Novel Hierarchical Decision Tree Frameworks Introducing Tree Method Bagging-Stump Integration and Height Optimization

Małgorzata Przybyła-Kasperek[1][0000−0003−0616−9694] and Benjamin Agyare Addo[2][0000−0003−0618−5221]

University of Silesia in Katowice, Institute of Computer Science,
Będzińska 39, 41-200 Sosnowiec, Poland
malgorzata.przybyla-kasperek@us.edu.pl, benjaminagyareaddo@gmail.com

Abstract. The paper introduces a novel hierarchical decision tree framework designed to enhance classification quality in dispersed and fragmented data. By integrating two levels of modeling, the framework employs local decision trees to generate prediction vectors, which are then synthesized through a global decision tree for final classification. Five distinct approaches – Tree, Tree & height, Bagging & stump, Bagging & height, and Bagging – are proposed and evaluated. Each method varies in how local models are constructed, focusing on factors such as tree depth, bagging methods, and tree stumps. Experimental results on data sets from the UCI Machine Learning Repository demonstrate that the Bagging approach, particularly with an optimized number of bags and trees height, consistently achieves superior performance across metrics including accuracy, F-measure, and balanced accuracy. These findings highlight the framework's robustness and effectiveness in managing dispersed data, offering significant potential for applications in high-dimensional, fragmented and multi-class classification scenarios.

Keywords: Dispersed Data Classification · Bagging · Decision Tree Optimization · Local and Global Models · Ensemble Learning.

1 Introduction

Machine learning is a powerful tool that is applied in various fields to predict and classify. A challenge arises when data is sourced from several entities, clients or individuals with varying attributes. With such data, a classical machine learning model is not an optimal tool since it requires aggregating and centralizing the data, which may be expensive in terms of privacy and time sensitivity [15]. From literature, classifying dispersed data can be obtained through Ensemble learning [2, 14] and Federated learning [4, 8]. Federated learning (FL) is a collaborative machine learning approach designed to overcome the challenges of working with dispersed data while preserving privacy [3, 5]. Ensemble Learning (EL) involves building local models using separate tables proceeded by generating a final prediction through a fusion method applied to these local models [11,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_1

13]. An alternative to the approaches mentioned above is the construction of a dual-level model that aggregates prediction vectors generated by separate local models. This method inherently preserves data privacy, as only the prediction vectors are shared and consolidated, rather than the raw data itself. The structure of the data in this approach is highly flexible, allowing for a wide range of formats. Notably, a global model is not created, and the algorithm operates in a non-iterative manner. Instead of refining a global model, the focus is on developing an aggregation model that simply consolidates the predictions from the local models. Furthermore, the local models can be different in type from the aggregation model, providing additional flexibility in how the system is designed. This method is implemented in [6, 9, 10] using decision trees embedded with bagging technique and K-nearest neighbors algorithm. This paper builds on the approach introduced by [9] to explore the different methods of performing dual-level hierarchical classification using decision trees. The paper's main contributions are: a hierarchical decision tree framework for classifying dispersed, high-dimensional, multi-class data, and a comparison of five methods – Tree, Tree & Height, Bagging & Stump, Bagging & Height, and Bagging – differing in bagging use and tree structures.

Section 2 describes the dual-level hierarchy and the five methods proposed. Section 3 presents the datasets, experiments, and results discussion. Section 4 outlines the conclusions and future research plans.

2 Methods

The approach consists of two main steps: the first is training local models from local tables and the second is generating a global model. The global model makes the final decision based on the prediction vectors generated by the local models. We compare five hierarchical (two-stage) decision tree frameworks labeled: tree, tree & height, bagging & stump, bagging & height and bagging. In all these approaches, the main difference lies in building local models. The global model is a classical decision tree.

2.1 Local decision trees

In the study, we consider local models that generate prediction vectors from the abstract level and from the measurement level. 1 gives a brief description of the local model for each approach.

