Data Classification based on Decision Tree, Rule Generation, Bayes and Statistical Methods: An Empirical Comparison

Sanjib Saha
Assistant Professor
Department of Computer Science and Engineering
Dr. B.C. Roy Engineering College, Durgapur, India

Debashis Nandi, PhD
Associate Professor
Department of Information Technology
National Institute of Technology, Durgapur, India

ABSTRACT
In this paper, twenty well known data mining classification methods are applied on ten UCI machine learning medical datasets and the performance of various classification methods are empirically compared while varying the number of categorical and numeric attributes, the types of attributes and the number of instances in datasets. In the performance study, Classification Accuracy (CA), Root Mean Square Error (RMSE) and Area Under Curve (AUC) of Receiver's Operational Characteristics (ROC) is used as the metric and come up with some findings: (i) performance of classification methods depends upon the type of dataset variables or attributes such as categorical, numeric and both (mixed), (ii) performance of classification methods on categorical attributes is superior than on numeric attributes of a dataset, (iii) classification accuracy, RMSE and AUC of a classification method depends on the number of instances in datasets, (iv) classification performance decreases in case of instances decreases for both categorical as well as numeric datasets, (v) top three classification methods are established after comparing the performance of twenty different classification methods for the categorical, numeric and both (mixed) attribute datasets, (vi) out of these twenty different classification methods Bayes Net, Naïve Bayes, Classification Via Regression, Logistic Regression and Random Forest method performs best on these medical datasets.

Keywords
Data Mining; Classification; Classification Accuracy; RMSE; ROC; Confusion Matrix

1. INTRODUCTION
Data mining is the technique of extracting previously unknown and potentially useful knowledge from large amount of data [1]. It is also known as Knowledge Discovery in Databases (KDD). Pattern discovery, Association & correlation, Classification, Clustering and Outlier analysis are the part of data mining.

Classification is the process to construct a model based on the training set and uses it to classify new data or test set. It is a supervised learning as observations; measurements are accompanied by known class labels in a large amount of training set and new data is classified based on training set [1].

The main aim of this research is to determine which learning algorithm to select based on the characteristics of the given dataset to perform better, without trial-and-error testing on different available algorithms.

Classification Accuracy is a metric which is defined as the percentage of number of correctly classified instances.

Root Mean Square Error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed.

Receiver Operating Characteristics (ROC) curve which is plotted with the probability of the class prediction has been introduced to evaluate performance of machine learning algorithms. Bradley [2] compared popular machine learning algorithms using area under the curve of ROC, and found that area under ROC exhibits several desirable properties compared to accuracy. A model with perfect accuracy will have an area 1.0.

Confusion Matrix [3] is a specific table layout that visualize the performance of machine learning algorithm, typically a supervised learning. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class.

2. LITERATURE SURVEY
There is several research papers have been studied related to comparison of classification methods based on datasets.

Statistical pattern recognition, neural nets, and machine learning classification methods were applied to four real-world medical datasets and empirically compared the true error rates and found machine learning procedures for rule induction or tree induction performed best [4].

An empirical study on 7 individual supervised machine learning methods and 9 different combined methods were applied on 4 different biological datasets and found combined methods perform better than the individual ones in terms of their specificity, sensitivity, positive predicted value and accuracy. Also, statistical methods (e.g. SVM, neural networks) tend to perform much better over multi-dimensions and numeric attributes but rule-based systems (e.g. Decision trees, PART) tend to perform better in discrete / categorical attributes [5].

An experimental investigation of the effect of discrete attributes on the precision of classification methods based on Area Under Curve (AUC) and found with increasing the number of the discrete attributes or with increasing the number of values in discrete attributes, the AUC of logistic regression is increased but linear classifier's AUC decreases, and AUC of the naïve-Bayes classifier remains constant [6].

Comparing Naive Bayes, Decision Trees (C4.5), and SVM with AUC and Accuracy and found that AUC was a better measure than accuracy and C4.5 gave better AUC score while accuracy was similar for these three methods [7].

