International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com
Volume 2, Issue 10, October 2013 ISSN 2319 - 4847

“Evaluating Performance of Data Mining Classification Algorithm in Weka”

Nikhil N. Salvithal[1], Dr. R. B. Kulkarni[2]

Walchand Institute of Technology, Solapur, Maharashtra, India.

Abstract

Data mining is a technique that uses different types of algorithms to find relationships and trends in large datasets to promote decision support. The data sizes accumulated from various fields are exponentially increasing; data mining techniques that extract information from large amount of data have become popular in commercial and scientific domains, including sales, marketing, customer relationship management, quality management. The aim of this paper is to judge the performance of different data mining classification algorithms on various datasets. The performance analysis depends on many factors test mode, different nature of data sets, type of class and size of data set.

Keywords-Data Mining, Classification, Performance, WEKA

1. Introduction:

As the data sizes accumulated from various fields are exponentially increasing, data mining techniques that extract information from large amount of data have become popular in commercial and scientific domains, including marketing, customer relationship management, quality management.

We studied various articles regarding performance evaluation of Data Mining algorithms on various tools, some of them are described here, Abdullah [3] compared various classifiers with different data mining tools & found WEKA as best tool, Mahendra Tiwari & Yashpal Singh [1] evaluated performance of 4 clustering algorithms on different datasets in WEKA with 2 test modes. Some people worked on use of classification algorithms in WEKA for datasets from specific areas such as Tanuja S, Dr. U. Dinesh Acharya, and Shailesh K R [9] compared different data mining classification techniques to predict length of stay for an inpatient in hospital**.**

Generally arff datasets have 2 types of attributes nominal & numeric. There is need to find suitable classifiers for datasets with different type of class (either nominal or numeric), so we focused on evaluating performance of different classifiers in WEKA on datasets with numeric & nominal class attribute. During the evaluation, the input datasets and the number of classifier used are varied to measure the performance of Data Mining algorithm. Datasets are varied with mainly type of class attribute either nominal or numeric. We present the results for performance of different classifiers based on characteristics such as accuracy, time taken to build model identify their characteristics in acclaimed Data Mining tool-WEKA.

Classification maps data into predefined classes often referred as supervised learning because classes are determined before examining data. A classification algorithm is to use a training data set to build a model such that the model can be used to assign unclassified records in to one of the defined classes. A test set is used to determine the accuracy of the model. Usually, the given dataset is divided in to training and test sets, with training set used to build the model and test set used to validate it.

There are various classifiers are an efficient and scalable variation of Decision tree classification. The Decision tree model is built by recursively splitting the training dataset based on an optimal criterion until all records belonging to each of the partitions bear the same class label. Among many trees are particularly suited For data mining, since they are built relatively fast compared to other methods, obtaining similar or often better accuracy.

Bayesian classifiers are statistical based on Bayes’ theorem, they predict the probability that a record belongs to a particular class. A simple Bayesian classifier, called Naïve Bayesian classifier is comparable in performance to decision tree and exhibits high accuracy and speed when applied to large databases.

Rule-based classification algorithms generate if-then rules to perform classification. PART, OneR & ZeroR of Rule, IBK, and KStar of Lazy learners, SMO of Function are also used in evaluation process.

2. Related Work:

We studied various articles regarding performance evaluation of Data Mining algorithms on different tools, some of them are described here, Osama abu abbas worked on clustering algorithm, and Abdullah [3] compared various classifiers with different data mining tools , Tanuja S, Dr. U. Dinesh Acharya, and Shailesh K R [9] compared different data mining classification techniques to predict length of stay for an inpatient in hospital**.** Mahendra Tiwari & Yashpal Singh [1] evaluated performance of 4 clustering algorithms on different datasets in WEKA with 2 test modes. We presented their result as well as about tool and data set which are used in performing evaluation.

Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, & Emad M. Al-Shawakfa [3] in there article "A Comparison Study between Data Mining Tools over some Classification Methods" compared performance of different data mining tools like WEKA, KNIME, Orange & Tanagra over six different classifiers on 9 different datasets & found that WEKA performed better than rest three tools.

