Decision Tree using Feature Grouping

Neamul islam Fahim*, Md. Awinul Haque Utsha*, Raj Shekhar Karmaker*, Md Oli Ullah*, and Dewan Md. Farid*
Department of Computer Science and Engineering, United International University,
United City, Madani Avenue, Badda, Dhaka 1212, Bangladesh
Email: {nfahim202216, mutsha202163, rkarmaker193149, moliullah142060}@bscse.uiu.ac.bd, dewanfarid@cse.uiu.ac.bd

***Abstract*—Machine learning (ML) is a crucial junction of statistics and computer science in the quickly developing field of artificial intelligence. In particular, when conventional algorithms fail, this research investigates the creation of algorithms that give priority to data-driven insights over rigid instructions. Among the many tools that ML provides, decision trees and random forests have drawn attention for their accuracy in making decisions. However, the introduction of the random forest technique, an ensemble approach that combines predictions from various trees for improved accuracy, was prompted by the limits of individual decision trees in handling large datasets and complicated relationships. A innovative data categorization method based on the idea of feature correlation is at the heart of our study. We found discrete clusters and created a decision tree specifically for each by analyzing the interactions between dataset elements, resulting in detailed and accurate classifications. A thorough analysis of numerous datasets showed that our technique consistently outperformed more traditional approaches. This work essentially emphasizes the transformational potential of ML in data classification, highlighting the ability of feature correlation to improve accuracy and efficiency in the constantly evolving field of artificial intelligence.[10]**

***Index Terms*—Keyword1;random forest Keyword2; decision trees**

I. Introduction

In the study of artificial intelligence, machine learning combines statistics and computer science to create algorithms that perform better when exposed to relevant data as opposed to particular instructions[1]. When it is challenging or impractical to develop conventional algorithms to implement necessary functionalities, machine learning techniques are used instead. [3]The search for accurate and efficient decision-making tools has sparked the development of numerous algorithmic techniques in the quickly developing field of machine learning. These technologies are essential for turning raw data into usable insights since they are made to process, analyze, and interpret enormous amounts of data. Decision trees and random forests have stood out among the profusion of tools as particularly effective ones, especially in the areas of data classification and decision-making. As their name implies, decision trees are tree-like models of decisions and potential outcomes. In large part because of their clear visual depiction, they have established themselves as a mainstay in data classification tasks. The probable results of each decision or test on an attribute are represented by a node, and these nodes branch out to reflect subsequent decisions or ultimate classifications. Knowledge is successfully contained by this hierarchical structure, which also makes it interpretable and usable. Decision trees' visual nature enables stakeholders, including those without a strong technical background, to follow a decision's development and comprehend its logic. The Decision Tree Classifier is one of the most well-known uses of decision trees in multistage decision-making settings. Despite the fact that decision trees are unquestionably effective, their simplicity occasionally has drawbacks. A single decision tree could find it difficult to fully represent the breadth and depth of the data when faced with large datasets that are characterized by complex connections and relationships. The random forest approach excels in these difficult settings. The fundamental ideas of decision trees are built upon by the ensemble machine learning technique known as random forests. Random forests produce several decision trees, each trained on a random subset of the data, as opposed to relying on a single tree. The algorithm aggregates the findings from each of these trees when making predictions, frequently producing outcomes that are more reliable and accurate than those from a single tree. By utilizing the strengths of various trees, this ensemble approach ensures that the model gains from a variety of viewpoints and reduces the biases of individual trees. Random forests perform well, especially in situations with plenty of variables. They combine the insights from various trees while maintaining the decision trees' natural branching logic, providing a more thorough grasp of the material. . Random forests perform well in scenarios with lots of variables because they maintain the natural branching of decision trees while also benefiting from the combined wisdom of several trees[5]

In research, we presented an original approach made just for data classification. The idea of feature correlation serves as the system's cornerstone. We were able to locate separate clusters by looking at how different data set properties interacted with one another. The degree of association between the features affects how many of these clusters there are. We created a decision tree for each cluster that was found. With this method, classification can be done more precisely and specifically because each decision tree is made to fit the specifics of the related feature cluster.We put our suggested system through extensive testing using various data sets to verify its effectiveness. When compared to traditional classification approaches, the results showed continuously improved performance indicators, which was encouraging. This emphasizes how effective our approach may be as a tool for data classification, utilizing the strength of feature correlation to increase efficiency and accuracy. I hope the additional content is still true to the core of your study. If you need any other changes, please let me know!

