Chapter 5
Performance Analysis of Various Machine Learning Classifiers on Diverse Datasets

Y. Jahnavi, V. Lokeswara Reddy, P. Nagendra Kumar, N. Sri Sishvik, and M. Srinivasa Prasad

Abstract Machine learning is used to analyze data from different perspectives, summarize it into useful information, and use that information to predict the likelihood of future events. Classification is one of the main problems in the field of machine learning. The aim here is to study various classification algorithms in machine learning applied on different kinds of datasets. The algorithms used for this analysis are J48, Naive Bayes, multilayer perceptron, and ZeroR. The performance is analyzed using various metrics such as true positive rate, false positive rate, and error rates such as root mean squared error and mean absolute error. The performance of J48 algorithm is better than other algorithms for large datasets. The proposed algorithm still increases the performance in terms of error rates for large datasets. The contemplated algorithm is eventuated by mutating the splitting paradigm in the tree-based algorithms. The experimental analysis demonstrates that the proposed algorithm has reduced error rate as compared with the traditional J48 algorithm.

Y. Jahnavi (✉)
Department of Computer Science, Dr. V S Krishna Govt Degree and PG College (Autonomous), Andhra University TDR-HUB, Visakhapatnam, Andhra Pradesh, India
e-mail: yjahnavi.2011@gmail.com

V. Lokeswara Reddy
Department of Computer Science and Engineering, K.S.R.M College of Engineering (Autonomous), Kadapa, Y. S. R (Dt), Andhra Pradesh, India

P. Nagendra Kumar
Department of Computer Science and Engineering, Geethanjali Institute of Science and Technology, Nellore, Andhra Pradesh, India

N. Sri Sishvik
Department of Computer Science and Engineering, Vellore Institute of Technology, Kelambakkam-Vandalur Road, Chennai, Tamil Nadu, India

M. Srinivasa Prasad
Department of Library Science, Dr. V S Krishna Govt Degree and PG College (Autonomous), Visakhapatnam, Andhra Pradesh, India

P. K. Jha et al. (eds.), *Proceedings of Congress on Control, Robotics, and Mechatronics*, Smart Innovation, Systems and Technologies 364,
https://doi.org/10.1007/978-981-99-5180-2_5

Keywords Machine learning · Classification · Tree based classifier · Splitting criteria

5.1 Introduction and Preliminaries

During the past several years, investigation has been concentrated on diverse groups based on the machine learning algorithms due to the extreme require of accurate prophecies. Machine learning is not about giving tight rules by analyzing the datasets rather it is used to predict the likelihood of future events with some certainty. Classification is a machine learning approach used to fore tell cluster association for documents illustration and is a widely used technique in various fields [1].

Machine learning, pervasive computing, statistical analysis, data analytics, etc., are the applications of artificial intelligence (AI), whereas machine learning allows training and strengthens from practice to estimate the eventualities [2–5, 5–8].

A well-known test sample label is correlated with a separate result from the model. The extent of precision of the proportion of instances of the test set is grouped consequently by the framework. If precision is tolerable, then this model is used to separate tuples of data class labels, which are unknown [9, 10].

5.2 Literature Work and Methodologies

Classification has been considered as a seminal issue in the area of machine learning [11]. All the time, there has been absolutely a number of enormous surveys on classification algorithms [12, 13], performance evaluation [14–16], collations, and assessment of various classification algorithms [2, 17] beside their uses in figuring out real-life problems in the applications of business [9, 18–21], engineering, medicine [1, 22, 23], etc.

Amudha and Abdul Rauf [24, 25] applied data mining techniques as an approach for intrusion detection to identify whether the deviation from normal usage patterns can be flagged as intrusions and performed a correlative investigation of various classification algorithms.

Voznika and Viana [26, 27], described different approximation algorithms such as statistical algorithms, genetic programming, neural networks and concluded that the best model can be found by trial and error trying different algorithms in order to obtain the best results possible.

Kesavaraj, Sukumaran [13] performed investigation on multifold categorization methods to furnish an exhaustive analysis of machine learning algorithms.

Chintan Shah and Anjali Jeevani [17] compared decision tree, K-nearest neighbor, Naive Bayesian using parameters like correctly classified illustrations, time taken, relative absolute error, kappa statistic, and root relative absolute error on breast cancer dataset.

Dogan and Tanrikulu [18], performed a study that collate and contrast the precision of the classification algorithms. The application of certain classification models on multiple datasets is done in three stages. The research addressed the reliability of the classifiers, studied by demonstration on various datasets.

