Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 35 (2019) 698–703

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

2nd International Conference on Sustainable Materials Processing and Manufacturing

(SMPM 2019)

Prediction performance of improved decision tree-based algorithms: a review

Ibomoiye Domor Mienye[a], Yanxia Sun[a,*], Zenghui Wang[b]

[a]*Department of Electrical and Electronics Engineering Science, University of Johannesburg, Auckland Park Kingsway Campus, Johannesburg 2006, South Africa*

[b]*Department of Electrical and Mining Engineering, University of south Africa, Florida 1710, South Africa*

Abstract

Applications of machine learning can be found in retail, banking, education, health sectors etc. To process the large data emanating from the various sectors, researchers are developing different algorithms using expertise from several fields and knowledge of existing algorithms. Machine learning decision tree algorithms which includes ID3, C4.5, C5.0, and CART (Classification and Regression Trees) are quite powerful. ID3 and C4.5 are mostly used in classification problems, and they are the focus of this research. C4.5 is an improved version of ID3 developed by Ross Quinlan. The prediction performance of these algorithms is very important. In this paper, the prediction performance of decision tree algorithms will be studied, an in-depth review will be conducted on relevant researches that attempted to improve the performance of the algorithms and the various methods used. Comparison will also be done between the various tree based algorithms. The major contribution of this review is to provide researchers with the progress made so far, as there is no available literature that has put together relevant improvements of decision tree based algorithms, and lastly lay the foundation for future research and improvements.

Peer-review under responsibility of the organizing committee of SMPM 2019.

Keywords: Machine learning; Data mining; C4.5; ID3; Algorithm; Decision tree; Classification; Information gain ratio; entropy.

* Corresponding author. Tel.: +27115592213
E-mail address: sunyanxia@gmail.com

2351-9789
Peer-review under responsibility of the organizing committee of SMPM 2019.
10.1016/j.promfg.2019.06.011

1. Introduction

Over the years, technological advancement has changed our approach to handling data. Data mining is now being applied to various datasets to learn hidden patterns and make appropriate predictions and/or descriptions, and it refers to a collection of techniques used to extract hidden knowledge such as patterns, relationships, or rules from large data sets (Almasoud, et al, 2015). This extracted knowledge can be analyzed and is capable of predicting future trends (Mhetre and Nagar, 2017). Data mining is an important step in the knowledge discovery in databases process. Also, it is an interdisciplinary field that derives its methods from machine learning, artificial intelligence, and statistics, among others.

Applications of machine learning are found in retail, banking, military, and health sectors etc., and in order to achieve its goal, researchers are developing different algorithms using expertise from various fields of study (Neto and Castro, 2017). These algorithms can be used to build models, which acquire knowledge from previous data. It can be applied to solve problems related to classification, regression, clustering and optimization using algorithms such as Decision trees (ID3, C4.5, C5.0, CART), Support Vector Machine (SVM), Neural networks, Naïve Bayes, Linear regression, K-Nearest Neighbor (KNN) etc. Among these approaches, the ID3 (Iterative Dichotomiser 3) and C4.5 are the most used decision tree algorithms (Anuradha and Velmurugan, 2014).

Ross Quinlan developed both ID3 and C4.5 algorithms, the latter algorithm is an improved version of the former one. In ID3 decision tree only categorical type features can be used, and numerical type features cannot be applied. The improvements in C4.5 include the use of information gain ratio (IGR) rather than information gain as in ID3. Secondly, pruning can be done when constructing the tree or after construction of tree. Thirdly, C4.5 can handle attributes that contains continuous features. And it is also able to handle missing data (Adhatrao et al, 2013). However, C4.5 algorithm came with some limitations, which researchers are trying to solve. It constructs empty branches with zero values, and also over fitting occurs when the algorithm picks up data with unusual features, especially noisy data. These limitations led researchers to develop improved versions of C4.5 and also reconsider ID3 in an attempt to further improve it. These improvements gave varying degrees of accuracy, and this paper evaluates the prediction performance of these improved algorithms with particular focus on the methods used to achieve the given result.