II. Related Works

In numerous machine learning and data mining tasks, a decision tree and random forest was utilized as a classifier. This study examines a number of recent DT and random forest related works.

Chinese hospitals in Fuzhou. There are 14 qualities involved. Data from 68994 is randomly selected by the training array. Diabetic patients and stable humans, respectively. They made use of the full meaning of minimal Reliability Maximum Principal Component Analysis and Relevance (Marmara) reduce the number of dimensions. When it comes to RF's impacts, appeared to be higher than the other when compared to each other classifiers. Additionally, 0.8084 is the Luzhou data's best result collection.[14]

Assegie and Nair classified the handwritten digits from the standard data set of kaggle digits using the DT classification technique and estimated the model's accuracy for each digit from 0 to 9. For machine learning, the kaggle features include 42,000 rows and 720 columns; vector features are utilized for pixels in digital images. They applied machine learning methods to map the classifier's success rate graph in the realization of handwritten digits using a very effective language called "python programming." suggested that the decision tree classifier and 83.4.[2]

Prior to therapy, De Felice et al. proposed a decision tree approach to identify both established and new clinical indications for survival in locally advanced rectal cancer (LARC). The analyses demonstrated that even non-specialists in the subject can easily grasp the tree-based machine learning process, particularly when it comes to categorization trees. To even reach their statistical potential, validation mistakes must be controlled. Patients with LARC who had been histologically proven had their records examined between 2007 and 2014. To calculate overall survival (OS), the Kaplan-Meier method was utilized. 100 patients in all were involved. The 5-year and 7-year OS points were respectively 76.4[4]

A context-aware behavioral decision tree called "BehavDT" was introduced by Sarker et al. It accounts for generalizations about consumer behavior based on the degree of individual choice. The BehavDT approach offered both comprehensive and context-specific conclusions in uncommon association scenarios. Through the effectiveness of the BehavDT model, experiments were conducted on actual smartphone datasets of individual users. The results showed that, in comparison to other traditional machine learning models, the Behav DT context-aware model is the one that is most energetic, with an accuracy rate of up to 90[11]

The first workable approach for decision tree optimization for binary variables was illustrated by Hu et al. A dedicated bit vector library, data structures that reduce the search area, and modern application technologies were included in the algorithm's co-design to meet analytical restrictions. They evaluated the accuracy and contrasted it with the Optimal Sparse Decision Trees (OSDT) approach using the Binary Optimal Classification Trees (BinOCT) method, which is the most recent publicly available method. Additionally, they used numeric datasets from the other ProPublica COMPAS datasets as well as text datasets from the University of California, Irvine (UCI) Machine Learning Repository. The results demonstrated that when a COMPAS dataset was used, OSDT's best decision tree produced an accuracy of 66.90In addition, when BinOCT and OSDT produced the UCI dataset,[6].

Dimensionality reduction is used when the majority of features are thought to be unimportant (in relation to the task at hand), when human interpretability is needed, or to speed up subsequent algorithms. Here, we are less concerned with these factors and more concerned with how to include features to increase learnability all around. The image processing community has recently focused on dimensionality expansion techniques, which use sparse encoding in high-dimensional spaces (also known as sparse-overcomplete representation) to improve the likelihood that image categories will be linearly separable (in a manner akin to the motivation behind kernels). [9]

The most common methods for adding or introducing new dimensions work by combining current variables in nonlinear ways to enhance prediction accuracy. The methods used range from simple combinations of all input variables using a family of algebraic operators (products, logs, etc.) to more complex selection techniques based on boosting. Contrary to existing approaches, here offer a method that is entirely non-parametric and more adaptable, meaning that it is not limited to producing fresh features as a pre-defined combination of a (limited) number of original features. The ability of our technique to be supervised, or guided by the learning task, is another significant difference.[7]

The most common methods for adding or introducing new dimensions work by combining current variables in nonlinear ways to enhance prediction accuracy. The methods used range from simple combinations of each input variable using a set of algebraic operators (products, logs, etc.) to more complex selection techniques based on boosting [22].

