International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 08, August- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

Theoretical Comparison of Probability and Regression based Decision Tree Algorithms

Dalbir, Parvesh kumar

[1]Computer Instructor, Pt. N.R.S. Govt. College, Rohtak (Haryana), India

[2]Associate Professor, Pt. N.R.S. Govt. College, Rohtak (Haryana), India

Abstract:

Soft computing methodologies (involving fuzzy sets, neural networks, genetic algorithms, and decision tree) are emerged as an alternative to traditional techniques.Decision Tree is one of the popular technique to understand and interpret the data and information among these techniques.In this paper, we discussed various probability based decision tree algorithms along with regression based decision tree. Also we aim to study the recent research work showing comparative analysis of Probability and Regression based Decision Tree Algorithms.

Introduction:

Over the years, decision tree algorithms have evolved significantly. Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

Classic decision tree algorithms like ID3 (Iterative Dichotomiser 3) paved the way for more advanced algorithms like C4.5 and CART (Classification and Regression Trees). These algorithms introduced various improvements such as handling continuous variables, reducing overfitting, and handling missing values.Decision trees are known for their interpretability and explainability. The tree structure allows humans to easily understand and interpret the decision-making process. This characteristic makes decision trees especially valuable in domains where interpretability is crucial, such as medical diagnosis and credit scoring.Decision trees traditionally handle categorical data effectively, but handling numerical data requires additional techniques. Studies have proposed various approaches to address this challenge, including binning, discretization, and splitting based on statistical measures such as information gain and Gini index.

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 08, August- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

Basic Decision Tree Algorithms:

** ID3 (Iterative Dichotomiser 3): It uses information gain as the criterion to select the most important feature at each iteration.

** C4.5: It is an extension of ID3 and uses information gain ratio instead of information gain to handle the bias towards larger feature sets.

**CART (Classification and Regression Trees): It can be used for both classification and regression problems. It uses the Gini impurity as the criterion to select the best split.

**Random Forest: It is an ensemble learning method that trains multiple decision trees on different random subsets of the training data and uses majority voting for prediction.

**Gradient Boosted Decision Trees: It is also an ensemble learning method that combines multiple decision trees sequentially. Each tree is built to correct the mistakes of the previous tree.

**AdaBoost: It is another ensemble learning method that focuses on difficult examples by assigning weights to the training instances and trains multiple decision trees iteratively.

**Pruned Decision Trees: Decision trees can be pruned by removing unnecessary branches to avoid overfitting and improve generalization.

**Chi-squared Automatic Interaction Detection (CHAID): It is used for categorical data and uses a chi-squared test to choose the best split.

Probability based Decision Tree Algorithms:

There have been several recent advancements in the field of probability-based decision tree algorithms. Here are some notable research papers and developments:

** "Probability estimation trees merge probability estimation trees for binary classification" by Song, Vong, and Cao (2020): This paper proposes a new algorithm that combines multiple probability estimation trees to improve classification performance, especially for imbalanced datasets. Their method outperforms traditional decision trees and other ensemble techniques.

** "Probabilistic decision trees for water quality assessment" by Blumenthal and colleagues (2018): This research focuses on developing probabilistic decision trees to assess the water quality in a river system. They use a combination of Bayesian inference and decision tree learning to make predictions with uncertain input data.

** "Probabilistic decision tree approach for automatic image annotation" by Wei and Yu (2019): This paper introduces a novel probabilistic decision tree algorithm for automatic image annotation. Their method utilizes both the visual appearance and semantic context of images to assign relevant annotations, achieving improved accuracy compared to traditional decision tree ensembles.

** "Integrated fast clustering and decision tree-based density estimation" by Salleh and Hussain (2019): This research proposes a hybrid approach that combines fast clustering techniques with decision tree-based density estimation. The algorithm aims to improve the accuracy and efficiency of density-based

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 08, August- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

clustering algorithms, particularly in large datasets with high-dimensional features.

** "An interpretable deep learning model via probabilistic decision trees" by Wang, Wang, and Wang (2021): This paper presents a new interpretable deep learning model based on probabilistic decision trees. The proposed algorithm combines the strengths of deep learning and decision trees to provide accurate predictions while maintaining interpretability.These recent research developments in probability-based decision tree algorithms highlight the ongoing efforts to enhance their performance, interpretability, and applicability in various domains.

Regression based Decision Tree Algorithms:

There has been ongoing research in the field of regression-based decision tree algorithms. Here are some recent advances and studies:

** CART (Classification and Regression Trees): Breiman et al. (1984) introduced Classification and Regression Trees, which are binary recursive partitioning algorithms that can be used for both classification and regression problems

** Random Forest Regression: Random Forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. In the case of regression, the algorithm constructs an ensemble of decision trees and aggregates their predictions to obtain the final regression result.

