Dual-Level Decision Tree-Based Model for Dispersed Data Classification

Małgorzata Przybyła-Kasperek
University of Silesia in Katowice
Katowice, Poland *malgorzata.przybyla-kasperek@us.edu.pl*

Benjamin Agyare Addo
University of Silesia in Katowice
Katowice, Poland *benjaminagyareaddo@gmail.com*

Katarzyna Kusztal
University of Silesia in Katowice
Katowice, Poland *katarzyna.kusztal@us.edu.pl*

Abstract

The paper proposes a decision tree-based model for dispersed data classification. The dispersed data are stored in tabular form and are collected independently. They may have different objects as well as attributes, but some of them may be common among the tables. The proposed model has a two-level hierarchical architecture that uses decision trees at each level. At the lower level, bagging is used with decision trees for each table. For a classified object, prediction vectors are generated for each table, showing the probabilities that the object belongs to various decision classes. A global tree is trained based on vectors generated for validation set and it makes the final classification for a test object. This paper outlines experimental findings for our proposed approach and contrasts them with established methodologies from the literature. Statistical analysis, based on 16 dispersed data sets, confirms that our model improves classification quality for dispersed data.

Keywords: Dispersed data, Bagging, Hierarchical model, Decision tree.

1. Introduction

The concept of classification based on dispersed data assumes growing significance. As software development teams struggle with increasingly complex data sets distributed across disparate sources, the need for effective classification methodologies becomes ever more apparent. Indeed, the ability to efficiently categorize and analyze dispersed data is central to unlocking valuable insights and driving informed decision-making in today's dynamic and interconnected digital landscape. In regulated industries such as healthcare and finance, dispersed sets may be required to comply with data protection regulations like GDPR or HIPAA. Using dispersed sets allows organizations to adhere to regulatory requirements while still leveraging valuable data for analysis and decision-making.

Classifying dispersed data stored in tabular form, where objects and attributes vary across tables, presents unique challenges in machine learning. Dispersed data finds application across diverse domains. In healthcare, patient records are often dispersed across various databases or systems. Integrating data from electronic health records, medical imaging systems, laboratory results, and genetic data allows for comprehensive patient profiling and personalized treatment strategies [6, 21]. Financial institutions deal with dispersed data sources such as transaction records, customer profiles, market data, and risk metrics. By employing dispersed financial data, institutions can enhance fraud detection, customer segmentation for targeted marketing, and risk assessment for loan approvals. Environmental monitoring involves collecting dispersed

data from various sensors, satellites, and monitoring stations to assess air quality, water quality, climate patterns, and ecological health [1]. Many different applications for dispersed data can be found, but the problem is that traditional machine learning methods cannot cope with such data.

In distributed learning, different approaches can be identified, where local models are independently constructed and then combined using fusion techniques for the final decision-making. This process can be carried out either in parallel [12] or hierarchically [4, 14]. A significant emphasis is placed on the diversity exhibited among the base classifiers [11, 13], and the efficacy of ensembles is contingent upon the specific approach employed for the generation of the final decision [7, 8]. Federated learning challenges are currently being explored in depth due to the increasing prevalence of decentralized organizational structures in various industries and widespread data collection practices across different domains. Moreover, the significance of safeguarding data privacy has become paramount. Diverse methodologies have been employed in this field, such as decision trees [9], neural networks [20], and principal component analysis [5]. Also, some contributions to classification based on distributed data have been described in papers [16, 15, 17]. However, the global tree approach has never been used before. The approach considered in this paper is different from the mentioned above. In this approach, iterative convergence of the global model is not employed, and data protection is not emphasized to the same extent. In addition, we do not have control over the dispersion of data (as in distributed learning) it is assumed that the dispersed data are collected separately.

This paper introduces a novel decision tree-based model designed specifically for dispersed data classification. The proposed model employs a two-level hierarchical architecture, leveraging decision trees at each level to effectively handle disparate data sources. At the lower level, a bagging technique is utilized with decision trees tailored for each table individually. This process yields prediction vectors for classified objects, representing probabilities of class membership across decision classes within each table. Subsequently, a global tree is constructed using these vectors from a validation set, enabling final classification for test objects. Experimental evaluations conducted on sixteen dispersed datasets demonstrate the efficacy of the proposed approach. Statistical analysis confirms its superiority over traditional methods found in the literature.

The paper is organized as follows. In Section 2, the proposed classification model using a two-level hierarchical architecture is described. Section 3 addresses the data sets that were used and presents the conducted experiments and discussion on obtained results. Section 4 is on conclusions and future research plans.