2.2 Global decision tree

The global model generates the final decision based on the prediction vectors generated by the local models. To do this, we use a decision tree to detect patterns in the set of prediction vectors generated by local models. A validation set is used for the examination of the prediction vectors produced by local models. A decision table (sub-table of D) – the validation set – is denoted as

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_1

Table 1. Description of local models for each approach and their output for prediction vectors

Approach	Local Model	Prediction Vector
Tree	Decision tree	Abstract level
Tree and Height	Decision tree with adjusted height	Abstract level
Bagging and Stump	Nodes generated from sampling with replacement	Measurement level
Bagging and Height	Adjusted height and sampling with replacement	Measurement level
Bagging	Decision tree and sampling with replacement	Measurement level

$D_{val} = (U_{val}, A_{val}, d)$, where $A_{val} = \bigcup_{i=1}^{n} A_i$ and $U_{val} \subseteq U$. The classification of the object $x \in U_{val}$ based on the local table D_i is done using a set of attributes A_i. Depending on the approach used to build the local model, either a single tree or each tree $Tree_i^j, j \in \{1, \ldots, k\}$ generated through the bagging method, classifies the entity x and casts a vote towards one of the decision classes. These votes are stored in a prediction vector $\mu_i(x) = [\mu_{i,1}(x), \ldots, \mu_{i,c}(x)]$, where c is the number of decision classes. We get n prediction vectors for each object x; one vector $\mu_i(x)$ for each local table i. For each object in the validation set, the prediction vectors are concatenated and stored together with the true decision class as one object in the table D_{pred}. This table serves a pivotal role in the second phase of training global model. Based on the table D_{pred} the decision tree is trained. This model learn how to classify the prediction vectors generated by the local models obtained in the previous stage. The DecisionTreeClassifier function from the scikit learn library in Python with the Gini index is used to build the global decision tree. This global tree will be used for the final classification of new objects.

2.3 Objects classification process

The training data, represented as multiple local tables with varying attributes and objects, is available initially. The process starts with training trees on these tables, either with or without bagging, and varying heights depending on the chosen method. Validation data is then used to generate prediction vectors for each local decision tree, assigning class probability/vector coefficient based on the number of trees supporting each class. These prediction vectors are subsequently combined into a unified vector, which forms the input for training a global decision tree. The global decision tree consolidates the outputs of the local decision trees to deliver the final classification, ensuring an efficient and scalable solution for high-dimensional and multi-class problems. In Figure 1, the structure of the proposed hierarchical decision tree framework is illustrated, highlighting the sequential steps involved in building the model. In Figure 2, the process for classifying new objects using the proposed hierarchical decision tree framework is depicted. The framework operates in two levels.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_1

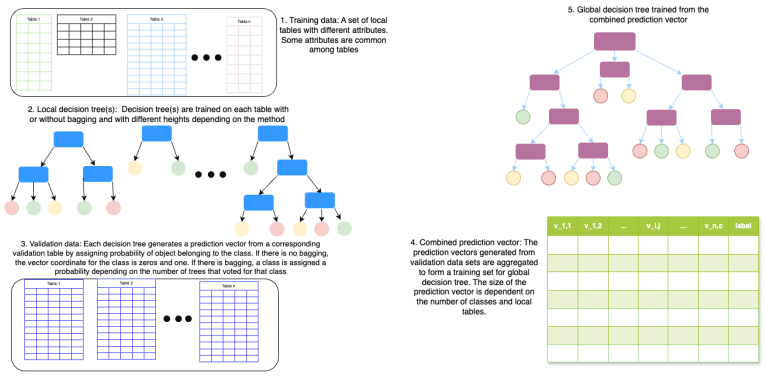

Fig. 1. Structure of the proposed hierarchical decision tree framework for dispersed data classification

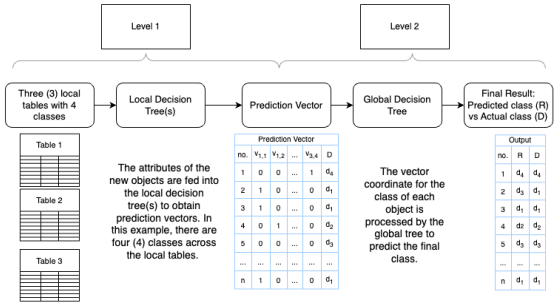

Fig. 2. Classification process for new objects using the hierarchical decision tree framework