Comparison of artificial neural network, decision tree and linear regression methods using RMSE value and found that

for numeric and categorical independent variables, linear regression was best when number of categorical variables was one and the artificial neural network was superior when the number of categorical variables was two or more regardless of the number of variables and sample size [8].

Performance measure using Area Under ROC Curve (AUC) for six machine learning algorithms were done on six real-world medical diagnostics datasets and found that AUC measure better than accuracy in evaluation of machine learning algorithms [2].

Using AUC and accuracy in evaluating machine learning algorithms, it was found that Naive Bayes and decision trees were very similar in predictive accuracy while Naive Bayes was significantly better than decision trees in AUC and AUC was a better measure than accuracy [9].

Comparative analysis of logistic regression and artificial neural network for computer-aided cancer diagnosis on breast sonograms was done. Performance measured by AUC was same for both. However, at a fixed 95% sensitivity, the artificial neural network had higher (12%) specificity compared with logistic regression value [10].

Popular rule-based classification techniques named Decision Tree, JRIP, NNGE, PART, and RIDOR are applied on eleven different medical datasets to interpret their applicability in classifying patients into groups. It was found that Decision Tree and PART algorithm outperforms than others [11].

Support Vector Machine (SVM) is a well-known soft computing and data mining classification technique. Several popular methods related to SVM are compared and analytically surveyed to find their best application areas [12].

3. DATA ANALYSIS

Datasets: Ten medical datasets (such as Colic, Heart-c, Hepatitis, Lymph, Audiology, Breast cancer, Primary Tumor, Breast cancer-w, Heart Stat-log and Diabetes) have been taken from UCI Machine learning repository [13] and all the datasets are in arff (attribute relation file format). It has been considered that some equivalent datasets, more or less similar in their ratio of categorical and numeric attributes and the instances. First four datasets are used for comparing performance of different classification methods and rest six datasets are used to check whether performance is dependent on number of instances in datasets.

Relationships of each dataset are denoted by c (categorical), n (numeric) and k (constant) and one constant should not be selected.

Table 1: Characteristics of the Data Sets used

Dataset	Numeric	Categorical	Instance	Relationship
Colic	7	16	150 of 368	y=1+15c+7n+k
Heart-c	6	8	150 of 303	y=1+7c+6n+k
Hepatitis	6	14	150 of 155	y=1+13c+6n+k
Lymph	3	16	150 of 150	y=1+15c+3n+k

Table 2: Characteristics of the Data Sets used

Dataset	Numeric	Categorical	Instance	Relationship
Audiology	0	70	226	y=1+69c+k
Breast cancer	0	10	286	y=1+9c+k
Primary Tumor	0	18	339	y=1+17c+k
Breast cancer - w	9	1	699	y=1+9n+k
Heart Stat-log	13	1	270	y=1+13n+k
Diabetes	8	1	768	y=1+8n+k

The objective of this paper is to analyze the change in performance of different classification methods using medical datasets as the training and test data, while considering categorical, numeric and both (mixed) attributes one at a time individually studying the effect it has on classification, even changing number of instances of the datasets.

Experiment is done by individually pruning each of four specially picked medical datasets (Colic, Heart-c, Hepatitis and Lymph) as purely categorical, purely numeric and both (mixed) attributes and run each of the mentioned rule based classification technique through the use of Weka tool [14] (a java based GUI tool).

Make a table, taking into consideration only the classification accuracy and ROC area metrics into consideration.

Further, used six more medical datasets to show how classification accuracy of different classification methods are dependent on the sample size by changing the number of instances in each dataset.

These ten medical datasets have been taken considering that their attributes may depend with each other.

