An Improved Decision Tree Algorithm for Text Classification and Visualization

IJIMSR, Vol.1, No.1, (2023) 23.1.1.003

K. Srinivas*
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
500075, India
Email: srirecw9@klh.edu.in
Optimistic Decision Tree Algorithm for Text Classification and Visualization

G. Madhukar Rao
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
500075, India
Email: madhusw511@klh.edu.in
Optimistic Decision Tree Algorithm for Text Classification and Visualization

***Corresponding author**

Received 29th August, 2023; Accepted 8th September, 2023

Keywords: Big Data Analytics, Data Mining, Decision Tree, Machine Learning, Predictive Analysis

An Improved Decision Tree Algorithm for Text Classification and Visualization

K. Srinivas
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, 500075, India
Email: srirecw9@klh.edu.in
Optimistic Decision Tree Algorithm for Text Classification and Visualization

G. Madhukar Rao
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, 500075, India
Email: madhusw511@klh.edu.in
Optimistic Decision Tree Algorithm for Text Classification and Visualization

ABSTRACT

For Big Data, the term "machine learning" must be reinvented. In this work, we give a survey on machine learning for large data processing in ongoing reviews, as well as a recommended approach for evaluating huge data. This examination provides a point of view on the area, identifies research gaps and opportunities, and provides a reliable foundation and encouragement for further research in the field of machine learning with Big Data through this process. We looked at various data kinds, learning methodologies, important difficulties in big data management, and the usage of machine learning algorithms in big data. To increase prediction accuracy, we suggested a decision tree merging process and an ODTA (Optimistic Decision Tree Algorithm) approach in this paper.

Keywords: Big Data Analytics, Data Mining, Decision Tree, Machine Learning, Predictive Analysis
Received 29th August, 2023; Accepted 8th September, 2023

BACKGROUND

Big Data (BD) is a word for a group of datasets so complex and huge that it gets hard to work using conventional data processing applications or available DB management tools. The difficulties join visualization capture, storage, curation, search, sharing, analysis, and transfer. The pattern to bigger data sets is required to the conventional data derivable from analysis of a single related data when contrasted with isolated smaller sets with a similar total sum of information, permitting connections to be found to spot business trends, choose the nature of research, forestall infections, connect legitimate citations, battle wrongdoing, and choose constant street traffic conditions. The term "Big Data" refers to the growth and application of technologies that provide the appropriate customer at the right time with the correct data from a massive amount of data that has been growing exponentially in our population for a long time. The difficulty is not only to manage the rapid expansion of data volumes, but also to manage progressive, varied enterprises as well as increasingly interconnected and sophisticated data. According to Madhukar Rao et al. (2016) and Lin et al. (2016), BD refers to large amounts of challenging data, both unstructured and structured, that traditional processing approaches or possibly algorithms can't handle (2019). Its goal is to expose hidden models, and it has resulted in a move from a scientific to a scientific worldview.

1.2 BIG DATA ANALYTICS

In general, Big Data is a term used to describe data that exceeds the standard storage, processing, and calculating the limits of conventional database systems Large-scale data models demand strategies and tools to separate and eliminate them. Because of the speed and variety of the data handled, consider structured data developing. Therefore, it is no longer sufficient to look at reports and statistics, as the vast amount of information available requires that the systems in place be capable of assisting in the analysis of the information.

Figure 1. Big Data Analytics

Big Data is data that exceeds the limits of typical databases and data analysis methodologies in terms of storage, processing, and Tools and approaches are needed to separate patterns in data at scale using big data as a resource. This is due to the variety and speed of structured data. This means that looking at data and producing reports is no longer enough, as there are so many different types. The system must be beneficial in helping with data analysis.