2. Methods and Concept

In this study, it is assumed that we have data in dispersed form. This means that we have access to a set of local decision tables

$$D_i = (U_i, A_i, d),\ i \in \{1, \ldots, n\}, \tag{1}$$

where U_i is a universe, a set of objects; A_i is a set of conditional attributes; d is a decision attribute. The decision table D_i is called a local table and is collected independently by one unit, which could be a hospital, bank or mobile application. Sets of objects and sets of attributes can be different between tables but some objects or attributes may be common. However, the concept that is described must be the same in all local tables, which is expressed by the occurrence of a common decision attribute d in all tables.

It is assumed that for new objects for which the classification should be done values for conditional attributes appearing in all local tables are given. Classification based on such dispersed data is not a simple task, since inconsistencies may appear in local tables, i.e. different decisions are made for the same object or a combination of conflicting decision values and conditional at-

tribute values can occur in tables. A dual-level decision trees model for such classification is proposed. This model consists of two stages. In the first stage, prediction vectors are generated using decision trees and bagging approach for local tables. For each local table, a bagging method is used to generate k training sets for the base classifiers, which are decision trees. In this way, the set of decision trees generates a prediction vector for the classified object based on one local table. The dimension of this vector is equal to the number of decision classes, and each coordinate corresponds to the probability of belonging the object to a given decision class. In the second stage, a global tree is trained based on the prediction vectors obtained from the first level. For each object, n vectors are obtained, each with a dimension equal to the number of decision classes. Based on this data, the global tree is built and makes the final classification of the object.

It can be said that the proposed method is a combination of bagging and stacking approaches from ensembles of classifiers applied to dispersed data. Figure 1 illustrates the steps involved in building the model. Local tables with different set of attributes, some condition attributes may be common, are available. In the first stage, k training sets are generated from each local table using a bagging approach. Then decision trees are generated based on these training sets. In the next step, a prediction is made for an object from validation set using the previously built decision trees. For objects from the validation set, values of attributes from all local tables are determined. The trees that are built based on one local table generate one prediction vector, where the coordinate is proportional to the number of votes cast by the trees for a given decision. The prediction vectors for one object from the validation set are then concatenated into a single sample in a new training set. This set is employed to train the global tree, which subsequently performs the final classification of the test object. The classification is based on prediction vectors generated from the trees obtained in the first stage. In the following sections, the steps of building the model will be described in more detail.

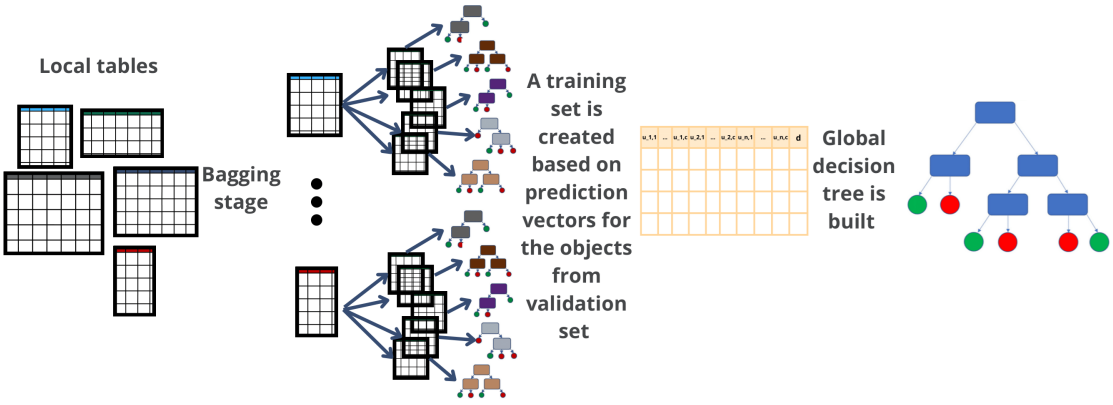

Fig. 1. Stages of model building.

2.1. Bagging method and prediction vectors

The Bagging method is employed individually for each local table D_i. A specific number of bags, let us assume k, are extracted from the decision table using the bootstrap sampling technique. This implies that the set of objects is drawn, with replacement, from the original set of objects from the local table. The size of each bag is identical to the original set. The set of conditional attributes in each bag corresponds to the original set of attributes from a given local table. A decision tree is constructed based on each bag using the CART algorithm with the Gini index [3].

