Volume 6, Issue 12, December 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Implementation of Novel Algorithm in the Rule Based Expert System for Classification Accuracy

C. Parthiban[*]
Research Scholar, PRIST University,
Tamil Nadu, India

M. Balakrishnan
Principal Scientist, (NAARM), Hyderabad,
Andhra Pradesh, India

Abstract– This research paper objective is to provide analysis of classifying the data set using novel naïve classification algorithm. Here this algorithm applied on two dataset which are the first one is training data set and other one is test data set, and this experiment attempts to compare the classification and accuracy of the proposed algorithm with the two dataset . Several constraints used for analytical purpose which are classification accuracy, True Positive Rate, False positive Rate sensitivity, Precision, Recall and f-measure using confusion matrix. Finally results are given in the tabular form to facilitate comparative analysis. From this study it is found that this algorithm given good result for both train as well as test data set but more accurate for train set. Section I discussed the general approach of classification, section II discussed Literature review, section III discussed the proposed work about the performance of train and test dataset, result and discussion briefed in the section IV and conclusion in section V.

Keywords – Classification, Novel Naïve, Confusion Matrix, Data set, Train set and Test set

I. INTRODUCTION

1.1 General Approach to Classification

A classification technique, or a classifier, is a systematic approach for the classification models from an input data set to output class. Examples include,

- Rule-Based Classifiers
- Naïve Bayes Classifiers.
- Neural Networks
- Decision Tree Classifiers
- Support Vector Machines

Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. The model generated by a learning algorithm should both fit the input data correctly predict the class labels of records it has never seen before. Therefore, a key objective of the learning algorithm is to build models with good generalization capability, i.e., models that accurately predict the class labels of previously unknown records, Fig. 1 shows the general approach of process classification for solving classification problems. Classification is a data mining method that assigns objects in a group to target class. The intention of classification is to accurately visualize the target class for each case in the data [1]. Data mining has been used to analyze large data sets and establish useful classification and patterns in the data sets. "Agricultural and biological research studies have used various techniques of data analysis including, natural trees, statistical machine learning and other analysis methods [2].

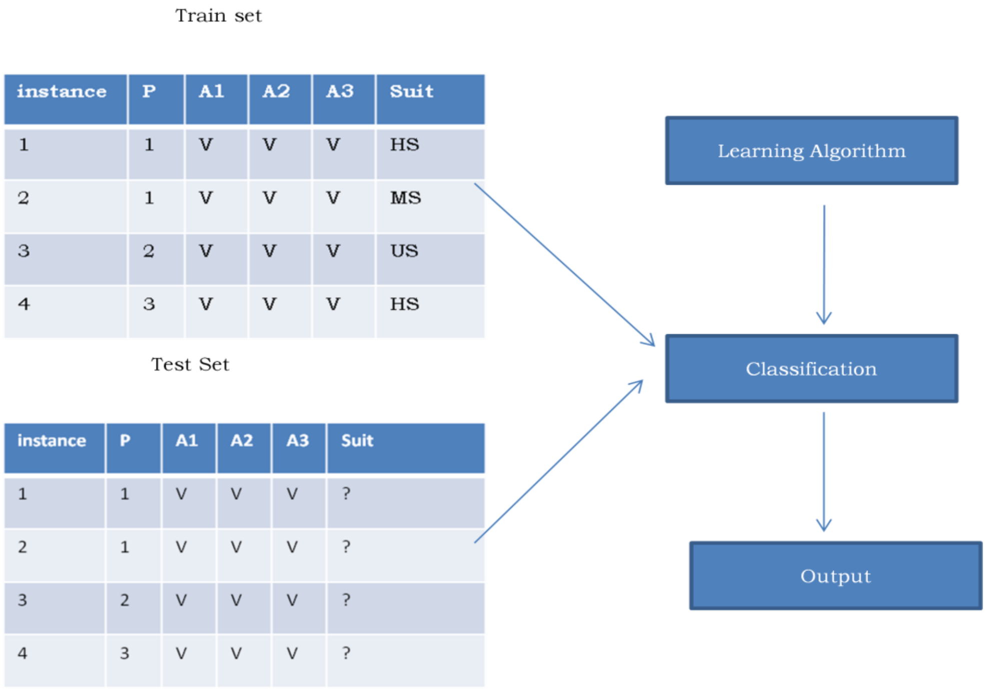
Bayes classifier proposes simple and powerful classification method. Naive Bayes classifier is based on the classical Bayes theorem which works on the probability theory. An important goal for many problems solving system is to collect evidence as the system goes along and to modify its behavior on the basis of the evidence. To model this behavior, we need a statistical theory of evidence. Bayes statistics is such a theory. The fundamental of notion of bayes statistics is conditional probability. The classifier is based on Bayes theorem, which is stated as:

Bayes Theorem:

$$P(d_i / E) = \frac{P(d_i) * P(E / d_i)}{\sum_j \mathrm{P(d_j) * P(E/d_j)}}$$

This is possible by modifying the Bayes theorem with novel approach.

