# METAANALYSIS OF OVERFITTING OF DECISION TREES

**Article** · June 2024



Abhijeet Mankar
Tuljaram Chaturchand College


Sudhakar Bhoite
Chh. Shahu Institute of Business Education & Research


Kabir Kharade
Shivaji University

**J**ournal of **N**onlinear **A**nalysis and **O**ptimization:
Theory and Applications
Vol. 15, No. 1(i), January-March 2024 ISSN: **1906-9685**

# METAANALYSIS OF OVERFITTING OF DECISION TREES

**A. D. Mankar**
Department of Computer Science, Tuljaram Chaturchand College of Arts, Science and Commerce, Baramati, Maharashtra, India
**S. D. Bhoite**
School of Computer Science and Applications, Chhatrapati Shahu Institute of Business Education and Research, Kolhapur, Maharashtra, India
**K. G. Kharade**
Department of Computer Science, Shivaji Univesity, Kolhapur, Maharashtra, India
**K. A. Raskar**
Department of Science and Computer Science, MIT College(Alandi), Pune, Maharashtra, India

***Abstract*:**

Decision tree algorithms are used in machine learning. Decision tree algorithms are widely used as they are easy to interpret. We are able to explain the reason behind the decision. While working with decision tree algorithms, the concept of overfitting comes in picture. Overfitting is addressed with the use of hyperparameters. How change in a hyperparameter value affects the model cannot be measured directly. In this study, we reviewed various papers related to overfitting, hyperparameters and pruning techniques in order to conduct metaanalysis of overfitting of decision trees.

***Keywords*:**

***Overfitting, decision trees, Pre-pruning, Post-pruning, reduced error pruning***

## 1. INTRODUCTION

Decision tree algorithms are comprehensible as they give an insight in the prediction.Decision trees are referred to as white box approach since they help us to understand intermediate decision nodes, as compared to artificial neural network since artificial neural networks are more complexto interpret.As a result of this artificial neural networks are referred to as black box approach. Decision trees are more popular in machine learning.In machine learning the dataset is divided into training and test sets andsometimes it is also divided into training, validation,and test data sets.There are popular decision tree algorithms like ID3, C4.5, C5.0, and CART. They are comparable with each other in terms of speed, pruning, and ability to deal with missing values. In this paper, we discuss the various research papers related to overfitting in decision trees and the methods used to deal with overfitting by the researchers.

## 2. IMPORTANCE OF PRUNING IN OVERFITTING

### 2.1 Description of overfitting

Decision trees sufferfrom the problem of overfitting. Overfitting is a phenomenon in which the decision tree model fits to the training data set and it becomes less accurate for the new data or for future data.Overfitting affects the performance of the model on the test data. It is due to the presence of noise or irrelevant features in the training data. Overfitting can also occur due to growth of a decision tree to a considerable depth, as deeper trees likely to result in overfitting.

### 2.2 Methods to avoid overfitting

The problem of overfitting can be solved by stopping to grow the tree earlier. This results in small training set which easily fits into the model.Another method uses training and validation set. This method works even if the training set is misinterpreted due to random errors.Another way is to use statistical test to determine whether to expand a node in a decision tree or not which ultimately improves performance beyond the training set. In another method, minimum description length principle is used. In this method, complexity for encoding training examples and the decision tree is explicitly measured. As the encoding size if found to be minimized, measuring is stopped.[1]

Journal of Nonlinear Analysis and Optimization:
Theory and Applications
Vol. 15, No. 1(i), January-March 2024 ISSN: **1906-9685**

### 2.3 Importance of pruning to avoid overfitting

Pruning of decision trees can be done in either of the two ways namely pre-pruning and post-pruning. Pre-pruning does not allow the decision tree to fully grow whereas post-pruning allows to fit the training data completely and is curtailed according to some criteria. The curtailed tree is a simplified version of the original with the minimal relevant branches removed. Pre-pruning can be implemented by using hyperparameter tuning.

Reduced error pruning removes the nodes including the subtree of that node that does not improve the accuracy on a validation set. It replaces the subtree into a leaf node labeled with the majority class at that node.[1] rpart package in RStudio (which is an implementation of CART algorithm) uses cost-complexity pruning.prune.rpart() uses ‘cp’ as complexity parameter.[2]

## 3. ANALYSIS OF PREVIOUS RESEARCH PAPERS

The techniques/concepts discussed in the papers reviewed and the contributions of study are discussed in the following table.