1.2.1 Big Data Analytics Tools for ML Algorithm

Apache Hadoop and Spark Most machine learning procedures can be paralleled by understanding an estimation technique for every data, Jordan et al. (2015). Such procedures can be clarified with the help of a MapReduce activity course: the data are divided into parts, each part is treated as equivalent, and the results are accumulated in the arrangement. Apache Hadoop is a broadly used open-source structure in Java for such purposes. Clients can send along with no other person clusters or server farms or can discover a solution for organizations using the Hadoop platform. Taking Hadoop into account, Apache Spark is a response striving to refocus execution by focusing specifically on features such as machine learning, diagram analysis, and data streaming, and programmable in Scala or Python. Apache Spark is a valuable and engaging analytics framework. It improves productivity through primitive memory registration, Pipelined computer and it improves usability through APIs in Scala, but Java, R API, and Python also introduced and works through Interactive Shell.

2. RELATED WORKS

With the development of innovation, data can be digitized into data and processed and dissected by a PC. In some fields, such as the Internet, money, and medication, numerous data sets can c reate numerous records every day. Moreover, products such as smart devices distribute multiple sensors, which thus create a great deal of data. Besides, the development of networks makes data storage more convenient. Numerous industries direct increasingly more analysis and processing of data in the request to acquire useful data from big data, to mine the useful data, and receive it. Consequently, the new development of a big data analysis calculation becomes the current focus. For the decision tree calculation, the big data acquisition features contain an enormous amount of sum; there is an enormous quantity of excess, and even some second-order quality or values of irrelevant components. The processing of them consumes a ton of registering resources as well as results in the oversize tree structure. It also affects the accuracy of one's predictions. In this way, it is a significant means of enhancing the ability and accuracy of machine learning to look at the characteristics of the datasets and discover the characteristics that meet the requirements.

2.1 Decision Tree: General Working

A decision tree is a tree structure used to classify instances. A decision tree is made up of nodes and oriented edges. There are two **kinds of nodes: the inner nodes and the leaf nodes.** When an internal hub represents a test condition for a component or property (used to separate records with different features), a leaf hub represents a classification, Wang et al. (2013). Once a decision tree template has been constructed, it will be very easy to classify it on its basis, Strecht et al. (2015) and Anyanwu et al. (2009). Figure 2 shows a basic structure of Decision Tree.

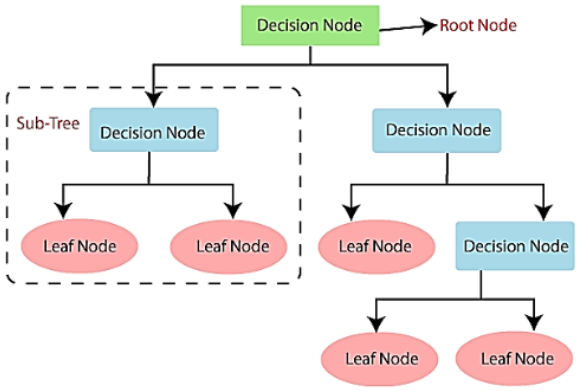

Figure 2. Basic Decision Tree Structure

To do this, start with the root hub, test an element of the nearby instance, and assign the instance to its youngsters (that is, select the proper branch) as indicated by the test structure. At the point when this branch may arrive at a leaf hub or another internal hub, the new test condition is used to recursively execute until it reaches a leaf hub. At the point when we arrive at the leaf hub, we get the last classification result in Pandey et al. (2013) and Naveen et al. (2015). A decision tree is used for predictive analysis when a forecast is performed by the construction of a decision tree with test points and branches, Harsh et al. (2018) and Rao et al. (2020). Attestation points, a decision is made to choose a classifier to choose a specific branch and browse the tree to arrive at the last goal of the decision. They can be categorized as classification trees and regression trees, when applied to earlier situations. When the yield variables are absolute, such as yes or no, while the latter applies to the yield variables that are numeric or continuous, such as the anticipated cost of goods, Srinivas et al. (2020). It introduces decision-making rules that are very simple. No hypothesis of a hidden relationship has been considered.

3. PROPOSED METHODOLOGY

The utility of a decision tree arises from its independence from specialized expertise during the design phase. Furthermore, it effectively manages higher dimensions without introducing intricate or time-intensive procedures into the learning process of the decision tree. Users can easily comprehend the tree's representation of accumulated knowledge, as the steps involved are neither complex nor time demanding. The precision exhibited by decision tree classifiers is truly remarkable. In the work by Meng et al. (2016), the "ODTA model" (Optimistic Decision Tree Algorithm) employs a four-step tree merging approach for implementation. Unlike traditional extension methods embraced by many other decision tree algorithms ODTA uses bootstrap statistical methodology, which does not rely on the unique data structure. Prior to ODTA, decision tree algorithms employed a technique involving scanning the database once for every level of the tree, as demonstrated by Sabah et al. (2019). However, when integrated with ODTA, this approach enhances both data collection and storage processes.