3 Data sets and experimental results

For the experimental evaluation, three data sets from the UC Irvine Machine Learning Repository [1] were employed: the Vehicle Silhouettes data set [12], the Soybean (Large) data set [7], and the Lymphography data set [16]. Initially, these data sets were available in a non-dispersed form. To simulate dispersion, each data set was divided into five separate versions, resulting in a total of 15 dispersed data sets. The Soybean data set included predefined training and testing sets available directly from the repository, while the Vehicle Silhouettes and Lymphography data sets were split into training (70%) and testing (30%) sets through stratified random sampling. Each training data set was divided into a collection of local decision tables. This dispersion process was performed in five variations for each data set, resulting in configurations with 3, 5, 7, 9, and 11 local decision tables per training set. The characteristics of the data sets are given in Table 2.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_1

Table 2. Data set characteristics

Data set	# The training set	# The test set	# Conditional attributes	Attributes type	# Decision classes
Vehicle Silhouettes	592	254	18	Integer	4
Soybean	307	376	35	Categorical	19
Lymphography	104	44	18	Categorical	4

Table 3. Performance Metrics Across Different Datasets and Methods (Highest values in blue)

Dataset	**Metric**	**Tree**	**Tree & H**	**Bag & Stump**	**Bag & H3**	**Bagging**
Vehicle	Prec.	0.5946	0.659	0.5838	0.6878	0.6978
	Recall	0.5684	0.6552	0.572	0.7072	0.6904
	F-m.	0.5748	0.6506	0.5604	0.6912	0.6868
	bacc	0.5532	0.6314	0.559	0.6732	0.6694
	acc	0.5684	0.6552	0.572	0.7072	0.6904
Soybean	Prec.	0.69	0.69	0.287	0.761	0.7516
	Recall	0.6158	0.6158	0.3678	0.7094	0.79
	F-m.	0.5668	0.5668	0.2794	0.655	0.7562
	bacc	0.6348	0.6348	0.2976	0.6926	0.65
	acc	0.6158	0.6158	0.3678	0.7094	0.79
Lymphography	Prec.	0.451	0.6722	0.6962	0.6858	0.7404
	Recall	0.5216	0.6434	0.7288	0.6958	0.7374
	F-m.	0.4334	0.5912	0.7048	0.677	0.7274
	bacc	0.62	0.7092	0.5314	0.5478	0.6956
	acc	0.5216	0.6434	0.7288	0.6958	0.7374

Classification performance was assessed on test sets, averaging results over five runs per dataset to address bagging's non-determinism. Multiple metrics were used for a comprehensive evaluation.

The experiments were carried out according to the following scheme. For 15 dispersed data sets (Vehicle Silhouettes, Lymphography and Soyabean with version 3, 5, 7, 9, 11 local tables) and approaches: Bagging and Stump; Bagging and Height 3; Bagging the bagging method was used with a different number of bags (10, 20, 30, 40, 50, 75, 100, 150, 200, 300, 500, 1000, 1500, 2000). A wide range of bag numbers were examined due to the goal of conducting broad comparisons. For the approach Tree and Height trees of different heights were used (2, 3, 4, 5, 6, 7, 8). The test set was divided in a stratified manner into a validation (50%) and test set (50%). The validation set was used to build a global decision tree. The model's evaluation was done using a test set.

Due to space limit, we do not present results obtained for all parameters (however, they will be made available upon request sent to the authors). Table 3 shows the average of the best (in terms of classification accuracy) results obtained for different data sets. The table also shows in blue the best result (from considered approaches) for each of the data sets.

Overall, the models incorporating Bagging method proved to be the most robust and effective across all data sets, particularly when the number of bags was tuned within the range of 500 to 1500, enhancing model stability and accuracy. Simple tree-based approaches struggled to handle the complexity and diversity of the data sets, while Bagging-Stump showed moderate success but remained inconsistent. The results suggest that Bagging is particularly well-suited for dis-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_1

persed hierarchical decision frameworks, while simpler methods may require further enhancements to handle such scenarios effectively. Also Bagging with trees reduced to a height of 3 gives quite good results. However, the use of Stump in combination with Bagging does not produce better results over other approaches involving bagging.