© 2016, IJARCSSE All Rights Reserved



Train set

instance	P	A1	A2	A3	Suit
1	1	V	V	V	HS
2	1	V	V	V	MS
3	2	V	V	V	US
4	3	V	V	V	HS

Test Set

instance	P	A1	A2	A3	Suit
1	1	V	V	V	?
2	1	V	V	V	?
3	2	V	V	V	?
4	3	V	V	V	?

Fig. 1 Process of Classification

First, a training set consisting of records whose class labels are known must be provided. The training set is used for a classification then which is used afterward applied to the test set, which is consist of records with unknown class label. The below fig. 2 show the proposed system's general approach for classification of the data set.

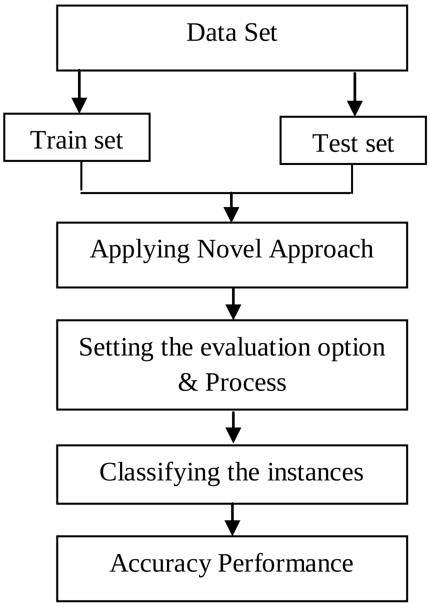


Fig. 2 Proposed System's Technique for classification

II. LITERATURE REVIEW

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining is automated data analysis techniques to discover earlier undetected relationships among data items [3]. Machine learning algorithms typically used in data mining have been applied to learn rules for an expert system based on examples provided by experts [4]. Classification is a supervised method that learns to classify the instances based on the knowledge learnt from a previously classified training set of instances. The rule based classification algorithms are: Decision Table, OneR and PART etc. Decision Table employs the wrapper method to find a good subset of attributes for inclusion in the table [5]. OneR creates one rule for each parameter in the training data, then selects the rule with the minimum error rate as its one rule [6]. PART algorithm producing sets of rules called decision lists which are ordered set of rules. Classification rule method in data mining are compared for envisage heart disease. The rule algorithms are Decision table, JRip, PART and OneR. By examine the experimental results of accuracy measure; it was observed that the Decision Table classification rule technique turned out to be best classifier for heart disease prediction because it has more accuracy. After analyzing the error rate it was concluded that the Decision table and OneR classification rule algorithm contains least error rate in possible two outcomes [7].

The comparative evaluation of classifiers NAIVE BAYES AND J48 in the context of bank dataset to maximize true positive rate and minimize false positive rate of defaulters rather than achieving only higher classification accuracy using WEKA tool. The experiments results shown are about classification accuracy, sensitivity and specificity. The results on this dataset show that the efficiency and accuracy of j48 is better than that of Naïve bayes [8].

III. METHODOLOGY

3.1 Train and Test Dataset Performance

Here two data sets were used which are train set with 337 instances and test set with 122 instances, which consists of records with class labels, to compare the accuracy (confusion matrix) of the data set. Here the data set is partitioned into a training and test set.

3.2 Confusion Matrix

The Confusion matrix exemplifies the correctness of the solution to the classification problem. The correct and incorrect instances that give us a most efficient method for classification by using the confusion matrix [9]. Confusion matrix contains information about actual and predicted classifications done by a classification system [10].The confusion matrix have the following meaning.

1. a is the number of correct predictions that an instance is negative,
2. b is the number of incorrect predictions that an instance is positive,
3. c is the number of incorrect of predictions that an instance negative,
4. d is the number of correct predictions that an instances positive [11].

