EJTAS
ISSN 2786-7447
EUROPEAN JOURNAL OF THEORETICAL AND APPLIED SCIENCES
OPEN ACCESS
PEER REVIEWED

Unraveling the Significance of the Classification Tree Algorithm in Machine Learning: A Literature Review

Michael E. Bensi
Graduate School, Angeles University Foundation, Philippines

Rossana A. Esquivel
Graduate School, Angeles University Foundation, Philippines

Suggested Citation

Bensi, M.E. & Esquivel, R.A. (2023). Unraveling the Significance of the Classification Tree Algorithm in Machine Learning: A Literature Review. *European Journal of Theoretical and Applied Sciences, 1*(5), 604-611.
DOI: 10.59324/ejtas.2023.1(5).49

Abstract:

Machine learning, an integral component of Artificial Intelligence (AI), empowers systems to autonomously enhance their performance through experiential learning. This paper presents a comprehensive overview of the Classification Tree Algorithm's pivotal role in the realm of machine learning. This algorithm simplifies the process of categorizing new instances into predefined classes, leveraging their unique attributes. It has firmly established itself as a cornerstone within the broader landscape of classification techniques. This paper delves into the multifaceted concepts, terminologies, principles, and ideas that orbit the Classification Tree Algorithm. It sheds light on the algorithm's essence, providing readers with a clearer and more profound understanding of its inner workings. By synthesizing a plethora of existing research, this endeavor contributes to the enrichment of the discourse surrounding classification tree algorithms. In summary, the Classification Tree Algorithm plays a fundamental role in machine learning, facilitating data classification, and empowering decision-making across domains. Its adaptability, alongside emerging variations and innovative techniques, ensures its continued relevance in the ever-evolving landscape of artificial intelligence and data analysis.

Keywords: *Classification Tree Algorithm, Literature Review, Machine Learning.*

Introduction

Machine learning, an integral component of Artificial Intelligence, empowers systems to autonomously enhance their performance through experiential learning. It encompasses three fundamental learning paradigms: supervised learning, unsupervised learning, and reinforcement learning. Within this domain, challenges take various forms such as classification, focusing on categorical solutions, regression to predict continuous values, and clustering to discern intricate data patterns. Notably, decision trees are a prevalent choice for classification tasks, with four primary tools — Naïve Bayes, Decision Trees, Logistic Regression, and Random Forest — serving as foundational pillars for solving these complex problems.

The realm of machine learning, particularly in the domain of classification tasks, has been profoundly shaped by the pivotal role that the classification tree algorithm plays. At its core, this algorithm encapsulates a foundational methodology that streamlines the process of categorizing new instances into predefined classes, leveraging the distinctive attributes exhibited by each instance. As a testament to its wide-ranging utility, the classification tree algorithm has firmly entrenched itself as a cornerstone within the broader landscape of classification techniques.

This paper undertakes the task of delving into the extensive literature and studies available, with the aim of shedding light on the multifaceted concepts, terminologies, principles, and ideas that orbit the classification tree algorithm. Through this review, the goal is to unravel the intricate layers that constitute the algorithm's essence, providing readers with a clearer and more profound understanding of its inner workings. By synthesizing and distilling a plethora of existing research, this endeavor seeks to provide a comprehensive framework that encompasses the algorithm's evolution, theoretical underpinnings, and practical applications. Through the amalgamation of diverse perspectives and insights, this review strives to contribute to the enrichment of the discourse surrounding classification tree algorithms.

The subsequent sections of this paper provide an overview of the classification tree and emphasize its significance in machine learning. Additionally, these sections explore popular methods employed for classification tasks. Moreover, a dedicated segment engages in a discussion of the importance of the classification tree algorithm. This discussion extends to explain the concepts of nodes, branches, leaves, and root nodes. Lastly, the splitting criteria and measures in the classification tree algorithm are explained in the final portion of the review.