The results of all five approaches were compared using statistical inference. Five dependent samples of 15 observations were created, representing the results for each data set and dispersion version. F-measure and balanced accuracy were selected as the metrics for comparison, as these measures account for multiple factors and are particularly suitable for unbalanced data. Since these metrics are ratio-scaled, statistical tests could be applied to assess the significance of the observed differences. To determine whether the differences in F-measure and balanced accuracy values among the approaches were statistically significant, the Friedman test was conducted.

Table 4. p-values for the post-hoc Dunn Bonferroni test for F-measure

	Tree	Tree and Height	Bagging and Stump	Bagging and Height 3	Bagging
Tree		0.83	1	0.01	0.0002
Tree and Height	0.83		1	1	0.11
Bagging and Stump	1	1		0.06	0.002
Bagging and Height 3	0.01	1	0.06		1
Bagging	0.0002	0.11	0.002	1	

Focusing first on the F-measure, the Friedman test indicated a statistically significant difference among the five approaches, $\chi^2(14, 4) = 26.8, p = 0.00002$. To identify specific differences, a post-hoc Dunn-Bonferroni test was performed, with significant results highlighted in blue in Table 4. The test revealed significant differences in the average F-measure values between the Tree approach and two other approaches: Bagging and Height 3, as well as Bagging. Additionally, a significant difference was identified between Bagging and Stump, and Bagging.

Table 5. p-values for the post-hoc Dunn Bonferroni test for balanced accuracy

	Tree	Tree and Height	Bagging and Stump	Bagging and Height 3	Bagging
Tree		1	1	0.43	0.21
Tree and Height	1		0.09	1	1
Bagging and Stump	1	0.09		0.005	0.002
Bagging and Height 3	0.43	1	0.005		1
Bagging	0.21	1	0.002	1	

Next, we analyze the differences in average balanced accuracy among the five approaches. Similar to the F-measure analysis, the Friedman test was conducted on balanced accuracy values, organized into five dependent samples of 15 observations each. The test confirmed a statistically significant difference in the averages among at least two of the approaches, $\chi^2(14, 4) = 19.6, p = 0.0006$. To

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_1

pinpoint the specific differences, a post-hoc Dunn-Bonferroni test was performed, with the significant results highlighted in blue in Table 5. The test revealed significant differences between the Bagging and Stump approach and two other approaches: Bagging and Height 3, as well as Bagging.

Based on these analyses, it can be concluded that the Tree approach and Bagging and Stump approach yielded the lowest F-measure values, while Bagging and Height 3, along with Bagging, achieved the best results. The Tree and Height approach demonstrated intermediate performance, with differences in average F-measure values that were not statistically significant compared to the other methods. These findings highlight the clear superiority of Bagging-based approaches for dispersed data in terms of F-measure performance.

The analysis of the five approaches – Tree, Tree with Height, Bagging and Stump, Bagging with Height, and Bagging – provides several key insights into their effectiveness across diverse data sets and configurations. Across all data sets, Bagging and Bagging with Height consistently delivered the best results for dispersed data. These methods demonstrated robustness in handling dispersed data sets with high dimensionality and unbalanced class distributions. The performance was particularly strong when the number of bags was appropriately tuned (500–1500 bags), showcasing their capability to generalize effectively.

4 Conclusion

This study introduces a hierarchical decision tree framework dedicated for classifying dispersed data. By combining local decision trees with a global decision tree, the framework successfully addresses challenges associated with data dispersion, high dimensionality, and multi-class classification.