2. Background of ID3 and C4.5 Decision Tree Algorithms

ID3 and C4.5 are information gain based algorithms developed by Ross Quinlan. ID3 algorithm constructs decision trees based on the information gain gotten from the training data, whereas C4.5 uses an additional information called gain ratio. The constructed decision tree is then used to classify test data. The dataset designed for training the decision tree serves as input to the algorithms, and it comprises of objects and various attributes. The decision tree building process for both algorithms is similar. Firstly, the class entropy and entropy of each attributes are calculated, then the information gain is calculated for all the attributes as seen in equation 1, 2, and 3 below. In ID3 algorithm the attribute with the highest information gain is considered the most informative attribute and is selected as the root node. And the process is repeated until all the attributes are in the tree.

$$info\,(D) = -\sum_{i=1}^{m} p_i * log_2\,(p_i) \tag{1}$$

$$info_A\,(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * info(D_j) \tag{2}$$

$$Gain(A) = info\,(D) - info_A\,(D) \tag{3}$$

There are some advantages in using ID3 including easy decomposability, strong intuitive nature etc. And some disadvantages include the following;

- Using information gain for feature selection, the algorithm tends to select attributes with more values, which is due to the fact that the value of the information gain of this kind of attribute will be bigger than others.
- In the decision tree building process, it is difficult to control the tree size. However, most researchers have tried to improve on this using various pruning methods to avoid the occurrence of over-fitting, which has led to the decision tree building process to be completed in two steps, that is modeling and pruning. Meanwhile, it will save a lot of time if a concise decision tree is built in onestep.
- There are several logarithmic calculations in the attribute selection process, this has made the computation of information gain time consuming.

C4.5 algorithm was developed to overcome some of the limitations of ID3. The IGR which is used in C4.5 algorithm is a less biased selection criterion, it takes the information gain and normalizes it with split entropy. In C4.5 two new parameters are calculated in addition to those of equation 1, 2, and 3, and they are the split entropy and IGR, as shown in equation 4 and 5 respectively. The IGR is used to evaluate the information value of the various attributes and to determine the best split attribute, and then the decision tree is generated with depth first strategy. In a similar way, every node of the tree initializes its information and produce their child nodes (Peng et al, 2017). The IGR is a modification of information gain to reduce feature bias towards attributes that has many branches. The gain ratio is large if the data is spread evenly and the value will be small if all data enters into one branch.

$$SplitInfo_A\ D = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * log_2 \frac{|D_j|}{|D|} \tag{4}$$

$$IGR\ (A) = \frac{Gain\,(A)}{SplitInfo_A\ D} \tag{5}$$

When the above calculations do the attribute with the highest IGR is selected as the root or split attribute. The process is repeated for every branch until all the cases in the branch have the same class. In C4.5 decision trees, calculating the best split point and best attribute are the two crucial aspects. They have high computational requirements since the dataset needs to be scanned several times. Unfortunately, this requires a large amount of serial operations which result in long execution time during the tree growth stage. Other limitations of the C4.5 is that over-fitting occurs when it faces with noisy data, and it constructs empty branches with zero values (Peng et al, 2017). From the above it can be deduced that both ID3 and C4.5 algorithms have limitations, and several researchers have attempted to improve on these algorithms using different methods. The intent of this paper is to review these improvements with particular focus on the method used and the performance obtained.

3. Literature Survey

3.1. Modifications to ID3

Yi-bin et al, 2017 proposed an improved ID3 algorithm, which merges the information entropy based on different weights with coordination degree in rough set theory. In ID3, selecting the optimal feature is based on the information gain formula (eq. 3), but the logarithm in the algorithm makes the computation complex. Their research was based on the fact that if a simpler formula can be used, the decision tree building process would be faster. In achieving this they replaced the logarithmic equation of information gain by the four arithmetic operations (i.e. addition, subtraction, multiplication, and division), thereby improving the running speed of the decision tree building process. This difficulty in information entropy computation in ID3 algorithm was reduced using the approximation formula of Maclaurin formula. The formula they deduced can be used to compute the information gain of the various attributes and the attribute with the highest information gain is selected as the root node, and so

on. They compared the traditional ID3 algorithm and their proposed version using three datasets. For the first two datasets which were small it was shown that their algorithm had better performance in terms of running time and tree structure, but the accuracy was not better than ID3. Meanwhile, for the third dataset which was large, their improved algorithm outperformed the standard ID3 in terms of running time, accuracy, and tree structure.