Rutvija and Pandya [28] performed the extensive analysis on various categorization techniques.

Keerthana [29] focused on image classification approach in order to identify better algorithm for medical image classification.

Classification is an approach of grouping or allocating class labels to a pattern set under the direction of an instructor. Classification is also termed as supervised learning. The patterns are primarily segregated into training and test sets. Training set is used to prepare the classifier, and the test set is prone to estimate the precision of a classifier. The classifiers are categorized into tree-based, rule-based, Bayes, functions, etc. The algorithms that have been chosen for this predictive data mining task include J48 from trees, multilayer perceptron from functions, ZeroR from rules, and Naive Bayes from Bayes.

The most popular supervised classifier which can work well on noisy data is decision tree classifier. There are various other types of classifiers such as Bayesian classification, neural network-based classifier, and support vector machine. A great deal of research has been done for developing efficient methods in the field of machine learning.

The J48 algorithm uses information gain and gain ratios to construct the decision tree for a given dataset. It works by recursively dividing the data on a single attribute, according to the information gain calculated. Each split in the tree represents a node where a decision must be taken, and you go to the following node and the next till you reach the leaf that expresses you the predicted output.

The steps in the J48 algorithm are as follows:

(i) If the requirements are the identical group, the tree illustrates the leaf so that the leaf is substituted by designating in the identical class.
(ii) The feasible information is intended for every characteristic, determined by a check on the attribute. Then, the gain in information is premeditated that would outcome from a examination on the characteristic (attribute).
(iii) Then the best characteristic (attribute) is identified on the foundation of the current selection criterion and that attribute is adopted for ramification.

5.3 Information Gain and Gain Ratio

The information gain is based on the entropy after a dataset is split on an attribute, where entropy is used to estimate the similarity of a sample. If the instance is completely identical, the entropy is zero, and if the instance is evenly separated, it has entropy of one.

The entropy is calculated using the following formula.

$$\text{entropy}(p1, p2, \ldots, pn) = -p1 \log p1 - p2 \log p2 - \cdots - pn \log pn \quad (5.1)$$

$$\text{entropy}(p1, p2, \ldots, pn) = -\sum pi \log pi \quad (5.2)$$

Entropy on the other hand is an estimate of *impurity*. It is characterized for a binary class with estimates a and b as:

$$\text{entropy} = -p(a) * \log(P(a)) - p(b) * \log(p(b)) \quad (5.3)$$

Using the above formula, we calculate two entropy values, namely entropy before and entropy after. The entropy before value is calculated before splitting, and entropy after is computed after considering the split. Now by assimilating the entropy before and after the split, we derive an estimate of information gain as denoted below:

$$\text{Information gain} = \text{entropy before} - \text{entropy after} \quad (5.4)$$

At each node of the tree, this computation is carried out for every feature, and the feature with the largest information gain is chosen for the split in a greedy manner. This process is applied recursively from the root-node down and stops when a leaf node contains instances all having the same class, i.e., it stops when the node cannot be divided further. Constructing a decision tree is all about finding attribute that has the highest information gain.

Gain ratio is a modification of the information gain that reduces its bias. It takes into account the number and size of branches while choosing an attribute. There are chances of getting negative values in the existing information gain and gain ratio algorithms.

The idea of the proposed algorithm is to eliminate the negative values. The accuracy of the algorithm can be improved by eliminating the negative values. The proposed algorithm checks if the entropy before value is less than entropy after value and return 0; otherwise, it returns the unknown rate calculated.

Because of the outliers pruning is a significant step to the result. Some instances are present in all datasets which are not well defined and differ from the other instances on its neighborhood.

The classification is performed on the instances of the training set, and tree is formed. There exist various algorithms for performing classification, extracting salient features, opinion mining, processing of scalable web log data using map reduce framework etc. [30–37]. The pruning is performed for decreasing classification errors which are being produced by specialization in the training set. Pruning is performed for the generalization of the tree.

5.4 Results and Discussion

Evaluation of the datasets has been done by using the proposed classification algorithm. Various classification algorithms analyzed are evaluated by the evaluation criteria such as true positive rate, false positive rate, mean absolute error, and root mean square error. Heart disease, mushrooms, and birds are the datasets used for the analysis. These two datasets are taken from UCI machine learning repository. The classification algorithms are applied on the data using tenfold cross validation technique, and the results are then recorded. A sample description of datasets has been represented in Table 5.1.