Table 1 :Tabular representation of findings of research papers

| References | Techniques/Concepts discussed | Findings/Contributions of study |
|---|---|---|
| [3] | Disguised Bayesian Analysis, Description noise, Classification noise | 1) The strategy of choosing the best fitting complex model produces rules with a higher average prediction accuracy than strategy of choosing the best fitting simple model in equivocal cases 2) Overfitting causes decline in prediction accuracy in case of the presence of notable description noise or when there is no sharp differentiation in classification patterns in description groups in typical problems. 3) The assumption by the statistical approach that the attributes are insignificant to classification is accepted by default in the absence of proof is favourable to the31uccesss of simplification techniques. |
| [4] | Cross validated cost complexity pruning (CV), No Pruning (NP) | 1) In presence of sparsest data, avoidance of overfitting through pruning reduces performance, as predicted instead of improving it. 2) For Cleveland Heart Disease data, as authors reduce the size of training set, the predictive accuracy of trees produced by NP jumps from one percent below those yielded by CV to higher than three percent above. 3) In case of Letter Recognition data, CV becomes better than NP as size of training sets go above 1000. 4) In case of Hypothyroid data, CV is superior than NP for training set of 45 cases. |
| [5] | naïve strategy, sophisticated strategy | 1) Accuracy of trees chosen by two strategies differs as classification noise increases 2) When the strategies differ, the tree selected by overlooking the chance of overfitting is most likely to be better regardless of the level of noise 3) The rise in classification noise declines the value of staying away from overfitting 4) The sophisticated strategy is expected to construct an inferior(subordinate) tree in case of high noise |
| [6] | C4.5, TBA (Tree building with Bonferroni Adjustment) | 1)TBA produces smaller trees as compared to C4.5 without compromising accuracy 2) TBA is easy, but its performance is poor while dealing with missing values |
| [7] | Fishers Pruning, Error-based Pruning, MDL, Bonferroni Pruning | 1) Bonferroni produces the least complicated and most correct trees 2) When N<20, no technique yields the correct tree 3) For N between 20 and 30, accuracy of all pruned trees is approximately the same and the average accuracy of unpruned trees is much lower 4) For N greater than 40, |

| | | |
|---|---|---|
| | | Bonferroni attains close to maximum average accuracy while other techniques have lower average accuracy |
| [8] | Reduced Error Pruning (REP), Pessimistic Error Pruning (PEP), Minimum Error Pruning (MEP), Critical Value Pruning (CVP), Cost-Complexity Pruning (CCP), Error-Based Pruning (EBP) are considered. Results of 3375 different experiments of pruning methods on various data sets are summarized and discussed. | 1)Pruning does not usually decrease the predictive accuracy 2) CV-1SE displays worst performance followed by 1SE and REP 3) 0SE, CV-0SE, EBP, and PEP performed equally well 4) No sign about the methods taking advantage of an independent pruning set surely performs better than others 5)REP tends to over prune |
| [9] | Alopex Perceptron Decision Tree (APDT) algorithm | 1) The authors present a new algorithm called the Alopex Perceptron Decision Tree (APDT) algorithm for learning a decision tree given a set of preclassified training patterns. 2) Simulation results show that the APDT algorithm learns trees that are superior in terms of accuracy and size compared to other oblique decision tree induction methods. 3) The APDT algorithm presented in this paper can address only 2-class problems. 4) The pruning algorithms recommended by the authors are entirely robust with respect to classification noise in the given sample of pattern vectors. |
| [10] | PrismTCS with J-Pruning, PrismTCS without pruning | Demonstration of the importance of the information theoretic J-measure as the basis for reducing overfitting |
| [11] | J-Pruning, No pruning | 1)J-pruning technique is robust in the presence of noise 2) J-pruning substantially reduces the number of classification rules |
| [12] | J-Pruning | The authors have illustrated prospective value of J-measure for reducing overfitting by pre-pruning branches during classification tree generation. |
| [13] | Cost-Complexity Pruning, Pessimistic Error Pruning, Reduced Error Pruning, Genetic algorithm | With the assumption of not losing accuracy, pruning method based on genetic algorithm can effectively reduce the size of the tree. |
| [14] | C4.5, MLPC (Multi Level Pruned Classifier) | 1) MLPC generates a smaller number of decision rules as compared to C4.5 2) The network bandwidth required to transfer the decision table is reduced 3) On an average MLPC gives 80% accuracy while C4.5 given 62% accuracy 4) False positive rate (FPR) of MLPC is 12% while C4.5 has 30% |