The results demonstrate that Bagging-based methods, particularly with optimal bagging parameters (500–1500 bags) and reduced tree height, consistently outperformed other approaches in terms of classification accuracy, F-measure, and balanced accuracy. The approach with bagging and trees with reduced height method also exhibited robust performance, emphasizing the utility of height-constrained trees in balancing complexity and accuracy. Simpler methods, such as a single tree as a local model and bagging with stumps, showed worse compliance of adaptability and accuracy and highlighted the limitations of traditional decision tree structures in handling dispersed data. One key finding is the significant impact of tree height and ensemble size on model performance. Ensemble learning techniques, when carefully tuned, are well-suited for hierarchical frameworks operating on dispersed data.

The current framework requires complete attribute values availability for classified objects. Future work could explore methods for managing incomplete data when, for some local tables, there are no specified values on the attributes for the classified object. Also, investigating strategies for dynamically selecting or weighting local models based on their relevance to specific data subsets could enhance performance.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_1

References

1. Asuncion, A., Newman, D. (2007). UCI Machine Learning Repository. Technical Report.
2. Ksieniewicz, P., Zyblewski, P., Burduk, R. (2021). Fusion of linear base classifiers in geometric space. Knowledge-Based Systems, 227, 107231.
3. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... Zhao, S. (2021). Advances and open problems in federated learning. Foundations and trends in machine learning, 14(1–2), 1-210.
4. Lewy, D., Mańdziuk, J., Ganzha, M., Paprzycki, M. (2022, November). StatMix: Data augmentation method that relies on image statistics in federated learning. In International Conference on Neural Information Processing (pp. 574-585). Singapore: Springer Nature Singapore.
5. Li, L., Fan, Y., Tse, M., Lin, K. Y. (2020). A review of applications in federated learning. Computers & Industrial Engineering, 149, 106854.
6. Marfo, K. F., Przybyła-Kasperek, M. (2022). Radial basis function network for aggregating predictions of k-nearest neighbors local models generated based on independent data sets. Procedia Computer Science, 207, 3234-3243.
7. Michalski, R. S., Chilausky, R. L. (1999). Knowledge acquisition by encoding expert rules versus computer induction from examples: a case study involving soybean pathology. International Journal of Human-Computer Studies, 51(2), 239-263.
8. Pedrycz, W. (2023). Advancing federated learning with granular computing. Fuzzy Information and Engineering, 15(1), 1-13.
9. Przybyła-Kasperek, M., Addo, B.A., Kusztal, K. (2024) Dual-Level Decision Tree-Based Model for Dispersed Data Classification. In Marcinkowski, B., Przybylek, A., Jarzębowicz, A., Iivari, N., Insfran, E., Lang, M., Linger, H., Schneider, C. (Eds.), Harnessing Opportunities: Reshaping ISD in the post-COVID-19 and Generative AI Era (ISD2024 Proceedings); University of Gdańsk: Gdańsk, Poland, 2024; ISBN: 978-83-972632-0-8. https://doi.org/10.62036/ISD.2024.44
10. Przybyła-Kasperek, M., Marfo, K. F. (2021). Neural network used for the fusion of predictions obtained by the K-Nearest neighbors algorithm based on independent data sources. Entropy, 23(12), 1568.
11. Rahim, N., El-Sappagh, S., Rizk, H., El-serafy, O. A., Abuhmed, T. (2024). Information fusion-based Bayesian optimized heterogeneous deep ensemble model based on longitudinal neuroimaging data. Applied Soft Computing, 162, 111749.
12. Siebert, J. P. (1987). Vehicle recognition using rule based methods. Turing Institute Research Memorandum TIRM-87-018, London, UK.
13. Seydi, S. T., Saeidi, V., Kalantar, B., Ueda, N., van Genderen, J. L., Maskouni, F. H., Aria, F. A. (2022). Fusion of the multisource datasets for flood extent mapping based on ensemble convolutional neural network (CNN) model. Journal of Sensors, 2022(1), 2887502.
14. Trajdos, P., Burduk, R. (2024). Ensemble of classifiers based on score function defined by clusters and decision boundary of linear base learners. Knowledge-Based Systems, 303, 112411.
15. Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., Khazaeni, Y. (2019, May). Bayesian nonparametric federated learning of neural networks. In International conference on machine learning (pp. 7252-7261). PMLR.
16. Zwitter, M., Soklic, M. (1988). Lymphography domain. University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_1