3.3 Standards and terms of Weka Data Analysis

- TP = true positives: number of examples predicted positive that are actually positive.
- FP = false positives: number of examples predicted positive that are actually negative.
- TN = true negatives: number of examples predicted negative that are actually negative.
- FN = false negatives: number of examples predicted negative that are actually positive.

IV. EXPERIMENTAL RESULTS

Here the classification has been performed using Novel naïve algorithm on train and test soil dataset .Here weka 3.7 was used and calculated accuracy concerning with correctly and incorrectly classified instances produced with confusion matrix. This algorithm has given the correct clasification.

4.1 Contribution 1:

Naïve Confusion Matrix for Train Set

Naïve is applied on the train data set and generated for class suitability which is having three possible values which are unsuitable, highly suitable and moderately suitable. Using confusion matrix remaining accuracy performance were measured using the below given equation. Fig 3 show the Visualization of Parameters of train Experimental Data

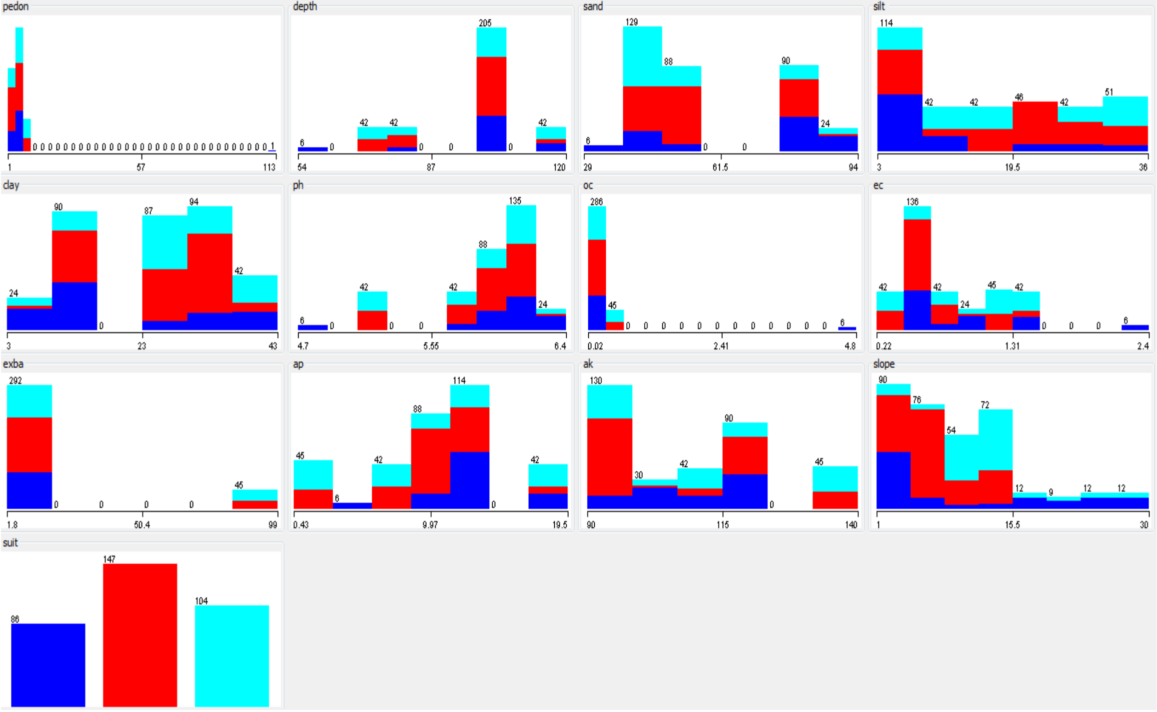


Fig. 3 Visualization of Parameters of train set Experimental Data

Confusion Matrix

```
  a          b         c       <-- classified as
 58          14       14 |     a = unsuitable
 17          105   25 |        b = highly suitable
 21           32   51 |        c = moderately suitable
```

Here finding the accuracy using below given equation i.e

- True Positive Rate = diagonal element/ sum of relevant row.
- False Positive Rate = non-diagonal element/ sum of relevant row.
- Precision = diagonal element/ sum of relevant column.
- F-measure = 2*precision*recall/ (precision+recall).

The above confusion matrix for train set:

True positive for class a = unsuitable is 58, false positive 28.
True positive for class b = highly suitable is 105, false positive 42.
True positive for class c = moderately suitable is 51, false positive 53.

Therefore the diagonal element of the matrix 58+105+51 = 214 stands for the correctly classified instances and the other elements are 17+21+14+32+14+25 = 123 represents the incorrectly classified instances.