Classification Tree Algorithm: Its Importance in Machine Learning

The utilization of a classification tree algorithm involves the assignment of a new instance to a predetermined category based on its attributes. This technique is frequently employed in the classification field. The algorithm implements a tree-like structure where each internal node represents an attribute test, and each leaf node corresponds to a class label. The data is recursively divided based on attribute values to generate the tree. The decision rules derived from the tree can be utilized to anticipate the class label of novel instances. Various adaptations of classification tree algorithms have been proposed, including the Direct Nonparametric Predictive Interface (D-NPI) algorithm (Özcan & Peker, 2023) and the Tree Penalized Linear Discriminant Analysis (TPLDA) algorithm (Alharbi, Coolen, & Coolen-Maturi, 2021). The goal of these algorithms is to enhance classification accuracy, interpretability, and variable selection in the tree construction process.

Overview of the classification tree algorithm as a popular method for classification tasks

Decision tree classifiers are widely acknowledged as one of the most prevalent techniques for data classification. Their extensive applicability ranges from medical disease analysis to text classification and image classification (Charbuty & Abdulazeez, 2021). Nevertheless, the conventional decision tree methodologies have certain limitations when it comes to predicting categorical variables based on groups of inputs.

To address these limitations and explore more sophisticated solutions, researchers have introduced novel tree-based approaches that offer enhanced accuracy, interpretability, and efficiency. One such method is the Tree Penalized Linear Discriminant Analysis (TPLDA) which generates a classification rule grounded on groups of variables, rendering resulting trees more readily understandable and computationally less demanding (Poterie, Dupuy, Monbet, & Rouviere, 2019).

Another avenue for enhancing classification accuracy and efficiency involves the adoption of a global decision tree paradigm within the context of quantum classifiers. This paradigm merges the Bayesian algorithm and the quantum decision tree classification algorithm, enabling the management of incremental data and achieving high accuracy and efficiency for big data classification tasks (Ji, Bao, Mu, Chen, Yang & Wang, 2023).

Additionally, researchers have explored an enhanced non-parametric kernel functions incorporating modified entropy to improve the classification, as demonstrated in a previous study (Rajeshkanna & Arunesh, 2021). Lastly, a novel modularity-based hierarchical classification tree (MHCT) has been introduced as a supervised learning technique to address multi-class classification challenges. Impressively, the MHCT methodology not only reduces training time but also achieves accuracy comparable to other established algorithms (Chengwei, Bofeng, Xinyue, Mingqing & Goubing, 2016). These innovative approaches signify the ongoing evaluation and diversification of the classification tree algorithm landscape, propelling the field forward with promising advancements.

Importance of Classification Tree Algorithm in Machine Learning

Classification algorithms are of utmost importance in the realm of machine learning. They serve the purpose of categorizing data into discrete classes or groups based on their attributes. These algorithms contribute significantly towards making informed predictions and decisions by discerning patterns and relationships from the given data. Their significance stems from the fact that they enable people to gain insights and comprehend intricate datasets, detect trends and patterns, and make precise predictions or classifications.

For instance, in the domain of education, classification algorithms have been employed to devise placement predictor systems that scrutinize previous placement history and various student attributes to ascertain their probability of securing employment (Khandelwal, Pareek, Dey, & Pareek, 2023). In the realm of medicine, these algorithms have been utilized to pre-categorize skin lesions into various types of skin cancers, thereby assisting medical professionals in rendering superior decisions for prompt diagnosis and treatment (S et al., 2022).

In the field of education, classification algorithms have been used to predict grade outcomes and evaluate student performance (Hongthong & Temdee, 2022). In the arena of integrated circuit testing, these algorithms have been harnessed to classify defects and curtail test duration (Song et al., 2022). All in all, classification algorithms represent indispensable tools in machine learning that empower people to extract meaningful insights and make well-informed decisions from intricate datasets.