It is assumed that a validation set is available to investigate the prediction vectors that are

generated by tree ensemble built based on local tables for each decision class. An object from the validation set that will be classified using dispersed data requires specified values for all attributes found in the local tables. Thus, a decision table called a validation set is given

$$D_{val} = (U_{val}, A_{val}, d), \tag{2}$$

where

$$A_{val} = \bigcup_{i=1}^{n} A_i. \tag{3}$$

The classification of the object $x \in U_{val}$ based on the single local table D_i involves using a subset of attributes A_i from that table. Each tree $Tree_i^j, j \in \{1, \ldots, k\}$, that is built using the bag, classifies the object x and contributes a vote towards one of the decision classes. These votes are then counted, with each coordinate in the vector corresponding to a decision class representing the number of votes cast by decision trees for that class. Finally, a prediction vector

$$\mu_i(x) = [\mu_{i,1}(x), \ldots, \mu_{i,c}(x)], \tag{4}$$

where c is the number of decision classes, is built based on each local table D_i.

Then, based on these prediction vectors, a training table is created. For each object in the validation set, predictions are made based on local tables. In this way, n prediction vectors are obtained

$$\mu_i(x), i \in \{1, \ldots, n\}. \tag{5}$$

Then a decision table is created

$$D_{pred} = (U_{pred}, A_{pred}, d) \tag{6}$$

in which the objects from the validation set

$$U_{pred} = U_{val} \tag{7}$$

are described by the attributes

$$A_{pred} = \{\mu_{1,1}, \ldots, \mu_{1,c}, \ldots, \mu_{n,1}, \ldots, \mu_{n,c}\} \tag{8}$$

correspond to the coordinates of the prediction vectors. For object $x \in U_{val}$, the values stored in the table are as follows

$$[\mu_{1,1}(x), \ldots, \mu_{1,c}(x), \ldots, \mu_{n,1}(x), \ldots, \mu_{n,c}(x), d(x)] \tag{9}$$

where $d(x)$ is the correct decision class for object x taken from the validation set. This table is then used in the second stage of model building to generate a global decision tree.

2.2. Global decision tree

The second stage involves training the decision tree based on the decision table D_{pred}. This model will learn how to classify the prediction vectors generated by the trees obtained in the previous stage. The CART algorithm with the Gini index is used to build the global decision tree. This tree will be used for the final classification of new objects.

When classifying a new object, it is processed at two levels. First, prediction vectors are generated using the decision trees of the first stage obtained using the bagging approach. Then, using these prediction vectors and the global decision tree trained in the second stage, the final decision is made.

The pseudo-code of algorithm generating model is given in Algorithm 1. In the first step, a bagging approach is used to generate k decision trees based on each local table. These trees

are then used to obtain prediction vectors for objects from the validation set. Based on the created prediction vectors, which form a single tuple, a global tree is generated to make the final decision. Thus, the constructed model consists of two levels. The lower level is a set of trees for each local table. Based on this level n prediction vectors are built for the test object. These vectors then form the input for classification using the decision tree from the second level.

Algorithm 1 Pseudo-code of algorithm generating model

Input: A set of local decision tables $D_i = (U_i, A_i, d), i \in \{1, \ldots, n\}$; validation set – a decision table $D_{val} = (U_{val}, A_{val}, d)$, where $A_{val} = \bigcup_{i=1}^{n} A_i$; k – the number of bags
Output: Hierarchical two-stage model. A set of ensemble decision trees and a global tree.
foreach $i \in \{1, \ldots, n\}$
    foreach $j \in \{1, \ldots, k\}$
        Create the j-th bag by randomizing with returning objects from the set U_i, define a decision table $D_i^j = (U_i^j, A_i, d)$, where $|U_i^j| = |U_i|$.
        Build a decision tree $Tree_i^j$ based on D_i^j.
    end foreach
end foreach
foreach $x \in U_{val}$
    foreach $i \in \{1, \ldots, n\}$
        foreach $j \in \{1, \ldots, k\}$
            Classify the object x based on the tree $Tree_i^j$.
        end foreach
        $\mu_{i,l}(x), l \in \{1, \ldots, c\}$ is equal to the number of votes cast by decision trees $Tree_i^j$, $j \in \{1, \ldots, k\}$, for the l-th decision class.
    end foreach
    Write a new tuple $[\mu_{1,1}(x), \ldots, \mu_{1,c}(x), \ldots, \mu_{n,1}(x), \ldots, \mu_{n,c}(x), d(x)]$ in a decision table $D_{pred} = (U_{pred}, A_{pred}, d)$
end foreach
Build a decision tree $Tree$ based on D_{pred}.
Return $Tree_i^j$, for $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, k\}$ and $Tree$

