IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"Comparison Analysis Of Classification Algorithms To Predict Customers' Awareness About Net Banking Security And Usefulness"

- Authors: Ms. Dhara Joshi[1], Dr. Shreya Shah[2],
- Affiliations: [1]Research Scholar, [2]Assistant Professor
[1]CVM University, V.V. Nagar Gujarat, India.

Abstract: Decision Tree is the most prominent classification algorithm for supervised machine learning. Various decision tree algorithms, such as J48, Random Tree, LMT, and Random Forest, are applied to the datasets depending on the types of data and the application requirements. This paper presents a comparison analysis of J48, Random Forest, and LMT algorithms based on multiple evaluation parameters. The results demonstrate that J48 consistently outperforms other algorithms in terms of accuracy and robustness, compared to algorithms such as Random Forest and LMT.

Keywords:
Supervised machine learning, J48, Random Forest, LMT, Classification Algorithm Comparison.

Abbreviations and Acronyms LMT: Logistic Model Tree

I. INTRODUCTION

II. Classification of data is a crucial task. The accuracy of results, efficiency, and error rate are the most important parameters in the selection of a classification algorithm for the proposed research. For supervised learning, the algorithm is selected based on a comparison of errors, where the misclassification error is the sum of false positives and false negatives [1]. Along with the error ratio and other parameters, the relative training time is equally important to consider. In recent years, like other fields, banking has also started to utilise data mining to improve its overall performance for its customers. For example, customer query answering, common question answering, e-learning, etc., are analysed through various data mining techniques. Basically, in data mining, data are extracted, utilised, and then processed for future needs. Although various data mining techniques are available, classification is the most widely used [2]. Classification comes with a variety of techniques, such as decision trees, neural networks, genetic algorithms, and k-nearest neighbours. The decision tree is one of the most popular techniques used for classification and applies to data whose values are known precisely. Some common algorithms for decision tree construction are J48, Simple Cart, ID3, Random Tree, and Random Forest etc.

III. This paper presents a comparative study of various classification algorithms, specifically focusing on decision tree algorithms. Each decision tree algorithm utilises a unique process for record classification, leading to distinct classification results that are independent of other algorithms.

IV. Selecting the most suitable algorithm is challenging, as the performance and accuracy of a classification algorithm are influenced by several factors. These include the nature of the data, the presence of noise, the data retrieval environment, and the training time allocated to the algorithm. Our study aims to provide insights into these variations.

Data Mining Overview

Data mining is the process of discovering knowledge, patterns, and correlations within large datasets using a combination of machine learning, statistics, and database systems. It takes raw data as input and generates information by extracting data from huge datasets. It is also known as knowledge mining. It transforms raw data into actionable insights that can be used for informed decision-making, predictive modelling, and understanding complex phenomena.

The key techniques used in data mining are clustering, classification, and regression. Classification is a supervised machine learning technique that builds a model based on previous datasets to predict future trends with new datasets and categorise data into predefined classes or groups based on their characteristics.

Classification

Classification involves building a model from a labelled dataset by applying a classification algorithm. The model is then used to test accuracy. Data mining classification is a process that involves the analysis of data to identify patterns and relationships. The objective of classification is to build a model that can be used to predict the class or category of new data instances based on their attributes or features [7]. Classification is a powerful technique for predicting the class or category of new data instances in data mining. The process involves several steps, including data pre-processing, feature selection, model selection, model training, and model tuning.

Decision Tree Induction Process

Using a training dataset of cases with attributes and class labels, the decision tree induction procedure builds a predictive model in the shape of a decision tree. This procedure is usually carried out in a divide-and-conquer, recursive, top-down fashion. Classification facilitates informed decision-making across several domains and allows precise forecasting of future trends for new data by identifying patterns in the old data.

The decision tree has a variety of implications in the process of data mining as a powerful predictive model. It is used widely in areas like statistics, machine learning, data mining, and artificial intelligence. As a prediction model, it maps observations about an item to predict the item's target value. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision-making.

DECISION TREE ALGORITHM

After selecting a dataset, three classification algorithms, namely Random Forest, J48, and LMT are were implemented. After implementation, the report containing the accuracy has been generated for all the algorithms. The process has been done using both the Weka tool and different Python libraries in Jupyter Notebook. Each algorithm's effectiveness and efficacy are analysed and compared from all aspects.

A. Random Forest

The Random Forest algorithm is a supervised machine learning algorithm that works on both classification and regression problems and is known for its accuracy, versatility, and robustness. It operates using a technology called ensemble learning, specifically bootstrap aggregation. As a flexible and powerful machine learning algorithm, it can handle complex problems, manage large datasets, and deliver accurate predictions. This feature makes it valuable for data science in various fields. The Algorithm builds multiple decision trees on different, randomly sampled subsets of the original training data.

B. J48

The J48 algorithm is one of the most widely used machine learning algorithms developed by Ross Quinlan, used for examining data categorically and continuously. It has the ability to work with both categorical and continuous-valued attributes. The machine learning process has two main phases: a learning phase, where the classification algorithm is trained, and a classification phase, where the algorithm labels new data. Classification is a data mining task that maps data into predefined groups and classes, also known as supervised learning. The algorithm uses a greedy technique to induce decision trees for classification and uses reduced-error pruning. The attribute to be predicted is known as the dependent variable, as its value is dependent on the values of other independent attributes. Therefore, the attributes that help in predicting the value of the dependent variable are known as the independent variables in the dataset [9].

C. LMT

The full name of LMT is the Logistic Model Tree. It is a classification model within a decision tree with an associated supervised training algorithm. The basic LMT induction algorithm uses cross-validation to find the number of logic boost iterations. The LMT algorithm is a classification model that combines the strengths of two distinct machine learning techniques: logistic regression and decision trees. LMT aims for both high predictive accuracy and interpretability by combining the two algorithms in a single tree. A leaf model is constructed by adding a partial linear model to the inherited model from its ancestral nodes while the tree is grown [11].