Importance of Classification Tree Algorithm in Machine Learning

A classification tree algorithm is comprised of various components such as nodes, branches, leaves, and a root node. These components play crucial roles in classifying records based on their attributes. The root node, which serves as the starting point of the tree, represents the entire dataset. Nodes within the tree symbolize attribute tests, where internal nodes represent tests conducted on attributes. Each branch emerging from an internal node corresponds to the potential outcomes of the test. At the end of the branches, the leaves can be found, which represent different classes. Each leaf node is associated with a specific class label.

In the classification process, the algorithm employes attribute tests to navigate from the root node along branches to reach a leaf node. The attribute associated with the leaf node then determines the final classification result. This structured process of moving from the root to a relevant leaf node enables the algorithm to categorize instances effectively. This definition of the key components lays the foundation for understanding their respective roles and interactions within the classification tree algorithm (Dai, Zhang, & Wu, 2016; Audrey, Jean-Francois, Valerie, & Laurent, 2019)

Importance of Classification Tree Algorithm in Machine Learning

The utilization of splitting criteria and measures in classification tree algorithms is essential in determining the appropriate divisions of data at each node of the tree. Different criteria are employed to evaluate the quality of split and identify the optimal attribute for division. Before delving into specific approaches. One possible approach involves using a similarity formula to identify the attribute that exhibits the highest degree of similarity, which then becomes the splitting node (Zaim, Ramdani, & Haddi, 2018). Another method utilizes a Penalized Linear Discriminant Analysis (PLDA), which is based on clusters of variables, to construct a classification rule (Poterie, Dupuy, Monbet, & Rouviere, 2019). In certain scenarios, decision trees employ distinct splitting mechanisms, such as categorizing and uncategorizing child nodes, to enhance performance (Zeng & Chen, 2019). Furthermore, it is noteworthy that swarm intelligence algorithms posses the potential to explore the optimal parameters for multivariate linear splitting criteria in the process of decision tree building (Jariyavajee, Polvichai, & Sirinaovakul, 2019). Additionally, confidence intervals may be employed for appraising the gain that is associated with every split, and it is worth mentioning that more precise intervals result in the generation of more accurate decision tree (De Rosa & Cesa-Bianchi, 2015).

The utilization of the Gini impurity serves as a criterion in the construction of decision tress for attribute selection. It gauges the extent of impurity present within a grouping of vectors that necessitate partitioning into disparate clusters. The objective is to minimize the overall Gini impurity of the partition. The task of determining the partition with minimum weighted Gini impurity (PMWGP) poses a challenge and exhibits ties to the geometric k-means clustering problem (Laber & Murtinho, 2019; Sany Laber & Murtinho, 2018).

The utilization of entropy in decision trees serves as a metric in determining the ideal attribute to utilize for data splitting. Conventional metrics such as information entropy and Pearson's correlation coefficient are frequently employed, although their efficacy in handling uncertainty is limited. Deng entropy, which draws inspiration from the concept of Basic Belief Assignment (BBA), is posited as a measure for splitting rules in the classification of fuzzy datasets (Li, Xu, & Deng, 2019). An alternative approach proposes a novel algorithm for generating decision trees that approximates natural intelligence. This approach scrutinizes the relationship between entropy and knowledge, comparing the suggested algorithm with traditional methods (Popova, Popov, Karandey, & Gerashchenko, 2019). In the realm of packet classification, a heuristic method predicted on information entropy is suggested for constructing a balanced decision tree that takes into account time and space complexity (Dong, Meng, & Jiang, 2018). Another paper presents a statistical examination of decision tree learning algorithms based on different parametric entropies (Arellano, Bory-Reyes, & Hernandez-Simon, 2018). In conclusion, a fresh metric by the name of "dysconnectivity" has been put forth as a heuristic approach to the process of attribute selection during decision tree construction, with a specific focus on polymorphic attributes (Wang, 2011).