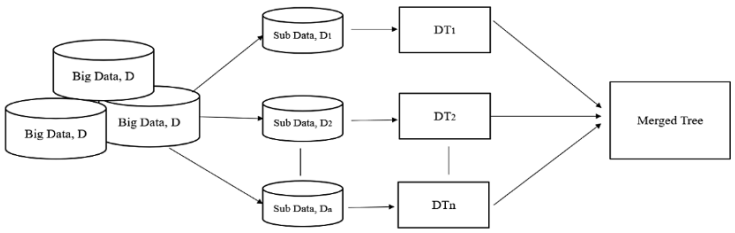

Figure 3. Proposed architecture model of Optimistic Decision Tree Algorithm (ODTA) model.

Utilizing ODTA, we gain the capability to incorporate new elements and modify the decision tree without the need to recreate the entire structure from scratch. The ODTA algorithm initiates by segmenting the raw data into more manageable subsets, each of which subsequently produces a separate tree, resulting in multiple decision trees. To recreate a tree similar to one that would have been generated if all initial preparatory data were stored in memory, an evaluation of each tree becomes essential. The foundational framework of the ODTA algorithm is depicted in Figure 3. Operating as a recursive refinement algorithm, the distinctive innovation of this approach lies in its continuous partitioning of heterogeneous sets into progressively more homogenous subsets. The algorithm's pivotal advancement lies in the persistent process of transforming non-uniform segments into subsets as uniform as possible. The individual steps of this algorithm are shown in Figure 4.

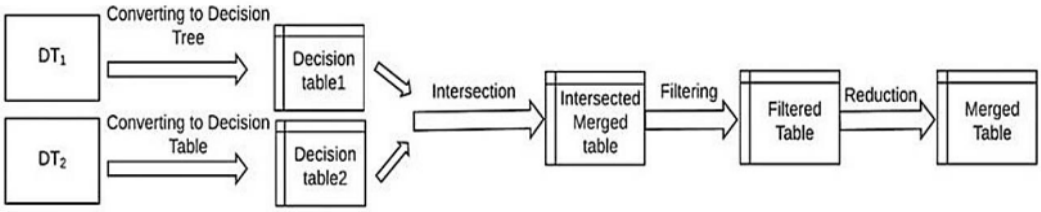

Figure 4. Proposed Decision Tree Merging Process

Decision trees find it simpler to make decisions based on hierarchy. Decisions related to hierarchy are more readily established using decision trees. This is attributed to the enhanced visibility into the modification of each criterion. Additionally, transitioning to rules eliminates the distinction between character tests conducted near the tree's root and those carried out near the leaves.

Input:

- It's a collection of drive tuples and their class tags.
- Candidate attributes are included in the attribute table below.
- To determine the fractionation criterion that "best" separates specific classes, use the Attribute selection method technique

4. EXPERIMENTAL RESULTS

4.1 Comparison between Optimistic Decision Tree Algorithm and Rainforest Algorithm

We've compiled seven datasets sourced from the Keel and UCI machine learning repositories. These datasets underwent pre-processing steps to handle missing values, eliminate duplicate instances, and select relevant features. In this section, we present the results of our refined decision tree in comparison to an idealistic decision tree. The findings, as depicted in Table-6, indicate that as complexity increases from 0% to 80%, the results obtained from the datasets demonstrate the potential to reduce the training set size by up to 90% without significantly impacting the classification outcomes. This observation highlights that an increase in the number of attributes leads to concurrent improvements in both accuracy and complexity. Because of these observations, we have introduced the following acronyms: TP for true positive, TN for true negative, FP for false positive, and FN for false negative. The Tropical Rainforest and ODTA algorithms are presented separately, with the CART and ODTA algorithms illustrated in Tables 2 and 3, respectively.