4. EXPERIMENTAL RESULT

In experiments using Weka 3.6.9 tool [14], attributes of the dataset have been first selected. Then cross validation of 10 folds have been chosen as test method. The fold determines the amount of data used for pruning; one fold is used for pruning and the rest for growing the rules. After that the particular classification method would be chosen and also the parameters would be specified as search method greedy stepwise, seed value 1 is used for randomizing the data, prune value True is used whether pruning is performed, debug value False is used whether debug information is output to the console, confidence factor 0.25 is used for pruning (smaller values giver more pruning).

While doing these experiments, it has been considered that performance of twenty data mining classification methods: Bayes Net [15], Naïve Bayes [16], Naïve Bayes Simple [16], Decision Tree J4.8 [17], Random Forest [18], Naïve Bayes-Tree [19], Simple CART [20], Decision Stump [21], Classification via Regression [22], Vote [23], Voted Perceptron [24], Multiclass Classifier [25], VFI [26], Logistic

Regression [27], K-NN [28] and rule based classification methods: Decision Table [29], JRIP [30], NNGE [31], PART [32], and RIDOR [33] applied on categorical, numeric and both (mixed) attributes. Accordingly, it has been found top five methods, and separately top three methods for each kind of dataset be it numeric, categorical or both (mixed). Weka has been chosen because all these classification methods are available in this tool and it is very user friendly in editing datasets and testing modes. It has been chosen 10 folds cross validation as test option and classification accuracy (%), RMSE (Root Mean Squared Error) and area under ROC used as a metric to measure the performance throughout these experiments.

Performance of Top 6 Classification methods on specific set of attributes

Table 3: Classification Accuracy (%) obtained [w.r.t.instance fixed to 150 on each datasets]

Method	Attribute	Colic	Heart-C	Hepatitis	Lymph
Naive Bayes	Numeric	51.33	77.33	81.93	73.64
	Mixed	76.00	86.00	84.51	83.10
	Categorical	76.00	81.33	83.87	83.78
Bayes Net	Numeric	64.00	71.33	81.64	65.54
	Mixed	77.33	82.67	83.22	85.81
	Categorical	77.33	81.33	83.22	83.10
Classification Regression	Numeric	62.66	80.66	81.93	72.29
	Mixed	77.33	82.00	81.93	79.05
	Categorical	76.00	84.00	81.93	83.78
Logistic Regression	Numeric	64.00	78.66	83.22	72.97
	Mixed	66.00	82.67	82.58	73.64
	Categorical	72.66	81.33	84.51	79.72
Random Forest	Numeric	62.00	76.00	77.41	68.91
	Mixed	78.00	79.33	85.80	83.78
	Categorical	78.00	73.33	79.35	81.08
PART	Numeric	69.79	65.10	78.71	67.57
	Mixed	**65.10**	68.46	**83.33**	76.35
	Categorical	64.43	**73.15**	81.29	**80.41**

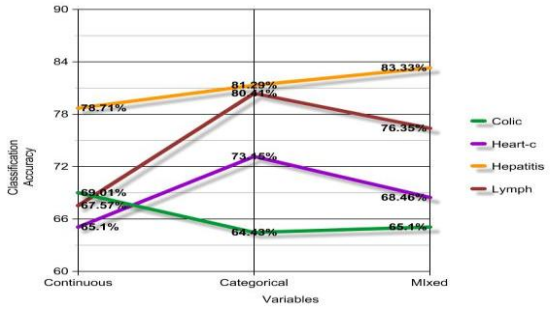

Figure 1: The above figure depicts graphically the stats of PART method, as presented above in [Table 3].

From the graphical representation of classification accuracy against the type of attributes in the particular datasets, it has been seen that categorical attributes have a very significant role in performance of classification. The general rule is that numeric attributes alone hampers the performance of most classification methods and whereas categorical attributes alone boost the performance of most classification methods.

Here analyzing these facts of performance of various classification methods while keeping the instance fixed to 150 on each medical datasets. The above table [Table 3] shows the list of top six classification methods which are shortlisted from these twenty data mining classification methods.