Information gain is a method used in decision tree algorithms to select attributes that are most information for classification. It measures the reduction in entropy or impurity achieved by splitting the data based on a particular attribute. The attribute with the highest information gain is chosen as the splitting criterion at each node of the decision tree. This helps in creating subsets that are pure as possible in terms of class labels, leading to more accurate classification. Information gain is used in various domains, such as intrusion detection systems (De Sousa, Veiga, De Oliveira Albuquerque & Giozza, 2022), COVID-19 surveillance (Ivandari, Maulana & Karomi, 2022).

Methodology

A narrative synthesis was utilized in this review to present findings from multiple studies which heavily relied on words and text to explain the significant information derived from them (Limna, 2022). The researchers employed a purposive sampling to narrow the coverage of the sources and only include the relevant studies which may contribute to the better result and deeper understanding of this endeavor. Employing content analysis, the researchers were able to objectively described and quantified the data from various sources. Only articles from peer-reviewed articles were included in this paper ranging from 2015 to 2023 year only.

Results and Discussion

The discourse surrounding the Classification Tree Algorithm and its import in the realm of machine learning offers noteworthy discernments into the weightiness of this approach in the classification of data. The classification tree algorithms form the bedrock of allotting fresh instances to predetermined categories on the basis of their attributes. The aforesaid algorithms have undergone enhancements over the course of time, whereby adaptations such as the Direct Nonparametric Predictive Interface (D-NPI) and Tree Penalized Linear Discriminant Analysis (TPLDA) have endeavored to augment precision, comprehensibility, and variable selection.

The utilization of classification tree algorithms is prevalent across a multitude of disciplines, ranging from medical disease analysis to text and image classification. There is a persistent pursuit among researchers to discover fresh tree-based methodologies, such as TPLDA and the amalgamation of quantum classifiers, in order to overcome constraints and enhance efficacy. Furthermore, novel techniques, including non-parametric kernel functions and modularity-based hierarchical classification trees (MHCT), have surfaced as innovative approaches to achieve precision and efficiency while simultaneously curtailing the duration of training.

The significance of classification algorithms in the realm of machine learning is underscored, as they occupy a central position in the process of categorizing data, facilitating informed prognostications, and supporting decision-making across various domains. Instances of their application include anticipatory analysis of student employability, identification of skin lesions, appraisal of student performance, and classification of faults in integrated circuits. Classification algorithms furnish users with the capability to extract insightful information and arrive at well-informed decisions from intricate datasets.

The discourse proceeds to delve deeper into the fundamental constituents of classification tree algorithms, which comprise nodes, branches, leaves, and the root node, expounding on their respective functions in the classification process. These constituents collectively bestow upon the algorithm its efficacy in accurately categorizing instances.

The segment concerning splitting criteria and measures in classification tree algorithms offers a thorough investigation of the methodologies employed in identifying the optimal attribute for data division. Various techniques, including similarity formulas, Penalized Linear Discriminant Analysis (PLDA), and swarm intelligence algorithms, are deliberated upon. Additionally, the relevance and exploration of Gini impurity and entropy as attributes for selection are discussed, as well as their effectiveness in addressing uncertainty and optimizing decision tree construction. Furthermore, the significance of information gain as a critical aspect of decision tree algorithms in diverse application domains is highlighted.

In brief, the present discourse provides an all-encompassing survey of the significance of the Classification Tree Algorithm in the realm of machine learning, encompassing its development, utilization, fundamental constituents, as well as the complexities entailed in the criteria and measures of splitting. These discernments accentuate the algorithm's pivotal function in the classification of data and its persistent advancement towards surmounting diverse challenges and augmenting its applicability across multiple domains.