Table 1: Dataset Descriptions

S.No	DataSets	Resources	Number of Attributes	Class
1	Austrian	692	12	2
2	Mammo-graphic	963	4	2
3	Diabetes	767	7	2
4	BreastCancer	563	25	2
5	Thyroids	7245	24	3
6	Spam base	489	56	2
7	Monk-2	424	5	2

Back then, a sizable dataset existed, but the available memory was limited. To tackle this challenge, a novel approach was proposed to handle the manageability of information-rich decision trees. All decision tree algorithms used in RainForest are designed to adjust the main memory's capacity. The decision tree algorithms within RainForest are structured in a way that emphasizes adaptability in shaping the tree's nature. This approach doesn't mandate utilizing the entire dataset, as demonstrated by Thangaparvathi and colleagues (2010). Figure 5 shows the performance of rain forest algorithm and ODTA on different datasets.

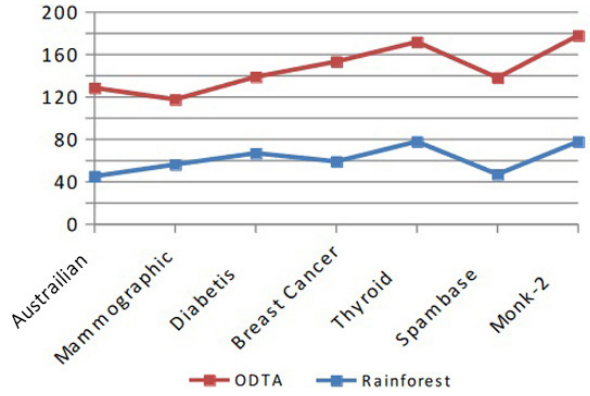

Figure 5. Performance of Rainforest Vs ODTA

This can be divided into subsections if several methods are described [2].

4.2 Comparison between Optimistic Decision Tree Algorithm and CART (Classification and Regression Trees)

Decision tree algorithms are employed to divide attributes for testing at each node, determining whether the split is "Optimal" for individual classes. The resulting partitions at each branch strive for maximum purity, necessitating that splitting patterns remain consistent. Based on the values of the splitting criterion, the training data is divided into multiple subsets. The algorithm continues recursively until all instances within a subset belong to the same class in any given decision tree.

Table 2: CART (Classification and Regression Trees)

Datasets	Classification Accuracy (%)	Precision Weighted Average	Recall Weighted Average	F-Score Weighted Average
Australian	56.34	0.42	0.36	0.37
Mammo-graphic	79.87	0.37	0.43	0.43
Diabetes	52.32	0.74	0.39	0.52
Breast Cancer	75.4	0.56	0.42	0.29
Thyroids	56.4	0.44	0.48	0.48
Spam base	48.98	0.58	0.35	0.39
Monk-2	25	3	6	5

Information gain suffers from the limitation of being biased towards univariate attributes, which constitutes its primary drawback. Another drawback arises from the uneven distribution of data, often favoring one child node with more entries than the other-a preference that Gain Ratio usually exhibits. Additionally, the Gini Index yields unfavorable outcomes when dealing with datasets comprising more than two categories. These constitute the limitations of the partitioning criteria.

Table 3: Optimistic Decision Tree Algorithm

Datasets	Classification Accuracy (%)	Precision Weighted Average	Recall Weighted Average	F-Score Weighted Average
Australian	88.65	0.76	0.45	0.49
Mammo-graphic	84.54	0.48	0.56	0.43
Diabetes	66.45	0.88	0.36	0.47
Breast Cancer	79.86	0.69	0.73	0.56
Thyroid	62.54	0.56	0.82	0.48
Spam base	56.72	0.67	0.45	0.57
Monk-2	38	5	8	12