In ID3 algorithm, other than the known problem of bias towards multi-value attributes, it also has an issue of computational complexity. In a bid to solving thes problems, Wang et al, 2017 proposed a novel approach to select the splitting attribute. The authors used rough set to implement the algorithm by introducing a concept of consistency, which is used as the criteria for splitting the data instead of information gain. By so doing their improved algorithm solved the issue of feature bias towards multi-valued attributes and reduced the computational complexity of the standard ID3 algorithm. In another research, Soni and Pawar, 2017 proposed an optimized ID3 algorithm aimed at overcoming the limitations of the standard ID3. This improved algorithm was also applied to solve the problem of ID3 algorithm's feature bias to multi-valued attributes by calculating attribute importance, which was combined with information gain to get a novel feature selection approach for decision tree construction. Their algorithm builds simple decision trees and reduces the dimension of the tree. The error rate and accuracy shows that the proposed method gives better performance than the traditional ID3 algorithm.

Several studies have shown that random forest has better classification accuracy than most tree based algorithms, it is not susceptible to over-fitting, and has a high tolerance for noise and outliers. Man et al, 2018 proposed a Random Forest algorithm that optimizes the decision tree algorithm node splitting using adaptive selection of parameters. This is aimed at improving the classification accuracy of the algorithm. The optimization of node splitting was done on ID3 and CART, this implies both information gain and Gini index were used to get the node splitting formula. Using adaptive parameter selection resulted in optimal selection of features. The improved Random Forest node splitting algorithm was used to improve the accuracy of image classification. From the results received, it was seen that selecting the optimal coefficient of the combined algorithm, the proposed method significantly improved the image classification accuracy and efficiency. Fang et al, 2017 proposed the introduction of the concept of mutual information in the decision tree building process. In their algorithm, mutual information replaced the information gain as the criterion for selecting the splitting attribute. The obtained results show that mutual information gave better accuracy and the performance was better than the traditional ID3 algorithm.

3.2. Modifications to C4.5

In (Chen et al, 2013), the researchers proposed an improved C4.5 decision tree algorithm based on sample selection in order to improve the classification accuracy, reduce the training time of large sample, and find the best training set. Their algorithm was based on the fact that decision tree only get local optimal solution and has the bigger relativity with initial sample. In sample selection, the authors used iteration process to find the best training set. When experimented on large datasets the accuracy and time consumption of their proposed algorithm was better than the standard C4.5 algorithm. Muslim et al (2018) conducted a research to improve the accuracy of C4.5 decision tree algorithm. Their research was based on increasing the accuracy of the algorithm to predict credit receipts. In this research, the authors applied the split feature reduction model in the preprocessing stage. First, the dataset was inserted into WEKA and the features assessed using gain ratio algorithm to obtain the gain values of each feature, then split feature reduction model was applied by dividing the optimal 16 features into 4 splits. The first split consists of 16 features, the second comprised of 4 features, the third split comprised of 8 features, and the last split had 4 features. When the C4.5 algorithm was applied to each split, they produced varying results with the third split giving the best accuracy of 73.1%. Then Bagging Ensemble was applied to further increase the accuracy and the third split gave accuracy of 75.1%. When compared with the standard C4.5 algorithm it was observed that after applying split feature reduction and bagging ensemble the accuracy was increased by 4.6%.

