified instances, therefore next step we proceed with finding out which instances was not correctly classified. For this we have to do a tuple wise analysis, so for the same I am taking only one classifier now i.e. DMNBtext Classifier.

III. RESULTS

Table 1: Consolidated Classifiers Sheet using Training Set as Test Options

	DATA CLASSIFICATION STATISTICS					
	Data File Description					
	Relation=temp	Attributes=62		Instances=41		
	CLASSIFIER 1: AI Classifier SMO	CLASSIFIER 2: Tree Classifier REPTree	CLASSIFIER 3: Ibk	CLASSIFIER 4: Logistic	CLASSIFIER 5: MultiLayer Perceptron	CLASSIFIER 6: DMNBText
Summary						
Number of Correctly Classified Instances	41	28	41	41	41	35
Correctly Classified Instances %	100	68.2927	100	100	100	85.3659
Number of Incorrectly Classified Instances	0	13	0	0	0	6
Incorrectly Classified Instances %	0	31.7073	0	0	0	14.6341
Kappa statistic	1	0.4336	1	1	1	0.7681
Mean absolute error	0.2222	0.2969	0.0303	0	0.0059	0.2135
Root mean squared error	0.2722	0.3853	0.0321	0	0.0091	0.2783
Relative absolute error	55.9901	74.8118	7.635	0.0001	1.4937	53.7995
Root relative squared error	61.3467	86.8492	7.2447	0.0001	2.0548	62.7209
Total Number of Instances	41	41	41	41	41	41
Detailed Accuracy By Class						
TP Rate (weighted avg)	1	0.683	1	1	1	0.854
FP Rate (weighted avg)	0	0.266	0	0	0	0.025
Precision (weighted avg)	1	0.585	1	1	1	0.927
Recall (weighted avg)	1	0.683	1	1	1	0.854
F-measure (weighted avg)	1	0.627	1	1	1	0.871
ROC Area (weighted avg)	1	0.746	1	1	1	0.997
Confusion Matrix						
a=0,b=L,c=N						
a-a	22	16	22	22	22	19
a-b	0	6	0	0	0	0
a-c	0	0	0	0	0	3
b-a	0	1	0	0	0	0
b-b	13	12	13	13	13	10
b-c	0	0	0	0	0	3
c-a	6	6	0	0	0	0
c-b	0	0	0	0	0	0
c-c	6	0	6	6	6	6

Table 2: Classifiers Comparison Chart

Sl. No.	Classifiers	ROC (Weighted Average)
1	SMO	1
2	REPTree	0.746
3	Ibk	1
4	Logistic	1
5	MultilayePerceptron	1
6	DMNBText	0.997

Table 3: Analysis and performance change for DMNBText Classifiers using cross-validation as Test Options

From the Figure given below we can see the ROC value for DMNBText Classifier without and With the Attributor Evaluator. Thus by Calculating ROC change we can Find out the performance Change as given below.

Performance Change= ($ROC_{With\ Evaluator}$ - $ROC_{Without\ Evaluator}$) / $ROC_{Without\ Evaluator}$

CLASSIFIER NAME: DMNBText		
	Without Attribute Selection	*With Attribute Selection: Attribute Evaluator: name SearchMethod: Ranker*
List of Selected Attributes---------->>	*All*	*RL-54, L-18, L-14, L-20, L-24,, L-16, L-15, H-35, RL-60, RL-62, P-49, H-32, L-17*
Summary		
Number of Correctly Classified Instances	33	31
Correctly Classified Instances %	80.487	75.6098
Number of Incorrectly Classified Instances	8	10
Incorrectly Classified Instances %	19.512	24.3902
Kappa statistic	0.6439	0.5514
Mean absolute error	0.2455	0.2827
Root mean squared error	0.3339	0.3463
Relative absolute error	61.4542	70.7685
Root relative squared error	74.8162	77.6059
Total Number of Instances	41	41
Detailed Accuracy By Class		
TP Rate (weighted avg)	0.805	0.756
FP Rate (weighted avg)	0.175	0.215
Precision (weighted avg)	0.829	0.645
Recall (weighted avg)	0.805	0.756
F-measure (weighted avg)	0.77	0.696
ROC Area (weighted avg)	0.897	0.909
Confusion Matrix		
a=O,b=L,c=N		
a-a	21	20
a-b	1	2
a-c	0	0
b-a	2	2
b-b	11	11
b-c	0	0
c-a	3	4
c-b	2	2
c-c	1	0
Performance change(%)		1.34