3. Data sets and experimental methodology

In the experimental part, three data sets from the UC Irvine Machine Learning Repository [2] were used: the Vehicle Silhouettes data set [19], the Soybean (Large) data set [10], and the Lymphography data set [22]. These data sets are originally available in a non-dispersed form at the UC Irvine Machine Learning Repository. However, each data set was dispersed into five versions, resulting in a total of fifteen dispersed data sets. The proposed approach was assessed using the train and test method. The Soybean set contains both a training set and a test set available in the repository. The Vehicle Silhouettes and Lymphography data sets were randomly but in the stratified mode divided into a training set (70% of objects) and a test set (30% of objects). Also, one real dispersed data set was used in the experiments. The Extrapulmonary Tuberculosis data set contains 3342 extra-pulmonary TB patients diagnosed in Ghana. The study was conducted to understand the predictors of extrapulmonary TB compared to pulmonary TB such as HIV status and gender and others: age, type of healthcare facility, health outcomes. Data was collected from four different hospitals which was also used as a natural dispersion of data: General Hospital (1433 objects), Polyclinic (775 objects), Regional Hospital (359 objects) and Teaching Hospital (775 objects). The conditional attributes are: age, sex, whether the patient is HIV-positive, site affected, has an x-ray taken, year of diagnosis. The decision attribute – TB

diagnosis – contains three decision classes: SNTB, SPTB, EPTB. This data set was also used in a similar dispersed way in the paper [18]. The characteristics of the data sets are given in Table 1.

Table 1. 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
TB	2338	1004	6	Categorical and Numerical	3

As can be noted, the analyzed data sets are multidimensional and have multiple decision classes. The training set was dispersed into a set of local decision tables. The dispersion resulted in the creation of five different version of dispersion for each considered data set, i.e. with 3, 5, 7, 9, and 11 local tables for each training set. The decision tables, derived from a single training set, exhibit diverse sets of attributes, with some attributes shared among them. The number of conditional attributes varied across individual local tables. In situations where the dispersion consisted of a smaller number of local tables, the tables contained more attributes, ranging from 6 for the Vehicle Silhouettes and Lymphography data sets to several or dozens for the Soybean data set. Conversely, when the dispersion involved a larger number of local tables, the tables contained fewer attributes, ranging from 3 to 6. All local tables store the complete set of objects, but no objects' identifiers were included, so the objects' identification is not possible across the local tables. The decision attribute was copied from the training set to all local tables.

As was mentioned, the evaluation of classification quality was conducted based on the test set, however, five repetitions of the experiments for each data set were performed. Due to non-determinism in the bagging approach, the results below report the average value from these five runs. Different measures have been employed for this purpose. The classification accuracy measure (acc), which denotes the proportion of correctly classified objects in the test set, is utilized. Recall signifies the ability of the classifier to accurately identify the given class. Precision (Prec.) indicates the frequency at which the classifier avoids misclassifying an object to a given class. The F-measure (F-m.) is a comprehensive measure that evaluates the classifier's capability to maintain balanced accuracies

$$\text{F-measure } = 2 \cdot \frac{\text{Precision} \cdot \text{ Recall}}{\text{Precision} + \text{ Recall}}. \tag{10}$$

Balanced Accuracy, on the other hand, is an average value of Recall for all decision classes. Balanced accuracy ($bacc$) ensures that the performance assessment gives equal consideration to the classification accuracy of all classes.

The experiments were carried out according to the following scheme:

- For 16 dispersed data sets (Vehicle Silhouettes, Lymphography and Soyabean with version 3, 5, 7, 9, 11 local tables and TB data set) the bagging method was used with a different number of bags (10, 20, 30, 40, 50, 75, 100, 150, 200, 300, 500). A wide range of bag numbers were examined due to the goal of conducting broad comparisons. As previously mentioned, the number of local tables was determined by the limited number of attributes within each table.

- 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.

Comparison of experimental results was made in terms of: the quality of classification for different number of bags; the quality of classification of the proposed model and other methods known from the literature.