In (Cherfi et al, 2018), a novel algorithm for building decision trees was proposed. This algorithm named VFC4.5 is an improvement of C4.5. It is an adaptation of the way C4.5, which finds the threshold of a continuous attribute. Rather than finding the threshold that maximizes gain ratio, the paper proposes to simply reduce the number of

candidate cut points by using arithmetic mean and median to improve a major weakness of the C4.5 algorithm when it deals with continuous attributes. When experimented using various datasets it was observed that in most tests the proposed VFC4.5 led to smaller decision trees, gave better accuracy, and is faster compared to the standard C4.5 algorithm. Furthermore, medical science has over the years embraced the computational ability of machine learning in diagnosing illness such as heart disease, breast cancer, and diabetes, among others. However, applications in the medical domain needs to be very efficient and accurate which led Muslim et al, 2017 to carry out a research on optimization of C4.5 algorithm for breast cancer diagnosis. To improve the performance, the researchers applied particle swarm optimization (PSO) as a feature option and to optimize the C4.5 algorithm. They used the Wisconsin Breast Cancer dataset gotten from the UCI repository. A performance evaluation was conducted using confusion matrix which produced accuracy of 95.61% for the standard C4.5 classification algorithm and 96.49% for the PSO based C4.5 algorithm, showing a 0.88% increase in accuracy. Lastly, several researchers have attempted to improve decision tree algorithms including (Yuan and Wang, 2016), (She et al., 2017), (Chandrasekar et al., 2017), (Li, 2017), and (Desai and Chaudhary, 2017).

3.3. Discussion on Reviewed Papers

The above discussion shows that many researchers have attempted to improve the traditional ID3 and C4.5 algorithms using various methods in several applications. A lot of these improved versions got better performance than their corresponding algorithms. The summary of these contributions are shown in table 1 below.

Table 1. Summary of various contributions

Author(s)	Algorithm Improved	Contribution(s)	Method
Yi-bin et al, 2017	ID3	Improvement of ID3 Algorithm Based on Simplified Information Entropy and Coordination Degree	Approximation using Maclaurin formula
Wang et al, 2017	ID3	Novel approach to select splitting attribute.	Rough set
Soni and Pawar, 2017	ID3	Novel feature selection approach for decision tree construction.	Attribute importance combined with information gain
Man et al, 2018	Random Forest	Optimization of decision tree algorithm node splitting	Adaptive Parameter Selection
Fang et al, 2017	ID3	Introduction of the concept of mutual information in selecting the splitting attribute instead of information gain	Mutual information
Chen et al, 2013	C4.5	Improving classification accuracy of C4.5 decision tree algorithm based on sample selection	Iteration process
Muslim et al, 2018	C4.5	Improving the accuracy of the algorithm to predict credit receipts	Split Feature reduction model and bagging ensemble.
Cherfi et al, 2018	C4.5	A novel algorithm for building smaller and more accurate decision trees	Arithmetic mean and median
Muslim et al, 2017	C4.5	Optimization of C4.5 algorithm for breast cancer diagnosis.	Particle Swarm Optimization

4. Conclusion and Future Works

Research in machine learning algorithms have enormous opportunities for researchers. It is an important tool used in finding new patterns and relationships in large datasets. Machine learning has enhanced different spheres of our lives and industries with efficient ways to improve effectiveness. Decision tree algorithms such as ID3 and C4.5 have over the years been the most used algorithms for classification. To this end many researchers have tried improving their effectiveness in order to get better predictions and to keep up to date with data which is

continuously changing. This paper researched and reviewed several of these improvements which are needed so as to present researchers with the work that has been done so far and lay the foundation for future research. Future research can examine how these improved algorithms have been applied in real world scenarios and their adoption by researchers. Currently, it seems the improved algorithms are isolated and their usefulness outside the research community cannot be ascertained. Furthermore, little research has been done on the use of evolutionary algorithms for optimal feature selection, further work needs to be done in this area as proper feature selection in large datasets can significantly improve the performance of the algorithms.

Acknowledgements

This research is supported partially by South African National Research Foundation Grants (No. 112108 and 112142), and South African National Research Foundation Incentive Grant (No. 95687), Eskom Tertiary Education Support Programme (Y. Sun, Z. Wang), Research grant from URC of University of Johannesburg.