The various accuracy measures are given below:

True positive rate for :

Class a (unsuitable) = 58/(58+14+14) = 0.674.
Class b (highly suitable) = 105/(105+17+25) = 0.714.
Class c (moderately suitable) = 51/(51+21+32) = 0.490.

False positive rate for:

Class a (unsuitable) = 38/(38+105+25+32+51) = 0.151.
Class b (highly suitable) = 46/(46+58+14+21+51) = 0.242.
Class c (moderately suitable) = 39/ (39+58+14+17+105) = 0.167.

Precision for:

Class a (unsuitable) = 58/ (58+17+21) = 0.604.
Class b (highly suitable) = 105/ (105+14+32) = 0.695.
Class c (moderately suitable) = 51/ (51+14+25) = 0.567.

F-measure for:

Class a (unsuitable)	= 2 * 0.604 *0.674/(0.604+0.674)	= 0.814/1.278 = 0.637.
Class b (highly suitable)	= 2 * 0.695 *0.714/ (0.695+0.714)	= 0.992/1.409 = 0.704.
Class c (moderately suitable)	= 2 * 0.567 * 0.49 / (0.567 + 0.49)	= 0.556 / 1.057 = 0.526.

After classification of the accuracy which are shown in the tabular form, which are given below in the table 1 and 2.

Table 1 Classification comparison on train and Test set of Novel Naive Classifier

Novel Naive	Correctly Classified Instances	Incorrectly Classified Instances
Train Set	214 (63.50%)	123 (36.49%)
Test Set	71 (63.39%)	41 (36.60%)

Table II Accuracy measure of train set of Novel Naive Classifier

Novel Naïve (Train set Class)	TP Rate	FP Rate	Precision	Recall	F- Measure
a = unsuitable	0.674	0.151	0.604	0.674	0.637
b= highly suitable	0.714	0.242	0.695	0.714	0.705
b=moderately suitable	0.490	0.167	0.567	0.49	0.526
Weighted Avg.	0.635	0.196	0.632	0.635	0.632

4.2 Contribution 2:

Naïve Confusion Matrix for Test Set

Here test data set were used and generated confusion matrix and using this remaining accuracy performance were measured using the above given equation. Figure 6.4 show the Visualization of Parameters of test set Experimental Data

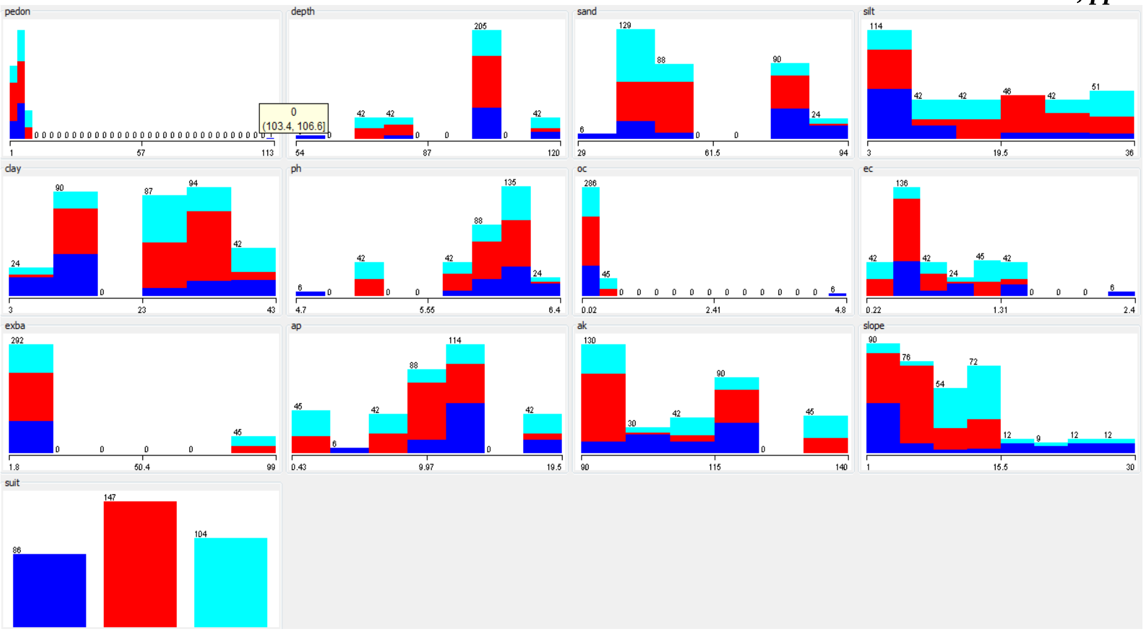


Fig. 4 Visualization of Parameters of test Experimental Data

```
 a        b     c                    <-- classified as
12        2     7 |                   a = unsuitable
 5       42     6 |                  b = highly suitable
15        6    17 |                  c = moderately suitable
```