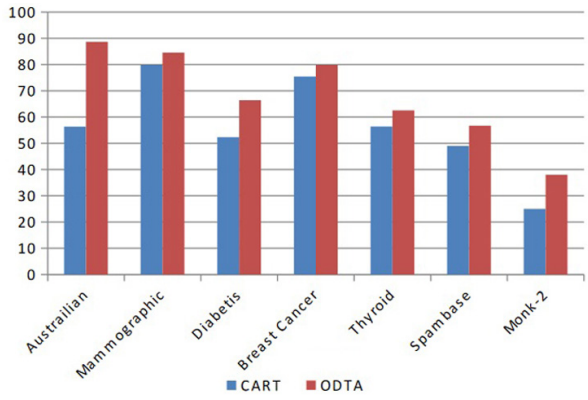

Figure 6: Performance of CART Vs ODTA

CONCLUSION

The realm of automated learning encompasses various methodologies, including the utilization of decision trees. Decision tree algorithms, when employed independently, have found application across diverse disciplines such as statistical replacement processes, text extraction, Australian Mammographic Diabetes, Breast Cancer, Thyroid, Spam Base, Monk-2, and others. Clinical validation fields and research robotics also stand to gain from these algorithms. Multiple decision tree algorithms have been devised, tailored to optimize accuracy and cost-effectiveness. This endeavor's primary objective is to establish a foundation for a comprehensive comprehension of machine learning within the context of Big Data. Extensive scrutiny has been directed toward decision tree algorithms, with this study specifically assessing the efficacy of ODTA using varied datasets. We hold the belief that our approach stands as one of the most accomplished in this domain.

REFERENCES

[1] Madhukar Rao, G., & Ramesh, D. (2016). Supervised learning techniques for big data: a survey. IJCTA. Int Sci Press, 9, 3811-3891.

[2] Wang, C. H., Zhou, L., Jiang, F., & Zhao, H. B. (2013). A Granular Computing Based Decision Tree Algorithm and its Application in Intrusion Detection. In Applied Mechanics and Materials (Vol. 268, pp. 1730-1734). Trans Tech Publications Ltd.

[3] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[4] Strecht, P. (2015, January). A survey of merging decision trees data mining approaches. In Proc. 10th Doctoral Symposium in Informatics Engineering (pp. 36-47).

[5] Anyanwu Matthew, N., & Shiva Sajjan, G. Comparative Analysis of Serial Decision Tree Classification Algorithms.

[6] Pandey, M., & Sharma, V. K. (2013). A decision tree algorithm pertaining to the student performance analysis and prediction. International Journal of Computer Applications, 61(13).

[7] Thangaparvathi, B., & Anandhavalli, D. (2010, October). An improved algorit hm of decision tree for classifying large data set based on rainforest framework. In 2010 International Conference On Communication Control And Computing Technologies (pp. 800-805). IEEE.

[8] Navin, G. V. (2015). Big Data Analytics for Gold Price Forecasting Based on Decision Tree Algorith m and Support Vector Regression (SVR). International Journal of Science and Research (IJSR).

[9] Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering, 6(10), 74-78.

[10] Rao, G. M., Ramesh, D., & Kumar, A. (2020). RRF-BD: ranger random forest algorithm for big data classification. In Computational Intelligence in Data Mining (pp. 15-25). Springer, Singapore.

[11] Meng, Q., Ke, G., Wang, T., Chen, W., Ye, Q., Ma, Z. M., & Liu, T. Y. (2016). A communication efficient parallel algorithm for decision tree. arXiv preprint arXiv:1611.01276.

[12] Srinivas, K., Rao, G. M., Vengatesan, K., Tanesh, P. S., Kumar, A., & Yuvaraj, S.(2020). An implementation of subsidy prediction system using machine learning logistical regression algorithm. Advances in Mathematics: Scientific Journal, 9(6), 3407-3415.

[13] Maraca, A. L., Casanova, D., & Teixeira, M. (2019). Assessing classification complexity of datasets using fractals. International Journal of Computational Science and Engineering, 20(1), 102-119.

[14] Lin, W., Zhang, Z., & Peng, S. (2019). Academic research trend analysis based on big data technology.

[15] Sabah, S., Anwar, S. Z. B., Afroze, S., Azad, M. A., Shatabda, S., & Farid, D. M. (2019, August). Big Data with Decision Tree Induction. In 2019 13th International Conference on Software, Knowledge,