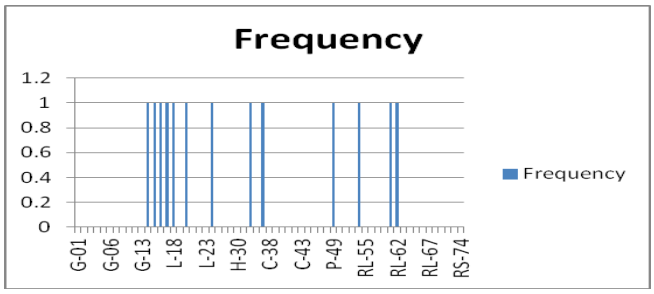

Fig. 3. Frequency

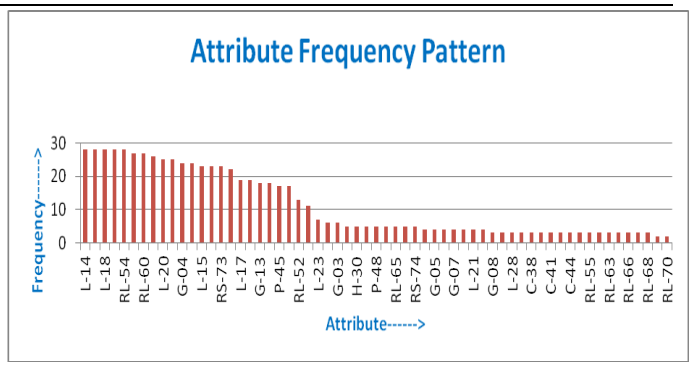

Fig. 4. Attribute Frequency Pattern

Table 4: MultilLayer Perceptron (Using Training Set)

Classifier Name: MultiLayerPerceptron (Using Traing Set)		
Summary	With all attributes	With only selected 35 attributes
Number of Correctly Classified Instances	41	41
Correctly Classified Instances %	100	100
Number of Incorrectly Classified Instances	0	0
Incorrectly Classified Instances %	0	0
Kappa statistic	1	1
Mean absolute error	0.0059	0.0075
Root mean squared error	0.0091	0.0119
Relative absolute error	1.4937	1.8879
Root relative squared error	2.0548	2.6893
Total Number of Instances	41	41
Detailed Accuracy By Class		
TP Rate (weighted avg)	1	1
FP Rate (weighted avg)	0	0
Precision (weighted avg)	1	1
Recall (weighted avg)	1	1
F-measure (weighted avg)	1	1
ROC Area (weighted avg)	1	1
Confusion Matrix		
a=O,b=L,c=N		
a-a	22	22
a-b	0	0
a-c	0	0
b-a	0	0
b-b	13	13
b-c	0	0
c-a	0	0
c-b	0	0
c-c	6	6
Performance change	0	0

Table 4: MultilLayer Perceptron (Using Cross Validation)