Osama Abu Abbas [2] in his article "comparison between data clustering algorithms by Osama Abu Abbas" compared four different clustering algorithms (K-means, hierarchical, SOM, EM) according to the size of the dataset, number of the clusters, type of S/W. Osama tested all the algorithms in LNKnet S/W- it is public domain S/W made available from MIT Lincoln lab www.li.mit.edu/ist/lnknet. For analyzing data from different data set, located at www.rana.lbl.gov/Eisensoftware.htm . The dataset that is used to test the clustering algorithms and compare among them is obtained from the site www.kdnuggets.com/dataset. This dataset is stored in an ASCII file 600 rows, 60 columns with a single chart per line.

Tanuja S, Dr. U. Dinesh Acharya, and Shailesh K R [9] in their article "Comparison of different data mining techniques to predict hospital length of Stay" compared four data mining classification techniques MLP, Naïve Bayes, K-NN, J48 to predict length of stay for an inpatient in hospital on preprocessed dataset derived from electronic discharge summaries with 401 instances & 16 parameters. In result they found that MLP performs better than rest three classifiers with 87.8% correctly classified instances.

Mahendra Tiwari & Yashpal Singh [1] in their article **"**Performance Evaluation of Data Mining clustering algorithm in WEKA**"** evaluated performance of 4 different clusterers (DBscan, EM, Hierarchical, Kmeans) on different datasets in WEKA with 2 test modes (Full training data, & Percentage split). They used 4 data sets for evaluation with clustering in WEKA, Two of them from UCI Data repository that are Zoo data set and Letter image recognition, rest two labor data set and supermarket data set is inbuilt in WEKA 3-6-6 .Zoo data set and letter image recognition are in csv file format, and labor and supermarket data set are in arff file format.

3. Evaluation Strategy/Methodology:

We used hardware as Pentium(R) D Processor platform which consist of 1 GB memory, Windows XP professional operating system, a 160GB secondary memory.

In all the experiments, We used Weka 3-7-7 , looked at different characteristics of the applications using classifiers to measure the accuracy in different data sets, time taken to build models etc.

Weka is acclaimed toolkit for machine learning and data mining that was originally developed at the University of Waikato in New Zealand. It contains 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 processing.

Input Data sets:

Input data is an important part of data mining applications. The data used in my experiment is real world data obtained from http://www.technologyforge.net/Datasets/ [9], during evaluation multiple data sizes were used, each dataset is described by the types of attributes, type of class (either nominal or numeric) the number of instances stored within the dataset, also the table demonstrates that all the selected data sets are used for the classification task. These datasets were chosen because they have different characteristics and have addressed different areas.

Details of data sets

All datasets have file type arff.

1. Dataset: EdibleMushrooms [4][5][6]

Attributes: 22 nominal,
Class: Nominal
Instances: 8124

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. There is no simple rule for determining the edibility of a mushroom; no rule like ``leaflets three, let it be'' for Poisonous Oak and Ivy.

Attribute Information:

1. cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
2. cap-surface: fibrous=f, grooves=g, scaly=y, smooth=s
3. cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
4. bruises?: bruises=t, no=f
5. odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
6. gill-attachment: attached=a, descending=d, free=f, notched=n
7. gill-spacing: close=c, crowded=w, distant=d
8. gill-size: broad=b, narrow=n

9. gill-color: black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e white=w, yellow=y
10. stalk-shape: enlarging=e, tapering=t
11. stalk-root: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
12. stalk-surface-above-ring: fibrous=f, scaly=y, silky=k, smooth=s
13. stalk-surface-below-ring: fibrous=f, scaly=y, silky=k, smooth=s
14. stalk-color-above-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
15. stalk-color-below-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
16. veil-type: partial=p, universal=u
17. veil-color: brown=n, orange=o, white=w, yellow=y
18. ring-number: none=n, one=o, two=t
19. ring-type: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
20. spore-print-color: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
21. population: abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
22. habitat: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

- Class:- Edible = Y, N

2. Dataset: LandformIdentification

Attributes: 6 numeric
Class: Nominal – 15 values
Instances: 300
Missing Values: None

This dataset contains satellite imaging data used for studying changes in terrain features on the Earth's surface. The goal is to correlate satellite measurements with terrain classification observations made by humans on the ground, so that changes in terrain can be tracked via satellite. The satellite data consists of numeric measurements of light intensity at six different wavelengths, which form the dataset attributes. The dataset contains 300 pixels of image data, which form the 300 instances in the dataset. For the LandformIdentification dataset, the terrain classifications are nominal values describing 16 different terrain types.