Table 4: Area under ROC obtained [w.r.t fixed instance 150]

Method	Attribute	Colic	Heart-C	Hepatitis	Lymph
Naive Bayes	Numeric	0.660	0.823	0.817	0.752
	Mixed	0.842	0.904	0.860	0.908
	Categorical	**0.846**	0.883	0.865	0.918
Bayes Net	Numeric	0.686	0.817	0.782	0.713
	Mixed	0.843	0.907	0.882	0.916
	Categorical	**0.848**	0.883	0.865	0.919
Classification Regression	Numeric	0.684	0.829	0.805	0.776
	Mixed	0.870	0.887	0.825	0.904
	Categorical	**0.884**	0.804	0.839	0.909
Logistic Regression	Numeric	0.662	0.839	0.835	0.753
	Mixed	0.828	0.909	0.802	0.830
	Categorical	**0.838**	0.873	0.804	0.822
Random Forest	Numeric	0.708	0.782	0.738	0.724
	Mixed	0.889	0.867	0.852	0.922
	Categorical	**0.881**	0.820	0.814	0.887

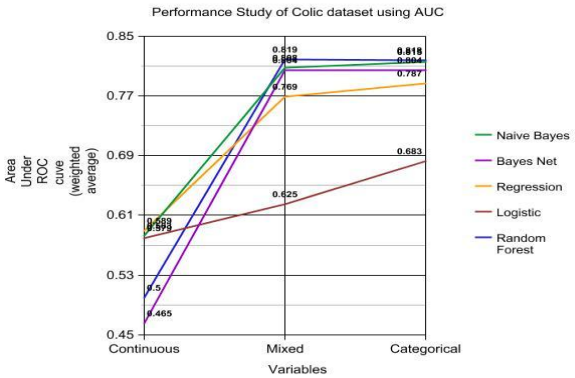

Figure 2: The above figure depicts graphically the stats of Colic dataset, as presented above in [Table 4].

From the graphical representation of Area Under ROC curve against the type of attributes in the particular datasets, it has been shown that categorical attributes have a significant roleas it increases visualization of the trade-off between the rate at which the model can accurately recognize 'yes' cases versus the rate at which it mistakenly identifies 'no' cases as 'yes' for different "portions" of the test set. Thus, it is making less error in classification.

Table 5: Area under ROC obtained [w.r.t. default number of instances as in each datasets]

Method	Attribute	Colic	Heart-C	Hepatitis	Lymph
Naive Bayes	Numeric	0.660	0.823	0.817	0.752
	Mixed	0.842	0.904	0.860	0.908
	Categorical	**0.846**	0.883	0.865	0.918
Bayes	Numeric	0.686	0.817	0.782	0.713

Net	Mixed	0.843	0.907	0.882	0.916
	Categorical	**0.848**	0.883	0.865	0.919
Classification Regression	Numeric	0.684	0.829	0.805	0.776
	Mixed	0.870	0.887	0.825	0.904
	Categorical	**0.884**	0.804	0.839	0.909
Logistic Regression	Numeric	0.662	0.839	0.835	0.753
	Mixed	0.828	0.909	0.802	0.830
	Categorical	**0.838**	0.873	0.804	0.822
Random Forest	Numeric	0.708	0.782	0.738	0.724
	Mixed	0.889	0.867	0.852	0.922
	Categorical	**0.881**	0.820	0.814	0.887

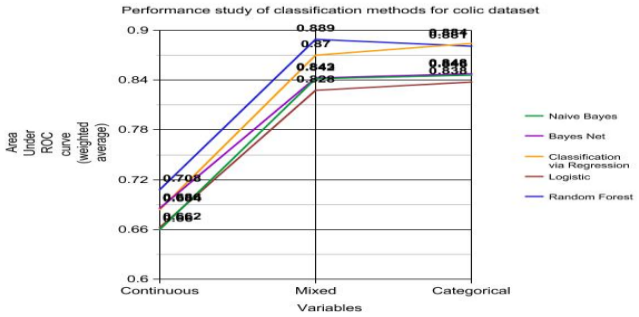

Figure 3: The above figure depicts graphically the stats of Colic dataset, as presented above in [Table 5].