Conclusion

In summary, it can be asserted that the Classification Tree Algorithm occupies a fundamental position within the domain of machine learning, fulfilling a pivotal function in the classification of data across a diverse range of domains. This discourse has provided illumination with regards to the algorithm's progression, highlighting its versatility through ingenious variations such as TPLDA, quantum classifiers, non-parametric kernel functions, and modularity-based hierarchical classification trees (MHCT). Furthermore, it has emphasized the algorithm's importance in facilitating informed decision-making, as evidenced by its employment in prognosticating student employability, medical diagnoses, performance appraisal, and defect classification in integrated circuits. Additionally, a comprehensive examination of the elements comprising classification trees, including nodes, branches, leaves, and root nodes, has illuminated their respective roles in accomplishing precise categorization.

Moreover, the discourse has explored the intricate domain of splitting criteria and measures, accentuating manifold techniques employed to optimize attribute selection, ranging from similarity formulas and PLDA to swarm intelligence algorithms. Gini impurity, entropy, and information gain have surfaced as pivotal instruments in the quest for data partitioning and the establishment of efficient decision trees. With the continuous advancement of machine learning, the Classification Tree Algorithm persists as an indispensable tool, constantly evolving to address emerging challenges and expand its applicability. Its enduring significance in the field emphasizes its role as a fundamental element in the ever-expanding realm of artificial intelligence and data analysis.

Conflict of interests

The researchers declare no conflict of interest.

References

Abdulmajeed, A.A., Coolen, F.P.A., & Coolen-Maturi, T. (2021). Direct Nonparametric Predictive Inference Classification Trees. *arXiv: Methodology*. https://doi.org/10.48550/arXiv.2108.11245

Arellano, A. R., Bory-Reyes, J., & Hernandez-Simon, L. M. (2018). Statistical Entropy Measures in C4.5 Trees. *International Journal of Data Warehousing and Mining, 14*(1), 1–14. https://doi.org/10.4018/ijdwm.2018010101

Poterie, A. Dupuy J.F., Monbet, V., Rouviere, L. (2019). Classification tree algorithms for grouped variables. Retrieved from https://hal.science/hal-01623570v1/file/classificationtreealgorithmsforgroupedvariables_apoterie_jfdupuy_vmonbet_lrouviere.pdf

Charbuty, B., & Abdulazeez, A. M. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01), 20–28. https://doi.org/10.38094/jastt20165

Chengwei, G., Bofeng, Z., Xinyue, W., Mingqing, H., & Guobing, Z. (2016). *The modularity-based Hierarchical tree algorithm for multi-class classification*. Proceedings from 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). https://doi.org/10.1109/SNPD.2016.7515969

Dai, Q., Zhang, C., & Wu, H. (2016). Research of Decision Tree Classification Algorithm in Data Mining. *International Journal of Database Theory and Application, 9*(5), 1–8. https://doi.org/10.14257/ijdta.2016.9.5.01

De Rosa, R., & Cesa-Bianchi, N. (2015). *Splitting with confidence in decision trees with application to stream mining*. Proceedings from International Joint Conference on Neural Networks (IJCNN), Killarney Ireland. https://doi.org/10.1109/IJCNN.2015.7280392

De Sousa, M. S., Veiga, C. E. L., De Oliveira Albuquerque, R., & Giozza, W. F. (2022). *Information Gain applied to reduce model-building time in decision-tree-based intrusion detection system*. Proceedings from 17th Iberian Conference on Information Systems and Technologies (CISTI). https://doi.org/10.23919/cisti54924.2022.9820579

Dong, X., Meng, Q., & Jiang, R. (2018). Packet classification based on the decision tree with information entropy. *The Journal of Supercomputing, 76*(6), 4117–4131. https://doi.org/10.1007/s11227-017-2227-z

Hongthong, T., & Temdee, P. (2022). *The classification-based machine learning algorithm to predict* *students' knowledge levels*. Proceedings from Joint International Conference on Digital Arts, Media and Technology With ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT &Amp; NCON). https://doi.org/10.1109/ectidamtncon53731.2022.9720334

Ivandari, I., Maulana, M. R., & Karomi, M. a. A. (2022). Improved Decision Tree Performance using Information Gain for Classification of Covid-19 Survillance Datasets. *JAICT (Journal of Applied Information and Communication Technologies), 7*(1), 74. https://doi.org/10.32497/jaict.v7i1.3501