The various accuracy measures are given in the below:

True positive rate for :
Class a (unsuitable) = 12/(12+2+7) = 0.571.
Class b (highly suitable) = 42/ (42+5+6) = 0.792.
Class c (moderately suitable) = 17/ (17+15+6) = 0.447.

False positive rate for:
Class a (unsuitable) = 20/ (20+42+6+6+17) = 0.22.
Class b (highly suitable) = 8/ (8+12+7+15+17) = 0.136.
Class c (moderately suitable) = 13/ (13+12+2+5+42) = 0.176.

Precision for:
Class a (unsuitable) = 12 / (12+5+15) = 0.375
Class b (highly suitable) = 42 / (42+2+6) = 0.84
Class c (moderately suitable) = 17 / (17+7+6) = 0.567

F-measure for :
Class a (unsuitable) = 2 * 0.375 * 0.571/ (0.375 + 0.571).
= 0.428/0.946 = 0.452
Class b (highly suitable) = 2 * 0.84 * 0.792 / (0.84+0.792)
= 1.331/1.632 = 0.816
Class c (moderately suitable) = 2 * 0.567 * 0.447 / (0.567 + 0.447)
= 0.556 / 1.014 = 0.548

The above accuracy measures shown in the below given tabular form in table 6.3.

Table III Accuracy measure of test set of Novel Naive Classifier

Novel Naïve (Test set Class)	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F- Measure**
a = unsuitable	0.571	0.22	0.375	0.571	0.453
b= highly suitable	0.792	0.136	0.84	0.792	0.816
b=moderately suitable	0.447	0.176	0.567	0.447	0.5
Weighted Avg.	0.634	0.165	0.66	0.634	0.64

© 2016, IJARCSSE All Rights Reserved

V. CONCLUSION

Novel algorithm applied in the both of the train set and test set. From the above experimental work we can conclude that correctly classified instances generated by train are 214 (63.50%), test 71 (63.39%) and incorrectly classified of train are 123 (36.49%), test are 41 (36.60%) as well as accuracy evaluation. The most outstanding result is the performance of both train and test set is good. But the perfect classification results achieved on the training set.

REFERENCES

[1] P. Andreeva, M. Dimitrova, and P. Radeva, “Data Mining Learning Models and Algorithms for Medical Application”, Proceedings of the 18th Conference on Saer, Pp 11-18, 2004.

[2] S.J. Cunningham and G. Holmes, The Proceedings of the Southeast Asia regional computer confederation conference, 1999.

[3] F.Wang and Y.Zhang, Ad Hoc and Sensor Networks, chapter A Survey on TCP over Mobile Ad-Hoc Networks, Nova Science Publishers, Pp. 267-281, 2005.

[4] S. Muggleton, Inductive Acquisition of Expert Knowledge, Addison-Wesley, Reading, Mass, USA, 1990.

[5] Ali Shawkat, and Kate A. Smith. "On learning algorithm selection for classification." Applied Soft Computing 6.2, pp. 119-138, 2006.

[6] Gaya Buddhinath and Damien Derry, “A Simple Enhancement to One Rule Classification.” Department of Computer Science & Software Engineering University of Melbourne, Australia, 2006.

[7] S.Vijayaran and Sudha, “An Effective Classification Rule Technique for Heart Disease Prediction”, International Journal of Engineering Associates (IJEA), Vol.1, Issue 4, pp.81-85, 2013.

[8] Tina R. Patil and S.S. Sherekar, “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification”, International Journal Of Computer Science And Applications, Vol. 6, No.2, 2013.

[9] Wenke Lee, Salvatore J. Stolfo, Kui W. Mok, “A Data Mining Framework for Building Intrusion Detection Models”.

[10] H. Kaur, “Actionable Rules: Issues and New Directions,” World Academy of Science, Engineering and Technology, Vol. 5, 2005, pp. 61-64.

[11] Xiang yang Li, Nong Ye, “A Supervised Clustering and Classification Algorithm for Mining Data With Mixed Variables”, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, Vol. 36, No. 2, 2006, pp. 396-406.