4. Results and comparisons

This section presents experimental results and comparisons. Tables 2, 3, 4, 5 show the experimental results for the Vehicle Silhouettes, the Soybean (Large), the Lymphography data set and the TB data set respectively. In the tables, the best results are highlighted in blue. As can be seen from the tables, depending on the data set and the degree of dispersion (number of local tables) a different number of bags is optimal. First, the quality of classification for different number of bags will be compared. We check whether there is a parameter value (number of bags) that generates a classification of better quality compared to other values. Since usually for a given data set and degree of dispersion all measures reach optimal values for the same parameter value, the comparison will be made using the balanced accuracy measure.

The received balanced accuracy were divided into eleven dependent data samples, results from Tables 2 – 5 obtained for different numbers of bags. The Friedman test was used to detect differences in multiple test samples. There was not a statistically significant difference in the results obtained for the eleven different numbers of bags, $\chi^2(15, 10) = 1.4, p = 0.33$. Additionally, comparative box-whiskers charts for the results with eleven different numbers of bags were created (Fig. 2). As can be observed, the values of the balanced accuracy for 100 and 200 bags stand out compared to other results. In the next step, the Wilcoxon each-pair test was used. This test confirmed that the differences in the balanced accuracy were significant between pairs: 30 and 150 bags with $p = 0.035$; 100 and 300 bags with $p = 0.035$; 150 and 200 bags with $p = 0.007$; 200 and 300 bags with $p = 0.04$. Thus, it can be concluded that using about 100 or 200 bags in the proposed approach yields good results. Further increasing the number of bags no longer brings significant improvement. Of course, the optimal value of the number of bags always depends on the data set and cannot be predetermined in advance.

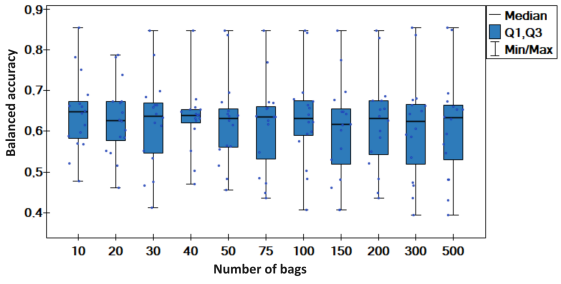

Fig. 2. Comparison of the results obtained for the eleven different numbers of bags.

Now, we will compare the proposed approach with other methods. Since the data is available in a dispersed version, it was not possible to apply the classical machine learning model directly. The baseline approach relied on applying the selected models to each local table separately. In this way, a set of classification models was built. The final classification for objects from the test set was done by simple voting. AdaBoost, Decision Tree and Naive Bayes algorithms were used to build the model based on the local tables. The selection aimed to contrast the proposed approach with tree-based methods and other models while also comparing its performance against simpler, interpretable models and more complex classifier ensembles. In each case, the models were homogeneous – the same type of models were built based on all local tables. The implementation of the models available in the ScikitLearn library with default parameters values were used for this purpose. A comparison of the results obtained for the baseline approach is given

Table 2. Results of precision (Prec.), recall, F-measure (F-m.), balanced accuracy ($bacc$) and classification accuracy (acc) for the proposed model **the Vehicle Silhouettes data set**.