| | | |
|---|---|---|
| [15] | TCSDT-SM, TCSDT-TP, TCSDT-FS | 1) Introduce feature selection before building decision tree, smoothing, and pruning process before calculating the class probability for each decision tree leaf 2) Three new methods TCSDT-SM, TCSDT-TP, TCSDT-FS obtained lower cost as compared to TCSDT algorithm. |
| [16] | Discrete-tree induction method,C4.5, Naïve Bayes, k-Nearest Neighbor, Support Vector Machine | 1) Heuristic-based decision tree induction method builds a good predictive model on categorical data set. 2) It also builds a compact tree model. |
| [17] | Random Forest, Decision Tree(J48) | 1) Random Forest proves better for the same number of attributes and large datasets while J48 is better with small datasets. 2) An increase in the number of instances from 286 to 699 results in an increase in correctly classified instances from 69.23% to 96.13% for the Random Forest. 3) The scenarios in the experiment cover the missing values in the datasets and also overcomes the overfitting problem caused because of missing values in the datasets. |
| [18] | Robust C4.5, PSRCG | 1) The authors looked into four PS (prototype selection) algorithms as data preprocessing for tree simplification and compared them to cutting-edge tree simplification techniques. 2) PSRCG has demonstrated on 22 datasets the highest efficiency of the PS approaches studies in this research (i) by reducing the decision tree size, (ii) by increasing the generalization accuracy, and (iii) by tolerating the presence of noise 3) When the decision tree is constructed using C4.5 and its error-based pruning, the authors have theoretically established the relationship between training set size and tree size in this study. |
| [19] | 1)Post Pruning: Reduced Error Pruning, Error complexity pruning, Minimum Error pruning, Cost based pruning 2) Pre-Pruning: Minimum no. of object pruning, Chi-square pruning | 1)Each time minnumobj increases, the accuracy is almost the same but the tree size changes. 2)Every time the number of folds are increased, accuracy for diabetes and glass datasets increases. 3) For glass and diabetes data set, as the confidence factor increases, the accuracy remains almost the same, but the tree size increases, so it appears ineffective. |
| [20] | Penalty methods, Early stopping, Comparison between overfitting and underfitting | 1)Early stopping method is superior as compared to penalty method that can avoid overfitting and underfitting with taking care to the validation time. 2) Penalty method is sensitive to variance and bias, on the other hand early stopping method is less sensitive to variance and bias. 3) Underfitting neural networks perform poorly on both training and test sets while overfitting networks may perform very well on training sets but poor on test sets. |
| [21] | Simple Random Search (RS), Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Estimation of | 1) Comparison of different hyperparameter tuning techniques for DT algorithms at large-scale 2) Thorough analysis of the effect of hyperparameters on the predictive performance of the induced models |

| | | |
|---|---|---|
| | Distribution Algorithm (EDA), Sequential Model-based Optimization (SMBO), Iterated F-race (Irace), Default hyperparameter values recommended for C4.5,CART, Ctree | |
| [22] | Lempel-Ziv-Markov chain algorithm, zlib algorithm, bz2 algorithm, CART, C4.5, CHAID, PUBLIC | 1)Minimum Surfeit and Inaccuracy or MSI algorithm designed 2) New algorithm requires an average of 5.7 nodes as compared to 23 nodes for CART algorithm 3) New algorithm has an average depth of 2.2 nodes while the average depth obtained by the CART algorithm is 4.8 nodes 4) New algorithm does not need optimization of hyperparameter whereas CART algorithm has such requirement |
| [23] | Adaptive overfitting | Results of the experiments raise question on statement that overfitting is a notable danger in the machine learning workflow. |
| [24] | Adaptive overfitting | 1) The overall competition is not affected by substantial overfitting. 2) For some competition authors observe proof of mild overfitting, but always with small effect size 3) Findings are evidence that overfitting does not cause a notable threat in the most favoured classification accuracy competitions on Kaggle. |
| [25] | Early stopping, Network reduction, Expansion of the training data, Regularization | 1)Early stopping strategy helps to stop training before learning noises, 2) Network reduction strategy provide us an approach to reduce noises in the training set, data expansion strategy is proposed for complicated models requiring abundant data to fine tune their hyperparameters, 3) Regularization helps us to distinguish noises, meaning and purposeless features and allot non-identical weights to them. |
| [26] | model without post-pruning, model after applying Bayes Risk Post-Pruning | 1) After applying Bayes Risk Post-Pruning, the model performance in the testing dataset was superior as compared to training dataset 2) The value of accuracy, precision, and recall for decision tree after applying Bayes Risk Post-Pruning were better than without post-pruning 3) The best model for both datasets after applying Bayes Risk Post-Pruning is achieved if the proportion of training dataset chosen is 80% |
| [27] | no reduction-no pruning (NR-NP), no reduction-pruning(NR-P), reduction-no pruning(R-N), reduction with pruning(R-P) | 1)Large scale empirical experiments on 13 UCI machine learning repository databases were conducted, 2)Noise filters(ENN,RENN and ALLKNN) surpassed decision trees learning without pruning in 9,9,8 datasets respectively. 3) In all over 7690 different experiments were carried out based on the combinations of reduction and pruning, 4) DROP3 and DROP5 performed better than instance reducers while they performed worse than noise filters. |
| [28] | C4.5 Decision Tree Algorithm, Genetic Algorithm, Information gain | 1) The proposed method without use of genetic algorithm resulted rise in the percentage of overfitting in three out of four datasets 2) The proposed method with genetic algorithm yields better success rate in all datasets. |