The considered sample heart disease dataset has 14 attributes and 303 instances that are categorized into 5 classes. Mushrooms dataset has 23 attributes and 8124 instances that are categorized into 2 classes. Experimentation has been done on each dataset.

5.4.1 Data Set 1 (Heart Disease Dataset)

The considered sample heart disease dataset has 14 attributes and 303 instances that have only 5 classes. This dataset has taken from UCI machine learning repository.

True positive rate, false positive rate, root mean square error, and mean absolute error are calculated for J48, Naive Bayesian, multilayer perceptron, ZeroR, and the proposed algorithm, which are represented in Table 5.2.

True positive rate of J48, Naive Bayesian, multilayer perceptron, ZeroR, and the proposed algorithm are 0.558, 0.559, 0.574, 0.541, and 0.558, respectively. It shows that the proposed modified algorithm is able to show the same performance as J48 algorithm.

Table 5.1 Sample description of datasets

Dataset	Attributes	Instances	Classes
Heart disease	14	303	5
Mushrooms	23	8124	2

Table 5.2 Results of the classification algorithms on heart disease dataset

	TP	FP	Mean absolute error	Root mean squared error
J48	0.558	0.238	0.2	0.3867
NB	0.559	0.273	0.1838	0.3368
MLP	0.574	0.189	0.2768	0.378
ZERO R	0.541	0.541	0.2591	0.3592
Modified algorithm	0.558	0.238	0.2	0.3367

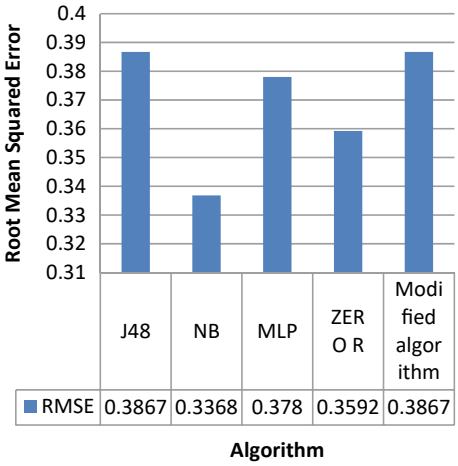

Fig. 5.1 Comparison of root mean squared error of classifiers on heart disease dataset

False positive rate of J48, Naive Bayesian, Multilayer Perceptron, ZeroR, and the proposed algorithm are 0.238, 0.273, 0.189, 0.541, and 0.238, respectively. It shows that the proposed modified algorithm is able to show the same performance as J48 algorithm.

Mean absolute error of J48, Naive Bayesian, multilayer perceptron, ZeroR, and the proposed algorithm are 0.2, 0.1838, 0.2768, 0.2591, and 0.2, respectively. It shows that J48 and the proposed modified algorithm are better in terms of mean absolute error, compared with ZeroR and multilayer perceptron. But for this dataset Naive Bayesian performs better by exhibiting low mean absolute error, i.e., 0.1838.

Root mean squared error of J48, Naive Bayesian, multilayer perceptron, ZeroR, and the proposed algorithm are 0.3867, 0.3368, 0.378, 0.3592, and 0.3367, respectively. It shows that the proposed modified algorithm is better in terms of root mean squared error, compared with the other algorithms, i.e., 0.3367.

For the considered dataset, the proposed modified algorithm shows better performance than other algorithms in terms of the root mean square error. The root mean square error values of various algorithms are pictorially represented in Fig. 5.1 on heart disease dataset.

5.4.2 Data Set 2 (Birds Dataset)

Birds dataset used here is an images dataset. It contains 600 images of 34 samples of 6 types of birds. We applied some filters before classifying the data. This dataset has taken from a Ponce research group repository. True positive rate, false positive rate, root mean square error, and mean absolute error are calculated for J48, Naive Bayesian, multilayer perceptron, ZeroR, and the proposed algorithm, which are represented in Table 5.3.

Table 5.3 Results of the classification algorithms on a birds dataset

	TP	FP	Mean absolute error	Root mean squared error
J48	0.372	0.126	0.213	0.2536
NB	0.325	0.095	0.2726	0.3421
MLP	0.390	0.102	0.2774	0.3784
ZERO R	0.165	0.169	0.2778	0.3727
Modified algorithm	0.372	0.026	0.213	0.2436

True positive rate of J48, Naive Bayesian, multilayer perceptron, ZeroR, and the proposed algorithm are 0.372, 0.325, 0.390, 0.165, and 0.372, respectively. It shows that the proposed modified algorithm is able to show the same performance as J48 algorithm.