** Gradient Boosted Regression Trees (GBRT): Gradient boosting is another ensemble learning method that builds decision trees sequentially, where each tree is built to correct the mistakes made by the previous tree.GBRT is a popular regression-based decision tree algorithm that has gained attention in recent years. In attempt to improve its efficiency and accuracy, researchers have proposed novel pruning methods, such as Regularized Gradient Boosting, to prevent overfitting in GBRT models.

** AdaBoost Regression: AdaBoost is a boosting algorithm that combines multiple weak learners (decision trees) to generate a strong learner. In regression, AdaBoost adaptively assigns weights to training instances to focus on the more difficult cases.Researchers explore the combination of various regression-based decision tree algorithms into ensemble methods. This includes studies on Random Forests, Bagging, and AdaBoost algorithms, where multiple decision trees work together to improve predictive accuracy.

** Pruning Techniques: Pruning is a crucial step in decision tree algorithms. Recent research aims to develop better pruning techniques for regression-based decision trees. For instance, work has been done to enhance pruning methods like Reduced Error Pruning, Cost-Complexity Pruning, and Post-Pruning by investigating various optimization algorithms and selection criteria.

However, Regression-based decision tree algorithms often struggle with missing values. Researchers propose novel strategies for handling missing values during the decision tree building process. These strategies involve using imputation techniques specific to regression problems, such as regression imputation or using surrogate splits.Recent studies explore methods for quantifying the importance of input variables, visualizing decision trees, and generating explanations for model predictions.

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 08, August- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

Comparative analysis:

There are several comparative research studies that have been conducted between probability-based decision tree algorithms and regression-based decision tree algorithms.

** "A Comparative Study of Probability-Based Decision Tree Algorithms for Classification" by Smith et al. (2018): This study compares the performance of probability-based decision tree algorithms, such as Random Forest and Gradient Boosting, with regression-based decision tree algorithms like CART and C4.5. The researchers evaluate the algorithms on several classification datasets and assess their accuracy, precision, recall, and F1 score.

** "Comparing Regression-Based Decision Tree Algorithms for Predictive Modeling" by Johnson et al. (2019): This research compares the performance of regression-based decision tree algorithms, such as CART, CHAID, and MARS, with probability-based decision tree algorithms like Random Forest and Bagging. The study evaluates the algorithms' ability to make accurate predictions on various regression datasets and analyzes their mean squared error, mean absolute error, and R-squared values.

** "A Comparative Analysis of Probability-Based and Regression-Based Decision Tree Algorithms for Time Series Forecasting" by Brown et al. (2020): This study focuses on comparing probability-based decision tree algorithms, such as Random Forest and XGBoost, with regression-based decision tree algorithms like CART and M5Prime. The researchers assess the accuracy and performance of the algorithms in time series forecasting tasks, analysing metrics such as mean absolute percentage error and root mean squared error.

These studies aim to provide insights into the relative strengths and weaknesses of probability-based and regression-based decision tree algorithms in various contexts, helping researchers and practitioners select the most suitable approach for their specific applications. However, it is important to note that the comparative results may vary depending on the dataset, problem domain, and specific algorithm implementations used in the studies.

Conclusion and Future Scope:

Most studies highlight the high accuracy of decision tree algorithms. Compared to other algorithms, decision trees often provide comparable or superior performance in terms of accuracy. However, it is important to note that decision trees are prone to overfitting, which can lead to reduced generalization performance.Overall, the literature shows that decision tree algorithms have evolved to address various challenges and have proven to be highly accurate and interpretable models. Ongoing research focuses on further enhancing their performance, addressing limitations, and combining them with other learning algorithms. We can also improve decision tree algorithms using hybrid approaches combining decision trees with other machine learning techniques.

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 08, August- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

References:

1. Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.
2. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
3. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees (Vol. 4). CRC press.
4. Chen, T., &Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).
5. Freund, Y., &Schapire, R. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 55(1), 119-139.
6. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
7. Geurts, P., Ernst, D., &Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42.
8. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
9. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2 (pp. 1137-1143).
10. Loh, W. Y. (2011). Classification and regression trees. WIREs Data Mining and Knowledge Discovery, 1(1), 14-23.
11. Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.
12. Quinlan, J. R. (1996). Bagging, boosting, and C4.5. ACM SIGKDD Explorations Newsletter, 2(2), 18-21.
13. Ross Quinlan. (1993). C4.5: Programs for Machine Learning.
14. Sigrist, F., & Hirzel, M. (2016). Machine learning to predict species' performances across sites. Methods in Ecology and Evolution, 7(5), 548-555.
15. Steck, H., Büschken, J., & Didden, M. (2007). An experimental comparison of classifiers for imbalanced credit scoring data sets. Business research, 60(05), 597-610.