Classifier Name: MultiLayerPerceptron (Cross-Validation)		
Summary	**With all attributes**	**With only selected 35 attributes**
Number of Correctly Classified Instances	30	32
Correctly Classified Instances %	73.1701	78.0488
Number of Incorrectly Classified Instances	11	9
Incorrectly Classified Instances %	26.8293	21.9512
Kappa statistic	0.5847	0.6506
Mean absolute error	0.1866	0.1636
Root mean squared error	0.3666	0.3474
Relative absolute error	46.7145	40.9598
Root relative squared error	82.148	77.8538
Total Number of Instances	41	41
Detailed Accuracy By Class		
TP Rate (weighted avg)	0.732	0.78
FP Rate (weighted avg)	0.101	0.093
Precision (weighted avg)	0.818	0.828
Recall (weighted avg)	0.732	0.78
F-measure (weighted avg)	0.754	0.795
ROC Area (weighted avg)	0.908	0.911
Confusion Matrix		
a=O,b=L,c=N		
a-a	15	16
a-b	1	1
a-c	6	5
b-a	1	0
b-b	10	12
b-c	2	1
c-a	1	2
c-b	0	0
c-c	5	4
Performance change	0	0.330396476

IV. Discussion

For the experimental analysis SMO, REPTree, IBK, Logistic and Multilayer perceptron classifiers are considered in this experiment/study. 41 instances in data sets were selected from the collected data. To cover this experimental I have taken Temp.csv datasets, and the observation and Calculation is done by considering following:

- Attribute selection
- Frequency
- ROC
- Confusion metrics

Table 5: Comparisons with graph without Feature Extraction and with Feature Extraction using DMNBText Classifier

Classifier: DMNBText	Type of Search Method	ROC Area(Wt. Avg.)
AttributeEvaluator: ChiSquaredAttributeEval	Without Feature Extraction	0.897
	Ranker	0.909

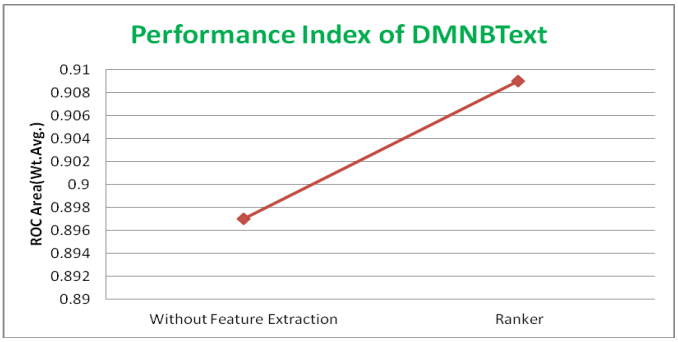

Fig. 5. Performance Index for DBMNText

V. Conclusion

As a conclusion we can tell that classification algorithms play a key role to solve real world problems. Selection of an application specific classifier is an emerging research area. In this paper, performance change is being evaluated and calculated using various popular classifiers. Initially, the percentage of correct classifications has been measured with the highest accuracy. Later, ranking performance has been estimated to select a suitable algorithm for this application. The ranking performance has shown that DMNBText performs the best for the given datasets. This also reduces computational complexity, and development and maintenance costs both in terms of hardware and human inspection.

Based on the results obtained in the various algorithms, we can conclude that the feature selection concept played an important role and can be useful component for many classifications. This is possible due to the low computational cost of this method, which is more efficient compared to the other ones. The main advantage of this method is that it makes no assumptions and these methods, not only improved the classification speed significantly, but they also improved the accuracy rate and the reliability in most of

the cases. Thus using the concept of Data Mining techniques we examine and calculate the performance using ROC Values.

VI. REFERENCES

[1] http://www.cs.waikato.ac.nz/~ml/weka/.
[2] http://ieeexplore.ieee.org/A Comparable Study employing WEKA Clustering/Classification Algorithms for Web Page Classification/Ioannis Charalampopoulos, Ioannis Anagnostopoulos/ 2011, Page(s): 235 – 239.
[3] http://ieeexplore.ieee.org/Rule-Based Classification Approach for Railway Wagon HealthMonitoring/G M Shafiullah, A B M Shawkat Ali, Adam Thompson, Peter J Wolfs/2010, Page(s): 1 – 7.
[4] http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf.
[5] http://www.ijcse.com/docs/IJCSE10-01-04-51.pdf
[6] http://en.wikipedia.org/wiki/Data_mining.
[7] Books: DATA MINING by Ian H. Witten & Eibe Frank, Second Edition.