Contrary to existing approaches, we here offer a method that is entirely non-parametric and more adaptable, meaning that it is not limited to producing fresh features as a pre-defined combination of a (limited) number of original features. The ability of our technique to be supervised, or guided by the learning task, is another significant difference. [7]

III. Methodology

We have used Random Forest, C4.5 and our proposed algorithm on the balanced dataset

A. Dataset

Dataset: To conduct this research, we used five datasets, which are diabetes, heart disease, voice classification, typhoid, and Irish. The diabetes dataset has 768 rows with 9 instances; the heart disease dataset has 1025 rows with 14 instances; the voice classification dataset has 3168 rows with 21 instances; the typhoid dataset has 3772 rows with 30 instances; and the Irish dataset has 150 rows with 6 instances. The dataset was cleaned and balanced before training the models.

B. Random Forest

Breiman developed the powerful Ensemble Learning (EL) approach known as Random Forest (RF).[13] Becision tree technique is random forest, which works by building several decision trees[8]. The machine learning algorithm Random Forest is flexible. This approach is based on bootstrap aggregation, often known as bagging[12]. To increase generalization and prediction accuracy, it builds several decision trees during training and integrates their results. In order to prevent overfitting, each tree is constructed using a unique subset of data and features. Given its stability and proficiency with handling large, complex datasets, Random Forest is a good choice for classification and regression problems.

Pseudocode of Random Forest:

Input:

- Training dataset: $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$ where $\mathbf{x}_i$ is a feature vector and y_i is the corresponding label.
- Number of decision trees to be created: N_{trees}
- Number of features to consider for each split: $N_{features}$

Output: A Random Forest ensemble of decision trees.

C. C4.5

The well-known algorithm for decision tree induction is C4.5. It is utilized for categorization tasks in data mining and machine learning. In order to optimize information gain or gain ratio, C4.5 builds decision trees by recursively partitioning data based on the optimum attribute splits. For the purpose of ranking potential tests, C4.5 employs two heuristic criteria:

Information Gain (IG):

$$IG(A) = Info(D) - Info_A(D)$$

where: - $IG(A)$ is the information gain for attribute A. - $Info(D)$ is the total entropy of the dataset D. - $Info_A(D)$ is the weighted sum of entropies of subsets S_i after splitting on attribute A.

However, information gain is strongly biased in favor of tests with many outcomes. To address this, C4.5 introduces the Gain Ratio:

Gain Ratio (GR):

$$GR(A) = \frac{IG(A)}{SplitInfo(A)}$$

where: - $GR(A)$ is the gain ratio for attribute A. - $IG(A)$ is the information gain for attribute A. - $SplitInfo(A)$ is a value constructed analogously to $Info(D)$ and is used to normalize information gain.

C4.5 can handle both categorical and numerical data, and the resulting decision trees are interpretable, aiding in the identification of trends in the data and making predictions. The difference between the initial information requirement and the new requirement is known as information gain.

D. Decision Tree using Feature Grouping

In this study, we provide a brand-new classification scheme for data. We have built decision trees on each group after utilizing correlation to group features from the entire dataset into the desired number of clusters.We have successfully applied this technique to numerous datasets.

IV. Results and Discussion

A. Dataset Description

To conduct this research, we used five datasets, which are diabetes, heart disease, voice classification, typhoid, and Irish. The diabetes dataset has 768 rows with 9 instances; the heart disease dataset has 1025 rows with 14 instances; the voice classification dataset has 3168 rows with 21 instances; the typhoid dataset has 3772 rows with 30 instances; and the Irish dataset has 150 rows with 6 instances. The dataset was cleaned and balanced before training the models.