Top 3 Classification methods on *mixed* attributes

Table 6: RMSE obtained [w.r.t. default number of instances as in each datasets]

Method	Colic	Heart-C	Hepatitis	Lymph
Naïve Bayes	0.419	0.229	0.363	0.266
Bayes Net	**0.405**	0.231	**0.371**	0.242
Random Forest	0.348	**0.239**	0.331	**0.251**

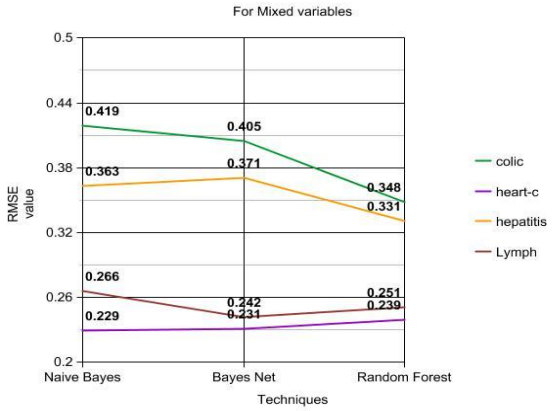

Figure 4: The below figure depicts graphically the stats of top 3 classification methods on *mixed* attributes, as presented above in [Table 6].

Top 3 Classification methods on *categorical* attributes

Table 7: RMSE obtained [w.r.t. default number of instances as in each datasets]

Methods	Colic	Heart-C	Hepatitis	Lymph
NB-Tree	0.374	0.235	**0.374**	0.239
Classification Regression	0.344	0.233	0.359	**0.255**
Bayes Net	**0.399**	**0.236**	0.365	0.244

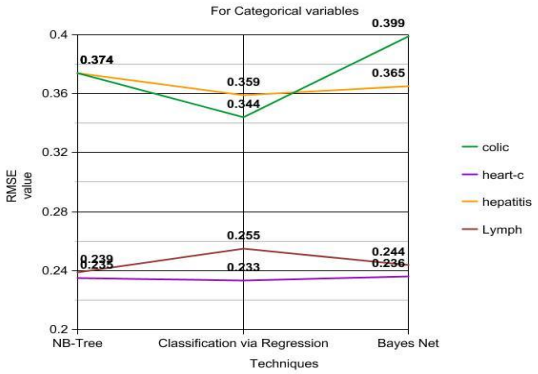

Figure 5: The above figure depicts graphically the stats of top 3 classification methods on *categorical* attributes, as presented above in [Table 7].

Top 3 Classification methods on *numeric* attributes

Table 8: RMSE obtained [w.r.t. default number of instances as in each datasets]

Methods	Colic	Heart-C	Hepatitis	Lymph
Classification Regression	0.456	0.259	0.355	0.310
NB-Tree	0.453	0.270	**0.360**	0.346
Multiclass Classifier	**0.468**	**0.371**	0.346	**0.392**

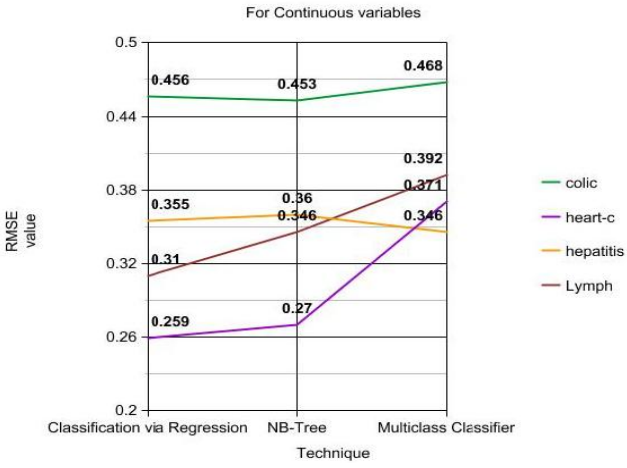

Figure 6: The above figure depicts graphically the stats of top 3 classification methods on *numeric* attributes, as presented above in [Table 8].