Attributes:

1. blue
2. green
3. red
4. nred
5. ir1
6. ir2

Classes:

1. Agriculture1
2. Agriculture2
3. Br_barren1
4. Br_barren2
5. Coniferous
6. Dark_barren
7. Deep_water
8. Marsh
9.N_deciduous
10.S_deciduous
11.Shallow_water
12.Shrub_swamp
13.Turf_grass
14.Urban
15.Wooded_swamp

3. Dataset: CPUPerformance [7]
Attributes: 6 numeric, 1 nominal,
Class: Numeric
Instances: 209
Missing Values: None
This dataset associates characteristics of CPU processor circuit boards, with the processing performance of the boards.
Attributes:

1. Vendor: Nominal, 30 vendor names.
2. MYCT: Numeric, cycle time in nanoseconds, 17-1500.
3. MMIN: Numeric, main memory minimum in KB, 64-32,000.
4. MMAX: Numeric, main memory maximum in KB, 64-64,000.
5. CACH: Numeric, cache memory in KB, 0-256.
6. CHMIN: Numeric, channels minimum, 0-52.
7. CHMAX: Numeric, channels maximum, 0-176.

Class:
Performance: Numeric, relative processing power, 15-1238.

4. Dataset: RedWhiteWine [8]
Attributes: 11 numeric, 1 nominal,
Class: Numeric
Instances: 6497
Missing Values: None

In the original form of this dataset, two datasets were created, using red and white wine samples. Here, these two datasets have been combined into one dataset. The inputs include objective tests (e.g. PH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).
Attributes:

1 - fixed acidity, numeric
2 - volatile acidity, numeric
3 - citric acid, numeric
4 - residual sugar, numeric
5 – chlorides, numeric
6 - free sulfur dioxide, numeric
7 - total sulfur dioxide, numeric
8 – density, numeric
9 – pH, numeric
10 – sulphates, numeric
11 – alcohol, numeric
12 – R/W, nominal – R= red, W = white
Class:
quality (score between 0 and 10) .

Experimental result and Discussion:-
To evaluate performance of selected tool using the given datasets, several experiments are conducted. For evaluation purpose, two test modes are used, the k-fold cross-validation (k-fold cv) mode, & percentage split (holdout method) mode. The k-fold cv refers to a widely used experimental testing procedure where the database is randomly divided in to k disjoint blocks of objects, then the data mining algorithm is trained using k-1 blocks and the remaining block is used to test the performance of the algorithm, this process is repeated k times. At the end, the recorded measures are averaged. It is common to choose k=10 or any other size depending mainly on the size of the original dataset.

In percentage split, the database is randomly split in to two disjoint datasets. The first set, which the data mining system tries to extract knowledge from called training set. The extracted knowledge may be tested against the second set which is called test set, it is common to randomly split a data set under the mining task in to 2 parts. It is common to have 66% of the objects of the original database as a training set and the rest of objects as a test set. Once the tests is carried out using the selected datasets, then using the available classification and test modes ,results are collected and an overall comparison is conducted.

4. Evaluation of classifier on Data sets:

We 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 3-7-7. The results after evaluation is described here:-

Table 1: Evaluation of classifiers on Edible Mushrooms dataset with Cross-validation mode.

Classifier model: - Full training set.

Classifier	**Time taken to build model**	**Test mode**	**Correctly classified instances**	**Incorrectly classified instances**	**Mean absolute error**	**Root Mean squared error**	**Relative absolute error**	**Root relative squared error**
Rules-ZeroR.	0 seconds	Cross-validation	4208/8124 (51.79%)	3916/8124 (48.20%)	0.4994	0.4997	100%	100%
Rules- PART	0.25 seconds	Cross-validation	8124/8124 (100%)	0/8124 (0%)	0	0	0%	0%
Decision table	4.73 seconds	Cross-validation	8124/8124 (100%)	0/8124 (0%)	0.0174	0.0296	3.4828%	5.916%
Lazy- IBk	0 seconds	Cross-validation	8124/8124 (100%)	0/8124 (0%)	0	0	0.0029%	0.003%
Bayes-NaiveBayes	0.2 seconds	Cross-validation	7785/8124 (95.82%)	339/8124 (4.17%)	0.0419	0.1757	8.3961%	35.16%
Functions-SMO	13.23 seconds	Cross-validation	8124/8124 (100%)	0/8124 (0%)	0	0	0%	0%
Trees-DecisionStump	0.05 seconds	Cross-validation	7204/8124 (88.67%)	920/8124 (11.32%)	0.1912	0.3092	38.29%	61.88%
Trees- J48	0.06 seconds	Cross-validation	8124/8124 (100%)	0/8124 (0%)	0	0	0%	0%

Table 2: Evaluation of classifiers on EdibleMushrooms dataset with Percentage split mode.