Vol. 15, No. 1(i), January-March 2024 ISSN: **1906-9685**

## 4. CONCLUSION

Based on the critical review of research articles related to overfitting of decision trees we draw following conclusions:

1. Majority of papers discussed pruning as the most sought technique to tackle overfitting problem.
2. About sixteen different types of pruning methods and its variations were cited in the paper.
3. WEKA data mining tool was used by four research articles.
4. Decision tree algorithms like C4.5, J48, CART, CHAID, PUBLIC were used by the researchers under references.
5. Increase in classification noise and missing values impacts the performance of decision trees.
6. Minimum Surfeit and Inaccuracy(MSI) algorithm required average 24.78% of nodes as required by CART. Average depth of tree obtained by MSI is 45.83% of the average depth of tree obtained by CART.
7. Early stopping strategy, Network reduction strategy, Regularization deal with noises while Data expansion strategy underlines the importance of expanded dataset for significant improvement in prediction accuracy.
8. Bayes Risk Post-Pruning improves accuracy as the size of the training dataset increases.
9. The article gives the summarised view as different researchers considered different techniques/concepts.

This paper gives a brief insight of the researches related to decision trees and overfitting problem by the researchers during 1991-2021.

## ACKNOWLEDGEMENT

It is our pleasure to acknowledge the valuable suggestions given by Late Dr. S. D. Gore. We also are thankful to our family members, higher authorities of out institutions and our colleagues who are motivation for us.

## REFERENCES

[1]Acharya, S., "*Data Analytics using R"*, McGraw Hill, 2018.

[2]The R Core Team (2023). *R : A Language and Environment for Statistical Computing Reference Index* (p. 3464). R Foundation for Statistical Computing. https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf Accessed on 14 December 2023.

[3] C. Schaffer, "When does overfitting decrease prediction accuracy in induced decision trees and rule sets?," in *Machine Learning—EWSL-91: European Working Session on Learning Porto, Portugal Proceedings,* Springer-Verlag Berlin Heidelberg, Vol. 482, pp. 192-205, 1991.

[4] Cullen Schaffer, "Sparse data and the effect of overfitting avoidance in decision tree induction," in *Proceedings of the tenth national conference on Artificial intelligence (AAAI'92) AAAI Press,* pp. 147–152, 1992.

[5]C. Schaffer, "Overfitting avoidance as bias", *Machine learning*, Vol.10, pp. 153-178,1993.

[6]D. Jensen, "Adjusting for multiple testing in decision tree pruning", in *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics, PMLR*, pp. 295-302, 1997.

[7]D. D. Jensen & M. D. Schmill, "Adjusting for Multiple Comparisons in Decision Tree Pruning," in *Proceedings of KDD-97*, pp. 195-198, 1997.

[8]F. Esposito, D. Malerba, G. Semeraro, & J. Kay, "A comparative analysis of methods for pruning decision trees", *IEEE transactions on pattern analysis and machine intelligence,* Vol. 19, No. 5, pp. 476-491,1997.

[9]S. Shah, S., & P. S. Sastry, "New algorithms for learning and pruning oblique decision trees", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* Vol. 29, No. 4, pp. 494-505, 1999.