False positive rate of J48, Naive Bayesian, multilayer perceptron, ZeroR, and the proposed algorithm are 0.126, 0.095, 0.102, 0.169, and 0.026, respectively. It shows that the proposed modified algorithm is able to show better performance compared with all the considered algorithms.

Mean absolute error of J48, Naive Bayesian, multilayer perceptron, ZeroR, and the proposed algorithm are 0.213, 0.2726, 0.2774, 0.2778, and 0.213, respectively. It shows that J48 and the proposed modified algorithm is better for the considered dataset in terms of mean absolute error, compared with Naive Bayesian, ZeroR, and multilayer perceptron.

Root mean squared error of J48, Naive Bayesian, multilayer perceptron, ZeroR, and the proposed algorithm are 0.2536, 0.3421, 0.3784, 0.3727, and 0.2436, respectively. It shows that the proposed modified algorithm is better in terms of root mean squared error, compared with the other algorithms, i.e., 0.2436.

For the considered dataset, the proposed modified algorithm shows better performance than other algorithms in terms of the root mean square error. The root mean square error values of various algorithms are pictorially represented in Fig. 5.2 on birds dataset.

The experimentation has shown that the proposed algorithm outperforms other existing algorithms on various considered datasets.

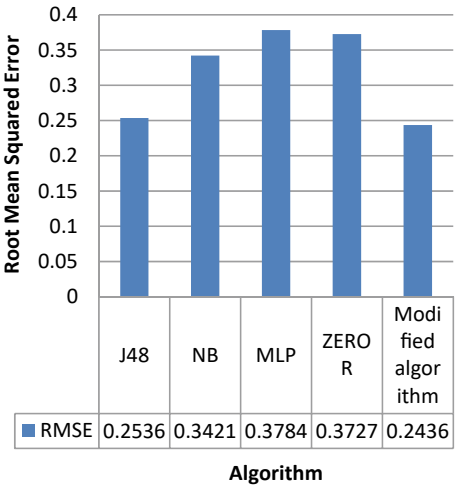

Fig. 5.2 Comparison of root mean squared error of classifiers on a bird's dataset

5.5 Conclusion

This study focuses on finding the right algorithm for classification of diverse datasets. The datasets that used are heart disease and mushrooms. For mushrooms dataset, the decision tree algorithm J48 and the proposed modified algorithm gave better results in terms of root mean squared error (RMSE) and for heart diseases dataset; the proposed modified algorithm gave reduced error rates than the traditional J48 algorithm, Naïve Bayes, multilayer perceptron, and ZeroR. However, it is noticed that the performance of a classifier depends on the dataset used. Although there are many algorithms available, the best one is often found by trial and error. For better results, one must compare or even combine the available algorithms. The performance of the existing algorithms can even be improved with minor modifications.

References

1. Vandehei, B., et al.: Leveraging the defects life cycle to label affected versions and defective classes. ACM Trans. Softw. Eng. Methodol. **30**, 2 (1–35) (2021), Online publication date: 1-Mar-2021
2. Jahnavi, Y., Radhika, Y.: A cogitate study on text mining. Int. J. Eng. Adv. Technol. **1**(6), 189–196 (2012)
3. Jahnavi, Y., Radhika, Y.: Hot topic extraction based on frequency, position, scattering and topical weight for time sliced news documents. In: 15th International Conference on Advanced Computing Technologies, ICACT 2013
4. Jahnavi, Y.: Statistical data mining technique for salient feature extraction. Int. J. Intell. Syst. Technol. Appl. (Inderscience Publishers), **18**(4) (2019)
5. Jahnavi, Y., et al.: A novel ensemble stacking classification of genetic variations using machine learning algorithms. Int. J. Image Graph. 2350015(2021)