Classifier	**Time taken to build model**	**Test mode**	**Correctly classified instances**	**Incorrectly classified instances**	**Mean absolute error**	**Root Mean squared error**	**Relative absolute error**	**Root relative squared error**
Rules-ZeroR.	0.02 seconds	Percentage split	1410/2762 (51.05%)	1352/2762 (48.95%)	0.4995	0.5	100%	100%
Rules- PART	0.13 seconds	Percentage split	2762/2762 (100%)	0/2762 (0%)	0	0	0%	0%
Decision table	4.75 seconds	Percentage split	2762/2762 (100%)	0/2762 (0%)	0.02	0.03	4.04%	6.17%
Trees-DecisionStump	0.02 seconds	Percentage split	2464/2762 (89.21%)	298/2762 (10.78%)	0.1902	0.3033	38.07%	60.66%
Trees- J48	0.06 seconds	Percentage split	2762/2762 (100%)	0/2762 (0%)	0	0	0%	0%
Functions-SMO	13.42 seconds	Percentage split	2762/2762 (100%)	0/2762 (0%)	0	0	0%	0%
Bayes-NaiveBayes	0.02 seconds	Percentage split	2625/2762 (95.03%)	137/2762 (4.96%)	0.0485	0.1922	9.70%	38.44%
Lazy- IBk	0 seconds	Percentage split	2762/2762 (100%)	0/2762 (0%)	0	0	0.005%	0.006%

Table 3: Evaluation of classifiers on LandformIdentification dataset with Cross-validation.

Classifier	**Time taken to build model**	**Test mode**	**Correctly classified instances**	**Incorrectly classified instances**	**Mean absolute error**	**Root Mean squared error**	**Relative absolute error**	**Root relative squared error**
Rules-ZeroR.	0 seconds	Cross-validation.	20/300 (6.66%)	280/300 (93.33%)	0.1244	0.2494	100%	100%
Rules- OneR	0.02 seconds	Cross-validation.	208/300 (69.33%)	92/300 (30.66%)	0.0409	0.2022	32.85%	81.06%
Rules- PART	0.06 seconds	Cross-validation	285/300 (95%)	15/300 (5%)	0.007	0.0809	5.6391%	32.45%
Trees-DecisionStump	0.02 seconds	Cross-validation	40/300 (13.33%)	260/300 (86.66%)	0.1157	0.2405	92.94%	96.41%
Trees- J48	0 seconds	Cross-validation	292/300 (97.33%)	8/300 (2.66%)	0.004	0.0596	3.17%	23.90%
Functions-SMO	1.8 seconds	Cross-validation	273/300 (91%)	27/300 (9%)	0.1157	0.2349	92.97%	94.16%
Lazy- IBk	0 seconds	Cross-validation	297/300 (99%)	3/300 (1%)	0.0077	0.0378	6.2099%	15.17%
Bayes-NaiveBayes	0 seconds	Cross-validation	297/300 (99%)	3/300 (1%)	0.0015	0.0347	1.19%	13.92%

Table 4: Evaluation of classifiers on LandformIdentification dataset with Percentage split.

Classifier	**Time taken to build model**	**Test mode**	**Correctly classified instances**	**Incorrectly classified instances**	**Mean absolute error**	**Root Mean squared error**	**Relative absolute error**	**Root relative squared error**
Rules-ZeroR.	0 seconds	Percentage split	2/102 (1.96%)	100/102 (98.03%)	0.1252	0.2512	100%	100%
Rules- OneR	0.02 seconds	Percentage split	55/102 (53.92%)	47/102 (46.08%)	0.0614	0.2479	49.08%	98.65%
Rules- PART	0.03 seconds	Percentage split	99/102 (97.05%)	3/102 (2.94%)	0.0039	0.0626	3.13%	24.92%
Trees-DecisionStump	0.02 seconds	Percentage split	6/102 (5.88%)	96/102 (94.11%)	0.1172	0.2447	93.66%	97.41%
Trees- J48	0 seconds	Percentage split	96/102 (94.11%)	6/102 (5.88%)	0.0085	0.0888	6.7888%	35.35%
Functions-SMO	2.03 seconds	Percentage split	66/102 (64.70%)	36/102 (35.30%)	0.1162	0.236	92.85%	93.91%
Bayes-NaiveBayes	0 seconds	Percentage split	99/102 (97.05%)	3/102 (2.94%)	0.004	0.0603	3.22%	24.01%
Lazy- IBk	0 seconds	Percentage split	101/102 (99.01%)	1/102 (0.98%)	0.0099	0.039	7.90%	15.51%

Table 5: Evaluation of classifiers on CPUPerformance dataset with Cross-validation.