5. PERFORMANCE ANALYSIS

It has been taken into consideration that medical datasets of different instances is used to determine the performance of various classification methods of different classes on categorical, numeric and both (mixed) types of attributes. For

comparing, each medical datasets are edited to 150 instances so that comparison can be done accurately. At first executing the above mentioned classification methods on datasets containing both types of attributes. After that categorical attributes are removed from datasets to check the performance of above mentioned methods on only numeric types of attributes and similarly remove numeric attributes from datasets to check the performance on only categorical attributes. After executing different methods on different attributes type of datasets, it has been noted down the value of classification accuracy, RMSE and ROC value into a table format as given above. After noting down the value in table, it is observed that the performance value of each classification methods for each type of attributes (categorical, numeric and mixed) and found that most of the classification methods gives better result for categorical type of attributes as compare to numeric and mixed attributes type and out of these twenty classification methods Bayes Net, Naïve Bayes, Classification via Regression, Logistic Regression and Random Forest are best. It has been seen that by using default instances/increase the instances of the medical datasets will get more accurate results in comparison to the fixed instances/decrease the instances of the same datasets. It is also checked that various classification methods of separate sets of same medical datasets i.e. in numeric, categorical and both (mixed), it is found that there are some methods which perform better on different categories of attributes.

6. CONCLUSION

It has observed from Table 3,4 that more or less every classification method of different categories are giving comparatively better result for categorical attributes than numeric attributes. It has also observed from Table 5 that classification performance decreases in case of instances decreases for both categorical as well as numeric datasets. Out of these twenty classification methods Bayes Net, Naïve Bayes, Classification via Regression, Logistic Regression and Random Forest classification methods are the best.

For mixed attribute datasets Naïve Bayes, Bayes Net and Random Forest classification methods are the best. For numeric attribute datasets Classification via Regression, NB-Tree and Multiclass Classifier methods are the best. For categorical attribute datasets NB-Tree, Classification via Regression and Bayes Net methods are the best. Out of these above five rule based classification methods PART and Decision Tree methods are the best.

7. REFERENCES

[1] Jiawei Han, Micheline Kember, Jian Pei, "Data Mining Concepts and Techniques", 3rd Edition, Morgan Kaufmann, 2012.

[2] Andrew P. Bradley, "The use of area under ROC curve in evaluation of machine learning algorithms", Pattern Recognition Society, 1997.

[3] Stehman, Stephen V, "Selecting and interpreting measures of thematic classification accuracy", Remote Sensing of Environment, 1997.

[4] Sholom M. Weiss, Ioannis Kapouleas, "An Empirical Comparison of Pattern Recognition, Neural Nets, and Machine Learning Classification Methods", Machine Learning.

[5] Aik Choon Tan, David Gilbert, "An empirical comparison of supervised machine learning techniques in bioinformatics", Proceedings of 1st Asia Pacific Bioinformatics Conference, 2003.

[6] Reza Entezari-Maleki, Seyyed Mehdi Iranmanesh, Behrouz Minaei-Bidgoli, "An Experimental Investigation of the Effect of Discrete Attributes on the Precision of classification Methods", World Applied Sciences Journal 7 (Special Issue of Computer & IT), 2009, 216-223.

[7] Jin Huang, Jingjing Lu, Charles X. Ling, "Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy", 3rd IEEE International Conference on Data Mining, 2003.

[8] Yong Soo Kim, "Comparison of the decision tree, artificial neural network and linear regression methods based on the number and types of independent variables and sample size", Expert Systems with Applications, 2008, 1227–1234.