Jariyavajee, C., Polvichai, J., & Sirinaovakul, B. (2019). *Searching for Splitting Criteria in Multivariate Decision Tree Using Adapted JADE Optimization Algorithm*. Proceedings from IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 2534-2540. https://doi.org/10.1109/SSCI44817.2019.9003063

Ji, N., Bao, R., Mu, X., Chen, Z., Yang, X., & Wang, S. (2023). Cost-sensitive classification algorithm combining the Bayesian algorithm and quantum decision tree. *Frontiers in Physics*, *11*. https://doi.org/10.3389/fphy.2023.1179868

Khandelwal, J., Pareek, G., Dey, R., & Pareek, S. (2023). *The Study of Machine Learning Classification Algorithm for Student Placement Prediction*. Proceedings from 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON). https://doi.org/10.1109/iemecon56962.2023.10092294

Laber, E. S., & Murtinho, L. (2019). Minimization of Gini Impurity: NP-completeness and Approximation Algorithm via Connections with the k-means Problem. *Electronic Notes in Theoretical Computer Science, 346,* 567–576. https://doi.org/10.1016/j.entcs.2019.08.050

Li, M., Xu, H., & Deng, Y. (2019). Evidential decision tree based on belief entropy. *Entropy*, *21*(9), 897. https://doi.org/10.3390/e21090897

Limna, P. (2022). Artificial Intelligence (AI) in the Hospitality Industry: A Review Article. *International Journal of Computing Sciences Research*, *6*, 1-12.

Özcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*, *3,* 100130. https://doi.org/10.1016/j.health.2022.100130

Popova, O., Popov, B., Karandey, V., & Gerashchenko, A. (2019). Entropy and algorithm of obtaining decision trees in a way approximated to the natural intelligence. *International Journal of Cognitive Informatics and Natural Intelligence*, *13*(3), 50–66. https://doi.org/10.4018/ijcini.2019070104

Poterie, A., Dupuy, J., Monbet, V., & Rouvière, L. (2019). Classification tree algorithm for grouped variables. *Computational Statistics*, *34*(4), 1613–1648. https://doi.org/10.1007/s00180-019-00894-y

Rajeshkanna, A., & Arunesh, K. (2021). *Optimizing Decision Tree Classification Algorithm with Kernel Density Estimation*. In Springer eBooks. https://doi.org/10.1007/978-981-15-9651-3_22

S, H. M., Raman, S., Sanjay, P., Latha, S., Muthu, P., & Dhanalakshmi, S. (2022). *Skin Lesion Classification using Machine Learning Algorithm for Differential Diagnosis*. Proceedings from 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS). https://doi.org/10.1109/ictacs56270.2022.9987971

Song, T., Huang, Z., & Yan, A. (2022). Machine learning classification algorithm for VLSI test cost reduction. *Integration, 87,* 40–48. https://doi.org/10.1016/j.vlsi.2022.06.005

Wang, Z. (2011). 'Entropy' on Covers and its application on Decision Tree Construction. *International Journal of Machine Learning and Computing*, 213–217. https://doi.org/10.7763/ijmlc.2011.v1.31

Zaim, H., Ramdani, M., Haddi, A. (2018). Splitting Method for Decision Tree Based on

Similarity with Mixed Fuzzy Categorical and Numeric Attributes. In: Tabii, Y., Lazaar, M., Al Achhab, M., Enneya, N. (eds) *Big Data, Cloud and Applications*. BDCA. https://doi.org/10.1007/978-3-319-96292-4_19

Zeng, H., & Chen, A. (2019). *Classification Tree with Hybrid Splitting Mechanism.* Proceedings from IEEE 17th International Symposium on Intelligent Systems and Informatics (SISY) (pp. 61-66). Subotica, Serbia. https://doi.org/10.1109/SISY47553.2019.9111639