Classifier	**Time taken to build model**	**Test mode**	**Correlation Coefficient**	**Mean absolute error**	**Root Mean squared error**	**Relative absolute error**	**Root relative squared error**
Rules-ZeroR	0 seconds	Cross-validation	-0.2486	88.189	155.49	100%	100%
Rules- M5Rules	0.2 seconds	Cross-	0.9839	13.081	27.6918	14.83%	17.80%

		validation					
Trees- REPTree	0.02 seconds	Cross-validation	0.9234	25.56	59.81	31.26%	38.46%
Trees-DecisionStump	0.02 seconds	Cross-validation	0.6147	70.91	121.94	80.41%	78.41%
Lazy- IBk	0 seconds	Cross-validation	0.9401	20.92	56.70	23.73%	36.46%
Lazy- KStar	0 seconds	Cross-validation	0.9566	13.52	46.41	15.33%	29.84%
Functions-MLP	6.39 seconds	Cross-validation	0.9925	6.576	19.13	7.45%	12.30%
Functions-LinearRegression	0.03 seconds	Cross-validation	0.9337	34.79	55.26	39.44%	35.54%

Table 6: Evaluation of classifiers on CPUPerformance dataset with Percentage Split.

Classifier	**Time taken to build model**	**Test mode**	**Correlation Coefficient**	**Mean absolute error**	**Root Mean squared error**	**Relative absolute error**	**Root relative squared error**
Rules-ZeroR	0 seconds	Percentage split	0	83.39	113.04	100%	100%
Rules- M5Rules	0.14 seconds	Percentage split	0.9841	12.30	30.89	14.75%	27.32%
Trees- REPTree	0 seconds	Percentage split	0.9334	29.51	44.83	35.39%	39.66%
Trees-DecisionStump	0 seconds	Percentage split	0	69.53	112.13	83.38%	99.18%
Lazy- IBk	0 seconds	Percentage split	0.9038	22.19	52.02	26.61%	46.02%
Lazy- KStar	0 seconds	Percentage split	0.9652	12.85	36.38	15.41%	32.18%
Functions-MLP	6.41 seconds	Percentage split	0.9979	6.438	9.778	7.72%	8.64%
Functions-LinearRegression	0.03 seconds	Percentage split	0.9642	32.71	41.10	39.22%	36.36%

Table 7: Evaluation of classifiers on RedWhiteWine dataset with Cross-validation.

Classifier	**Time taken to build model**	**Test mode**	**Correlation Coefficient**	**Mean absolute error**	**Root Mean squared error**	**Relative absolute error**	**Root relative squared error**
Rules-ZeroR	0.02 seconds	Cross-validation	-0.0361	0.6856	0.8733	100%	100%
Rules- M5Rules	13.44 seconds	Cross-validation	0.5802	0.5532	0.7115	80.68%	81.47%
Trees- REPTree	0.22 seconds	Cross-validation	0.5566	0.5542	0.7322	80.82%	83.83%
Trees-DecisionStump	0.08 seconds	Cross-validation	0.3963	0.6605	0.8017	96.33%	91.79%
Trees- M5P	4.28 seconds	Cross-validation	0.5885	0.5467	0.7064	79.74%	80.88%
Functions-MLP	31.45 seconds	Cross-validation	0.505	5993	0.7624	87.40%	87.29%
Functions-LinearRegression	0.09 seconds	Cross-validation	0.5412	0.57	0.7343	83.12%	84.07%
Lazy- IBk	0 seconds	Cross-validation	0.5994	0.4299	0.7731	62.69%	88.52%

Table 8: Evaluation of classifiers on RedWhiteWine dataset with Percentage Split.