[9] Jin Huang, Charles X. Ling, "Using AUC and Accuracy in Evaluating Learning Algorithms", IEEE Transactions on Knowledge and Data Engineering, March, 2005, Vol. 17, No. 3.

[10] Jae H. Song, Santosh S. Venkatesh, Emily A. Conant, Peter H. Arger, Chandra M. Sehgal, "Comparative Analysis of Logistic Regression and Artificial Neural Network for Computer-Aided Diagnosis of Breast Masses", Academic Radiology, April, 2005, Vol. 12.

[11] R P Datta, Sanjib Saha, "Applying rule based classification techniques to medical databases: An empirical study", International Journal of Business Intelligence and Systems Engineering (IJBISE), Inderscience Publishers, 2015.

[12] Subhankar Das, Sanjib Saha, "Data Mining and Soft Computing using Support Vector Machine: A Survey", International Journal of Computer Applications (0975-8887), Volume 77-No.14, September 2013.

[13] Blake C, Merz C, "UCI repository of machine learning datasets", 2000.

[14] WEKA 3.6.9 java based GUI tool popularly used for machine learning and knowledge analysis (http://www.cs.waikato.ac.nz/~ml/weka/). Provided by the Machine Learning Group at the University of Waikato, Hamilton, New Zealand, 1999-2013.

[15] N. Friedman, D. Geiger, M. Goldszmidt, "Bayesian network classifiers", Machine Learning, 1997, 29:131-163.

[16] George H. John, Pat Langley, "Estimating Continuous Distributions in Bayesian Classifiers", Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, 1995, pp. 338-345.

[17] J R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993, San Mateo, CA.

[18] Leo Breiman, "Random Forests", Machine Learning, 2001, 45(1):5-32.

[19] Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid", In: 2nd International Conference on Knowledge Discovery and Data Mining, 1996, 202-207.

[20] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone, "Classification and Regression Trees,

1984, Wadsworth International Group, Belmont, California.

[21] Iba. Wayne, Langley. Pat, "Induction of One-Level Decision Trees", Proceedings of 9th International Conference on Machine Learning, Aberdeen, Scotland, 1992, San Francisco, CA: Morgan Kaufmann.

[22] E. Frank, Y. Wang, S. Inglis, G. Holmes, I.H. Witten, "Using model trees for classification", Machine Learning, 1998, 32(1):63-76.

[23] Eric Bauer, Ron Kohavi, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants", Machine Learning, 1998, vv, 1-38.

[24] Y. Freund, R. E. Schapire, "Large margin classification using the perceptron algorithm", In 11th Annual Conference on Computational Learning Theory, New York, 1998, 209-217.

[25] David M.J. Tax, Robert P.W. Duin, "Using two-class classifiers for multiclass classification", IEEE International Conference on Pattern Recognition, 2002.

[26] G. Demiroz, A. Guvenir, "Classification by voting feature intervals", In 9th European Conference on Machine Learning, 1997, 85-92.

[27] Le Cessie, Van Houwelingen, "Ridge Estimators in Logistic Regression", Appl. Statist, 1992, 41, No. 1, pp. 191-201.

[28] D. Aha, D. Kibler, "Instance-based learning algorithms", Machine Learning, 1991, 6:37-66.

[29] Kohavi Ron, "The Power of Decision Tables", In: 8th European Conference on Machine Learning, 1995, 174-189.

[30] Cohen William W, "Fast Effective Rule Induction", In: 12th International Conference on Machine Learning, 1995, 115-123.

[31] Martin Brent, "Instance-Based learning: Nearest Neighbor With Generalization", Hamilton, New Zealand, 1995.

[32] Frank Eibe, Ian H. Witten, "Generating Accurate Rule Sets Without Global Optimization", In: 15th International Conference on Machine Learning, 1998, 144-151.

[33] Gaines Brian R, Compton Paul, "Induction of Ripple-Down Rules Applied to Modeling Large Databases", J. Intell. Inf. Syst., 1995.