[10]M. Bramer, "An Information-Theoretic Approach to the Pre-Pruning of Classification Rules," in: *Musen, M.A., Neumann, B., Studer, R. (eds) Intelligent Information Processing. IIP 2002. IFIP — The International Federation for Information Processing, Springer, Boston, MA,* Vol .93, pp. 201-212, 2002, 10.1007/978-0-387-35602-0_18.

[11]M.Bramer, "Pre-pruning Classification Trees to Reduce Overfitting in Noisy Domains," in *Proceedings of the Third International Conference on Intelligent Data Engineering and Automated Learning (IDEAL '02). Springer-Verlag, Berlin, Heidelberg*, pp. 7–12, 2002.

[12]M. Bramer, "Using J-pruning to reduce overfitting in classification trees", *Knowledge-Based Systems*, Vol. 15, No. (5-6), pp. 301-308, 2002.

[13]J. Chen, X. Wang & J. Zhai, "Pruning decision tree using genetic algorithms," in *2009 International Conference on Artificial Intelligence and Computational Intelligence*,Vol. 3, pp. 244-248, IEEE, 2009.

[14]D. D. Patil, V. M. Wadhai, & J. A. Gokhale, "Evaluation of decision tree pruning algorithms for complexity and classification accuracy", *International Journal of Computer Applications*, Vol. 11, No. 2, pp. 23-30, 2010.

[15]T. Wang, Z. Qin, Z. Jin, & S. Zhang, "Handling over-fitting in test cost-sensitive decision tree learning by feature selection, smoothing and pruning", *Journal of Systems and Software*, Vol. 83, No. 7, pp. 1137-1147, 2010.

[16]N. Kerdprasop, & K. Kerdprasop, "Discrete decision tree induction to avoid overfitting on categorical data," in *Proceedings of the Mamectis/Nolasc/Control/Wamus*, pp. 247-252, 2011.

[17]J. Ali, R. Khan, N. Ahmad, & I. Maqsood, "Random forests and decision trees", *International Journal of Computer Science Issues (IJCSI)*, Vol. 9, No. 5, pp. 272-278, 2012.

[18]M. Sebban et al., "Impact of learning set quality and size on decision tree performances", *IJCSS*, Vol. 1, No. 1, pp. 85-105, 2000.

[19]N. Patel, & S. Upadhyay, "Study of various decision tree pruning methods with their empirical comparison in WEKA", *International journal of computer applications*, Vol. 60, No. 12, pp. 20-25, 2012.

[20]H. Jabbar, & R. Z. Khan, "Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)", *Computer Science, Communication and Instrumentation Devices*, pp. 163-172, 2014, doi:10.3850/978-981-09-5247-1_017.

[21]R. Mantovani et al., "An empirical study on hyperparameter tuning of decision trees", *arXiv:1812.02207v1*, 2018, Available also from: [1812.02207v1] An empirical study on hyperparameter tuning of decision trees (arxiv.org)

[22]R. G. Leiva et al., "A novel hyperparameter-free approach to decision tree construction that avoids overfitting by design", *Ieee Access*, Vol. 7, pp. 99978-99987, 2019.

[23]R. Roelofs, "Measuring Generalization and overfitting in Machine learning", Technical Report, UCB/EECS-2019-102, University of California, Berkeley, 2019.

[24]R. Roelofs et al., "A meta-analysis of overfitting in machine learning", in Advances in Neural Information Processing Systems, Vol. 32, 2019.

[25]X. Ying, "An overview of overfitting and its solutions", *In Journal of physics: Conference series,* Vol. 1168, No. 2, p. 022022, IOP Publishing, 2019, DOI 10.1088/1742-6596/1168/2/022022

[26]D. Christianti, S. Abdullah, & S. Nurrohmah, "Bayes Risk Post-Pruning in Decision Tree to Overcome Overfitting Problem on Customer Churn Classification" in *Proceedings of the 1st International Conference on Statistics and Analytics, ICSA* 2020. http://dx.doi.org/10.4108/eai.2-8-2019.2290487

[27]A. Amro et al., "Instance reduction for avoiding overfitting in decision trees", *Journal of Intelligent Systems*, Vol. 30, No. 1, pp. 438-459, 2021.

[28]M. M. Mijwil, & R. A. Abttan, "Utilizing the genetic algorithm to pruning the C4. 5 decision tree algorithm", *Asian Journal of Applied Sciences*, Vol. 9, No. 1, pp. 45-52, 2021.