Classifier	Time taken to build model	Test mode	Correlation Coefficient	Mean absolute error	Root Mean squared error	Relative absolute error	Root relative squared error
Rules-ZeroR	0 seconds	Percentage split	0	0.6932	0.8863	100%	100%
Rules- M5Rules	13.56 seconds	Percentage split	0.5486	0.5653	0.7414	81.54%	83.64%
Trees- REPTree	0.2 seconds	Percentage split	0.5204	0.5771	0.7617	83.25%	85.93%
Trees-DecisionStump	0.09 seconds	Percentage split	0.3962	0.6664	0.8138	96.13%	91.81%
Trees- M5P	4.49 seconds	Percentage split	0.5821	0.5604	0.7212	80.85%	81.36%
Functions-MLP	31.16 seconds	Percentage split	0.5775	0.6267	0.7971	90.40%	89.93%
Functions-LinearRegression	0.08 seconds	Percentage split	0.5454	0.5722	0.743	82.54%	83.82%
Lazy- IBk	0 seconds	Percentage split	0.5709	0.4599	0.8085	66.35%	91.21%

5. Conclusions:

For each characteristic, we analyzed how the results vary whenever test mode is changed. Our measure of interest includes the analysis of classifiers on different datasets, the results are described in value of correctly classified instances & incorrectly classified instances (for dataset with nominal class value), correlation coefficient (for dataset with numeric class value), mean absolute error, root mean squared error, relative absolute error, root relative squared error after applying the cross-validation or Percentage split method.

Different classifiers like rule based(ZeroR, OneR, PART), tree based(Decisionstump,J48,REP), function(SMO,MLP), bayes(NaiveBayes), Lazy(IBk,Kstar) are evaluated on four different datasets.

Two datasets (EdibleMushrooms &LandformIdentification) have nominal class value & other two (CPUPerformance & RedWhiteWine) have numeric class value.

Most algorithms can classify both types of datasets with nominal & numeric class value. But there are also some algorithms that can only classify datasets with either nominal or numeric class value, such as Bayes algorithms able to classify datasets only with nominal class whereas Linear regression, M5Rules able to classify datasets only with numeric class value.

For all datasets, results with both test modes i.e. K-fold cross-validation & percentage split are nearly same. For dataset Edible Mushroom PART, SMO, J48 performed well with 100% correctly classified instances, from these J48 has been observed best due to least time taken to build model. For Landform Identification IBk & Naïve Bayes performed well with 99% correctly classified instances. For RedWhiteWine all algorithms performed average. For CPU Performance M5Rules, MLP performs well with 0.98 & 0.99 correlation coefficient. Remaining algorithms except ZeroR have performed moderate. As ZeroR predicts only one class value it performs poor for almost all datasets.

References:

[1] Mahendra Tiwari, Yashpal Singh (2012), Performance Evaluation of Data Mining clustering algorithms in WEKA, Global Journal of Enterprise Information System, vol 4, issue I.

[2] Osama A. Abbas (2008), Comparison between data clustering algorithm, The International Arab journal of InformationTechnology, vol 5, N0. 3.

[3] Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa , A Comparison Study between Data Mining Tools over some Classification Methods, (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence.

[4] Iba,W., Wogulis,J., & Langley,P. (1988). Trading off Simplicity and Coverage in Incremental Concept Learning. In Proceedings of the 5th International Conference on Machine Learning, 73-79. Ann Arbor, Michigan: Morgan Kaufmann.

[5] Duch W, Adamczak R, Grabczewski K (1996) Extraction of logical rules from training data using back propagation networks, in: Proc. of the 1st Online Workshop on Soft Computing, 19-30.Aug.1996, pp. 25-30

[6] Duch W, Adamczak R, Grabczewski K, Ishikawa M, Ueda H, Extraction of crisp logical rules using constrained backpropagation networks - comparison of two new approaches, in: Proc. of the European Symposium on Artificial Neural Networks (ESANN'97), Bruge, Belgium 16-18.4.1997.

[7] Kilpatrick, D. & Cameron-Jones, M. (1998), Numeric prediction using instance-based learning with encoding length selection, Progress In Connectionist-Based Information Systems. Singapore: Springer-Verlag.

[8] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553.

[9] Tanuja S, Dr. U. Dinesh Acharya, and Shailesh K R (2011) , Comparison of different data mining techniques to predict hospital length of Stay, Journal of Pharmaceutical and Biomedical Sciences (JPBMS), Vol. 07, Issue 07

[10] http://www.technologyforge.net/Datasets/

[11] http://www.let.rug.nl/tiedeman/ml06/InterpretingWekaOutput

[12] http://www.cs.utoronto.ca/~delve/data/mushrooms/mushroomDetail.html

[13] http://weka.wikispaces.com/Primer