B. Experimental Setup

We used the Google Colab platform, which gave us the necessary computational resources and collaborative tools, to complete this research project. Python was one of our main tools for conducting this experiment, and it served as the basis for our work on data analysis and modeling. To create and assess our predictive models, we also made use of Scikit-learn, a powerful machine learning package written in Python. Additionally, to aid in the display of our results and discoveries, we made insightful graphs and charts using Matplotlib, a well-known data visualization library. We were able to rapidly investigate, analyse, and analyze the data thanks to the combination of these technologies and tools, which eventually made it possible for us to derive useful findings from our study.

C. Experimental Results

Sample tables and figures are given below.

TABLE I
PERFORMANCE METRICS FOR DIFFERENT ALGORITHMS ON VARIOUS DATASETS HAVING FULL SIZE

Dataset	Random Forest				Correlation Feature Grouping with Decision Tree				C4.5			
	Acc.	Prec.	Recall	F1-Score	Acc.	Prec.	Recall	F1-Score	Acc.	Prec.	Recall	F1-Score
Diabetes	0.69	0.75	0.45	0.56	0.70	0.71	0.70	0.70	0.70	0.65	0.70	0.67
Heart Disease	0.73	0.74	0.65	0.69	0.73	0.73	0.73	0.73	0.73	0.70	0.73	0.72
Voice Classification	0.92	0.99	0.87	0.92	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Typhoid	0.54	0.58	0.55	0.51	0.56	0.57	0.56	0.56	0.56	0.52	0.56	0.54
Iris	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

TABLE II
PERFORMANCE METRICS FOR DIFFERENT ALGORITHMS ON VARIOUS DATASETS HAVING 50% OF TOTAL SIZE

Dataset	Random Forest				Correlation Feature Grouping with Decision Tree				C4.5			
	Acc.	Prec.	Recall	F1-Score	Acc.	Prec.	Recall	F1-Score	Acc.	Prec.	Recall	F1-Score
Diabetes	0.60	0.57	0.71	0.63	0.69	0.70	0.69	0.70	0.69	0.77	0.75	0.76
Heart Disease	0.82	0.82	0.78	0.80	0.99	0.99	0.99	0.99	0.99	1.00	0.97	0.99
Voice Classification	0.96	0.96	0.97	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
Typhoid	0.56	0.57	0.56	0.56	0.54	0.54	0.54	0.54	0.54	0.53	0.51	0.52
Irish	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