RESULT AND EFFICIENCY COMPARISON BETWEEN ALGORITHM*

The data are collected from one of my previous collection of data set for research work. It is used for the comparison purpose. The collected data is converted into attribute relation file format (ARFF). The results demonstrate the overall performance of all the algorithms.

Following is the list of attributes involved in the data mining process.

Attribute

@attribute Gender {Male, Female}
@attribute Category {Cooperative, Public, Private}
@attribute Education Level {PG, Graduate, UG}
@attribute Use of frequency {Daily, Weekly, Monthly, Rarely}
@attribute Convenience {neutral, Convenient, in Convenient}
@attribute Satisfaction level for security features {Satisfied, Dissatisfied, Neutral}
@attribute Satisfaction with the Reliability and Security of net banking Authentication {Satisfied, Dissatisfied, Neutral}
@attribute Concerned about the potential risks associated {Concerned, Very Concerned, Slight Concerned}
@attribute Impacts of digital banking on financial inclusion in India {Agree, Neutral, Dis Agree}
@attribute age {Below 20, 20 – 34, 35 – 44, 45 – 54, 55 and above}

Result of Algorithm:

Table 1 : Results of the Algorithm

J 48	Instances	152
	Test Mode	10-fold cross validation
	Number of leaves	21
	Size of tree	30
	Correctly classified instances	97 63.8158
	Incorrectly classified instances	55 36.1842
	Kappa Statistic	0.3182
	Confusion Matrix	a b c d e <-- classified as 26 19 0 1 0 | a = 35 - 44 12 71 0 1 0 | b = 20 - 34 3 4 0 0 0 | c = 55 and above 6 8 0 0 0 | d = 45 -54 0 1 0 0 0 | e = Below 20
	Time taken to build model	0 Seconds
LMT	Instances	152
	Test Mode	10-fold cross validation
	Number of leaves	1
	Size of tree	1
	Correctly classified instances	89 58.5526
	Incorrectly classified instances	63 41.4487
	Kappa Statistic	0.2384
	Confusion Matrix	a b c d e <-- classified as 24 21 0 1 0 | a = 35 - 44 16 65 1 1 1 | b = 20 - 34 4 3 0 0 0 | c = 55 and above 7 7 0 0 0 | d = 45 -54 0 1 0 0 0 | e = Below 20
	Time taken to build model	0.03 Seconds
Random Forest	Instances	152
	Test Mode	10-fold cross validation
	Number of leaves	0
	Size of tree	0
	Correctly classified instances	77 50.6579
	Incorrectly classified instances	75 49.3421
	Kappa Statistic	0.115
	Confusion Matrix	a b c d e <-- classified as 15 24 3 4 0 | a = 35 - 44 17 62 0 4 1 | b = 20 - 34 6 1 0 0 0 | c = 55 and above 6 8 0 0 0 | d = 45 -54 0 1 0 0 0 | e = Below 20
	Time taken to build model	0.03 Seconds

Tester: weka.experiment.Paired Corrected TTester
Analysing: Percent_correct
Datasets: 1
Result sets: 3
Confidence: 0.05 (two tailed)
Sorted by: -
Date: 16/08/25, 7:21 pm

Dataset	(1) trees.Si |	(2) trees	(3) trees
'Respondent'	(152) 61.29 |	56.75	48.75
	(v/ /*) |	(0/1/0)	(0/1/0)

Key:
(1) trees.Random Forest '-K 0 -M 1.0 -V 0.001 -S 1' -9051119597407396024
(2) trees. J48 '-C 0.25 -M 2' -217733168393644444
(3) trees. LMT '-I -1 -M 15 -W 0.0' -1113212459618104943

CONCLUSIONS

Here in this paper, we compare 3 classification algorithms to choose the best algorithm for the prediction of users' net banking awareness. The compared algorithms are J48, Random Forest and LMT. The results are analyzed and demonstrate the fact that the J48 algorithm works better for the proposed system of predicting customers' awareness about net banking security and usefulness.

REFERENCES

[1] T. O. a. a. E. Olcay Taner Yildiz, "Multivariate Statistical Tests for Comparing," Springer-Verlag Berlin Heidelberg, pp. 1-15, 2011.
[2] S. A. K. Manpreet Singh, "Performance Analysis of Decision Trees," International Journal of Computer Applications, vol. 71, no. 19, 2013.
[3] S. D. P. R. a. L. V. N. S. Aruna, "AN EMPIRICAL COMPARISON OF SUPERVISED LEARNING ALGORITHMS IN DISEASE DETECTION," International Journal of Information Technology Convergence and Services, vol. 1, no. 4, pp. 81-92, 2011.
[4] D. S. D. S. My Chau Tu, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms," in Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009.
[5] S. S. Aman Kumar Sharma, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis," International Journal on Computer Science and Engineering, vol. 3, no. 5, pp. 1890-1895, 2011.
[6] R. S. B. R. R. Kabra, "Performance Prediction of Engineering Student Using Decision Tree," International Journal of Computer Application, vol. 36, no. 11, pp. 8-12, 2011.
[7] https://mobigaurav.medium.com/introduction-to-data-mining-classification-c44c02b28dda.
[8] https://www.vldb.org/conf/1998/p404.pdf
[9] Jaymin N Undavia , "Comparison of classification Algorithms to predict Students' Post Graduation Course in Weka Environment", International Journal of Advanced Research in computer science and software Engineering. Vol. 3, Issue 9 September 2013 page 1250 - 1253.
[10] https://www.sciencedirect.com/science/article/abs/pii/S0957417417308308