References

[1] A. Desai, S. Chaudhary, Distributed decision tree v.2.0, in: 2017 IEEE International Conference on Big Data (Big Data). Presented at the 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 929–934.
[2] A. M. Almasoud, H. S. Al-Khalifa, and A. Al-Salman, "Recent developments in data mining applications and techniques," in 2015 Tenth International Conference on Digital Information Management (ICDIM), 2015, pp. 36–42.
[3] C. Anuradha and T. Velmurugan, A data mining based survey on student performance evaluation system, 2014 IEEE International Conference on Computational Intelligence and Computing Research, 2014, pp. 1–4.
[4] Cherfi, A., Nouira, K., and Ferchichi, A. (2018). Very Fast C4.5 Decision Tree Algorithm. Journal of Applied Artificial Intelligence, 2018, 32(2), pp. 119-139
[5] F. A. de A. Neto and A. Castro, A reference architecture for educational data mining, in 2017 IEEE Frontiers in Education Conference (FIE), 2017, pp. 1–8.
[6] F. Chen, X. Li, and L. Liu, Improved C4.5 decision tree algorithm based on sample selection, in 2013 IEEE 4th International Conference on Software Engineering and Service Science, 2013, pp. 779–782.
[7] H. Peng, X. Zhang, and L. Huang, An Energy Efficient Approach for C4.5 Algorithm using OpenCL Design Flow, 2017 IEEE International Conference on Field Programmable Technology (ICFPT), 2017
[8] K. Adhatrao, A. Gaykar, A. Dhawan, R. Jha, and V. Honrao, Predicting Students' Performance using ID3 and C4.5 Classification Algorithms. International Journal of Data Mining & Knowledge Management Process (IJDKP), 2013, 3(5)
[9] L. Fang, H. Jiang, and S. Cui, An improved decision tree algorithm based on mutual information, in 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2017, pp. 1615–1620.
[10] L. Yi-bin, W. Ying-ying, and R. Xue-wen, Improvement of ID3 algorithm based on simplified information entropy and coordination degree, in 2017 Chinese Automation Congress (CAC), 2017, pp. 1526–1530.
[11] M. A. Muslim, S. H. Rukmana, E. Sugiharti, B. Prasetiyo, and S. Alimah, Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis, J. Phys.: Conf. Ser., 983(1), p. 012063, 2018.
[12] M. A. Muslim, A. Nurzahputra, and B. Prasetiyo, Improving accuracy of C4.5 algorithm using split feature reduction model and bagging ensemble for credit card risk prediction, in 2018 IEEE International Conference on ICT (ICOIACT), 2018, pp. 141–145.
[13] M. Li, Application of CART decision tree combined with PCA algorithm in intrusion detection, Presented at the 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2017, pp. 38–41.
[14] P. Chandrasekar, K. Qian, H. Shahriar, and P. Bhattacharya, Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing, 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2017, pp. 481– 484.
[15] V. Mhetre and M. Nagar, Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA, in 2017 International Conference on Computing Methodologies and Communication (ICCMC), 2017, pp. 475–479.
[16] V. K. Soni and S. Pawar, Emotion based social media text classification using optimized improved ID3 classifier, in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 1500–1505.
[17] X. She, T. Lv, X. Liu. The Pruning Algorithm of Parallel Shared Decision Tree Based on Hadoop, Presented at the 2017 10th International Symposium on Computational Intelligence and Design (ISCID), 2017, pp. 480–483.
[18] W. Man, Y. Ji, and Z. Zhang, Image classification based on improved random forest algorithm, in 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2018, pp. 346–350.
[19] Z. Wang, Y. Liu, and L. Liu, A new way to choose splitting attribute in ID3 algorithm, in 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2017, pp. 659–663
[20] Z. Yuan, C. Wang, An improved network traffic classification algorithm based on Hadoop decision tree, Presented at the 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), 2016, pp. 53–56.