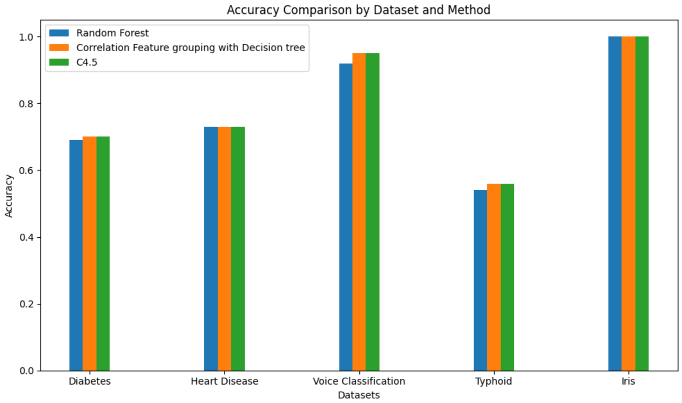

Fig. 1. Applying different Algorithm on full size dataset

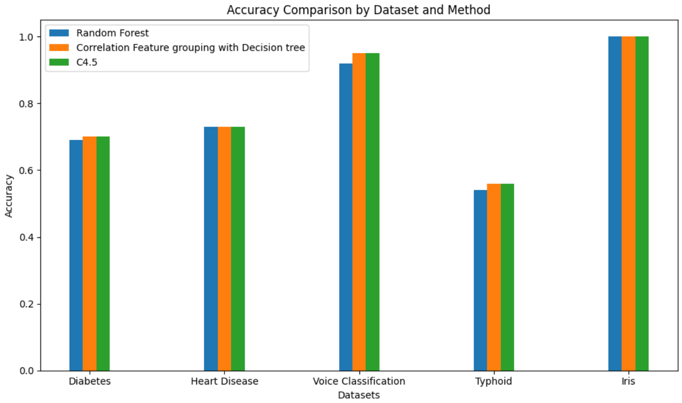

Fig. 2. Applying different Algorithm on 50% of total size of dataset

V. Conclusion and Future Work

Machine learning (ML), which bridges the gap between statistics and computer science, is at the forefront of the ongoing evolution of artificial intelligence. We have seen the emergence of algorithms using ML that rely more on data than strict instructions and provide fixes when conventional algorithms fail. Decision trees and random forests have established a niche for themselves among the plethora of tools offered by ML, proving vital for precision decision-making.

Our study delves into the complexity of data classification and introduces a fresh method based on feature correlation. We found separate clusters by analyzing how dataset features interact, and we then specifically tailored a decision tree for each to assure accurate classifications. Our system's thorough testing on a variety of datasets demonstrated its superiority to traditional approaches. The steadily rising performance metrics highlight the approach's promise.

Our study essentially confirms the transformational potential of machine learning for data classification. We have created the foundation for a system that promises increased accuracy and effectiveness by utilizing the subtleties of feature correlation. Such novel approaches will be crucial in determining the course of artificial intelligence as its environment continues to develop.

References

[1] Dildar Masood Abdulqader, A Mohsin Abdulazeez, and Diyar Qader Zeebaree. Machine learning supervised algorithms of gene selection: A review. *Machine Learning*, 62(03):233–244, 2020.

[2] Tsehay Admassu Assegie and Pramod Sekharan Nair. Handwritten digits recognition with decision tree classification: a machine learning approach. *International journal of electrical and computer engineering (IJECE)*, 9(5):4446–4451, 2019.

[3] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

[4] Francesca De Felice, D Crocetti, M Parisi, V Maiuri, E Moscarelli, R Caiazzo, N Bulzonetti, D Musio, and V Tombolini. Decision tree algorithm in locally advanced rectal cancer: an example of over-interpretation and misuse of a machine learning approach. *Journal of cancer research and clinical oncology*, 146:761–765, 2020.

[5] Adel Sabry Eesa, Zeynep Orman, and Adnan Mohsin Abdulazeez Brifcani. A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. *Expert systems with applications*, 42(5):2670–2679, 2015.

[6] Xiyang Hu, Cynthia Rudin, and Margo Seltzer. Optimal sparse decision trees. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Rong Jin and Huan Liu. Robust feature induction for support vector machines. In *Proceedings of the twenty-first international conference on Machine learning*, page 57, 2004.

[8] Sotiris Kotsiantis. A hybrid decision tree classifier. *Journal of Intelligent & Fuzzy Systems*, 26(1):327–336, 2014.

[9] Marc'Aurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann Cun. Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, 19, 2006.

[10] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[11] Iqbal H Sarker, Alan Colman, Jun Han, Asif Irshad Khan, Yoosef B Abushark, and Khaled Salah. Behavdt: a behavioral decision tree learning to build user-centric context-aware predictive model. *Mobile Networks and Applications*, 25:1151–1161, 2020.

[12] Shiju Sathyadevan and Remya R Nair. Comparative analysis of decision tree algorithms: Id3, c4. 5 and random forest. In *Computational Intelligence in Data Mining-Volume 1: Proceedings of the International Conference on CIDM, 20-21 December 2014*, pages 549–562. Springer, 2015.

[13] Mohamed El Amine Ben Seghier, Vagelis Plevris, and German Solorzano. Random forest-based algorithms for accurate evaluation of ultimate bending capacity of steel tubes. In *Structures*, volume 44, pages 261–273. Elsevier, 2022.

[14] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang. Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, 9:515, 2018.