6. Jahnavi, Y., Radhika, Y.: FPST: a new term weighting algorithm for long running and short lived events'. Int. J. Data Anal. Tech. Strategies **7**(4), 366–383 (2015)
7. Jahnavi, Y.: A new algorithm for time series prediction using machine learning models. Evol. Intell. (Springer), Accepted, 2022
8. Jahnavi, Y.: Analysis of weather data using various regression algorithms. Int. J. Data Sci. (Inderscience Publishers), **4**(2) (2019)
9. Parneetkaur, et al.: Classification and prediction based data mining algorithms to predict slow learns in education sector. In: 3rd International Conference on Recent Trends in Computing, 2015
10. Shiqun, et al.: Research and implementation of classification algorithm on web text mining. In: IEEE International Conference on Semantics Knowledge and Grid, 2007, pp. 446–449
11. Singh, P., et al.: Machine learning: a comprehensive survey on existing algorithms (2021)
12. Gulati, H.: Predictive analytics using data mining technique. In: 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 713–716
13. Binjubeir, M., et al.: Comprehensive survey on Big Data privacy protection. Access IEEE **8**, 20067–20079 (2020)
14. Kumar, et al.: Performance evaluation of decision tree versus artificial neural network based classifiers in diversity of datasets. In: 2011 World Congress on Information and Communication Technologies, 2011, pp. 798–803. https://doi.org/10.1109/WICT.2011.6141349
15. Rjeily, B., Andres, E., et al.: Medical data mining for heart diseases and the future of sequential mining in medical field. In: Tsihrintzis, G., Sotiropoulos, D., Jain, L. (eds.) Machine Learning Paradigms. Intelligent Systems Reference Library, vol 149. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-94030-4_4.
16. Sharma, S., et al.: Machine learning techniques for data mining: a survey. In: IEEE National Conference on Computational Intelligence and Computing Research (ICCIC), 2013
17. Panda, M., et al.: International proceedings on advances in soft computing. Intell. Syst. Appl. **628**, 363 (2018)
18. Dogan, W., Tanrikulu, Z.: A comparitive analysis of classification algorithms in data mining for accuracy, speed & robustness. Springer **14**(2), 105–124 (2013)
19. Panchal, G., et al.: Performance analysis of classification techniques using different parameters. In: Second International Conference, ICDEM, 2010
20. Tamizharasi, K., UmaRani, Performance analysis of various data mining algorithms. Int. J. Comput. Commun. Inf. Syst. (IJCCIS) **6**(3) (2014)
21. Korting, T.S.: "C4.5 algorithm and multivariate decision trees", image processing division. In: National Institute for Space Research--INPE
22. Ian, et al.: Data Mining Practical Machine Learning Tools and Techniques, 3rd edn. Elsevier (2011)
23. Zuveda, Comparison of the performance of several data mining methods for bad debt recovery in the health care industry. J. Appl. Bus. Res. **21** (2005)
24. Gayatri, N., et al.: Performance analysis of data mining algorithms for software quality prediction. In: International Conference on Advances in Recent Technologies in Communication and Computing, ARTCom'09, India, October 2009
25. Amudha, P., Abdul Rauf, H.: Performance analysis of data mining approaches in intrusion detection. In: PACC, International Conference, 2011
26. Voznika, F., Viana, L.: Data Mining Classification. Springer (2001)
27. Peter, T.J., Somasundaram, K.: An empirical study on prediction of heart disease using classification data mining techniques. In: IEEE-International Conference on Advances in Engineering Service and Management, 2012
28. Pandya, R., Pandya, J.: C5.0 Algorithm to improved decision tree with feature selection and reduced error pruning. Int. J. Comput. Appl. **117**(16) (2015), (0975-8887)
29. Keerthana, P., et al.: Performance analysis of data mining algorithms for medical image classification. Int. J. Comput. Sci. Mob. Comput. **5**(3) (2016)
30. Jahnavi, Y.: A New Term Weighting Algorithm for Identifying Salient Events. LAP LAMBERT Academic Publishing (2018)

31. Tiwari, V., et al.: Applications of the Internet of Things in healthcare: a review. Turk. J. Comput. Math. Educ. **12**(12), 2883–2890 (2021)
32. Haripriya, et al.: Using social media to promote E-commerce business. Int. J. Recent Res. Aspects **5**(1), 211–214 (2018)
33. Sukanya, et al.: Country location classification on Tweets. Indian J. Public Health Res. Dev. **10**(5) (2019)
34. Vijaya, et al.: Community-based health service for Lexis Gap in online health seekers
35. Bhargav, et al.: An extensive study for the development of web pages. Indian J. Public Health Res. Dev. **10**(5) (2019)
36. Srivani, et al.: An approach for opinion mining by acumening the data through exerting the insights
37. Jahnavi, Y., et al.: A novel processing of scalable web log data using map reduce framework. In: Proceedings of CVR 2022, Computer Vision and Robotics, ISBN: 978-981-19-7891-3