No. of tables		No. of baggs										
		10	20	30	40	50	75	100	150	200	300	500
3	Prec.	0.691	0.695	0.709	0.676	0.727	0.708	0.706	0.693	0.708	0.724	0.704
	Recall	0.669	0.672	0.683	0.665	0.717	0.691	0.671	0.665	0.698	0.691	0.674
	F-m.	0.675	0.678	0.691	0.668	0.720	0.697	0.680	0.673	0.700	0.701	0.685
	$bacc$	0.650	0.647	0.665	0.639	0.696	0.672	0.657	0.648	0.675	0.677	0.656
	acc	0.669	0.672	0.683	0.665	0.717	0.691	0.671	0.665	0.698	0.691	0.674
5	Prec.	0.679	0.701	0.708	0.686	0.696	0.668	0.723	0.692	0.692	0.683	0.698
	Recall	0.687	0.691	0.702	0.674	0.698	0.658	0.721	0.701	0.704	0.683	0.687
	F-m.	0.672	0.683	0.702	0.675	0.693	0.653	0.709	0.692	0.694	0.677	0.680
	$bacc$	0.662	0.674	0.684	0.653	0.672	0.637	0.695	0.677	0.675	0.660	0.670
	acc	0.687	0.691	0.702	0.674	0.698	0.658	0.721	0.701	0.704	0.683	0.687
7	Prec.	0.656	0.632	0.666	0.681	0.653	0.670	0.664	0.672	0.679	0.653	0.653
	Recall	0.633	0.630	0.654	0.660	0.643	0.660	0.646	0.669	0.683	0.671	0.658
	F-m.	0.637	0.625	0.655	0.661	0.643	0.662	0.648	0.667	0.679	0.658	0.650
	$bacc$	0.616	0.603	0.633	0.640	0.615	0.636	0.622	0.642	0.655	0.639	0.632
	acc	0.633	0.630	0.654	0.660	0.643	0.660	0.646	0.669	0.683	0.671	0.658
9	Prec.	0.685	0.666	0.661	0.686	0.690	0.682	0.694	0.678	0.680	0.686	0.679
	Recall	0.661	0.644	0.638	0.679	0.669	0.680	0.682	0.677	0.671	0.680	0.669
	F-m.	0.667	0.646	0.637	0.675	0.674	0.672	0.684	0.674	0.663	0.677	0.666
	$bacc$	0.645	0.624	0.621	0.659	0.650	0.658	0.665	0.649	0.653	0.658	0.649
	acc	0.661	0.644	0.638	0.679	0.669	0.680	0.682	0.677	0.671	0.680	0.669
11	Prec.	0.596	0.618	0.641	0.664	0.584	0.666	0.620	0.643	0.669	0.625	0.664
	Recall	0.586	0.597	0.624	0.644	0.584	0.649	0.614	0.639	0.641	0.620	0.644
	F-m.	0.581	0.599	0.620	0.642	0.573	0.642	0.610	0.636	0.639	0.614	0.643
	$bacc$	0.568	0.585	0.613	0.632	0.563	0.636	0.599	0.618	0.626	0.604	0.627
	acc	0.586	0.597	0.624	0.644	0.584	0.649	0.614	0.639	0.641	0.620	0.644

Table 3. Results of precision (Prec.), recall, F-measure (F-m.), balanced accuracy ($bacc$) and classification accuracy (acc) for the proposed model **the Soybean data set**.

No. of tables		No. of baggs										
		10	20	30	40	50	75	100	150	200	300	500
3	Prec.	0.755	0.733	0.758	0.740	0.742	0.750	0.724	0.724	0.736	0.708	0.694
	Recall	0.777	0.776	0.779	0.772	0.786	0.793	0.758	0.770	0.777	0.761	0.761
	F-m.	0.749	0.737	0.756	0.739	0.751	0.759	0.726	0.735	0.741	0.724	0.717
	$bacc$	0.668	0.628	0.659	0.632	0.643	0.655	0.594	0.602	0.637	0.584	0.590
	acc	0.777	0.776	0.779	0.772	0.786	0.793	0.758	0.770	0.777	0.761	0.761
5	Prec.	0.756	0.787	0.776	0.741	0.735	0.760	0.761	0.774	0.784	0.769	0.796
	Recall	0.759	0.790	0.786	0.771	0.748	0.790	0.778	0.806	0.814	0.807	0.830
	F-m.	0.736	0.769	0.759	0.740	0.720	0.760	0.750	0.776	0.785	0.773	0.801
	$bacc$	0.689	0.739	0.699	0.654	0.639	0.670	0.673	0.697	0.686	0.672	0.689
	acc	0.759	0.790	0.786	0.771	0.748	0.790	0.778	0.806	0.814	0.807	0.830
7	Prec.	0.717	0.722	0.772	0.744	0.775	0.758	0.760	0.747	0.742	0.754	0.755
	Recall	0.716	0.720	0.780	0.782	0.772	0.790	0.787	0.779	0.767	0.792	0.803
	F-m.	0.694	0.695	0.758	0.747	0.753	0.758	0.756	0.749	0.739	0.760	0.767
	$bacc$	0.597	0.587	0.664	0.627	0.637	0.626	0.641	0.616	0.601	0.646	0.649
	acc	0.716	0.720	0.780	0.782	0.772	0.790	0.787	0.779	0.767	0.792	0.803
9	Prec.	0.714	0.681	0.670	0.687	0.694	0.677	0.706	0.686	0.697	0.689	0.701
	Recall	0.745	0.702	0.729	0.724	0.735	0.714	0.735	0.723	0.724	0.733	0.738
	F-m.	0.709	0.673	0.684	0.686	0.697	0.674	0.699	0.685	0.691	0.694	0.702
	$bacc$	0.588	0.547	0.551	0.551	0.556	0.548	0.576	0.531	0.550	0.533	0.545
	acc	0.745	0.702	0.729	0.724	0.735	0.714	0.735	0.723	0.724	0.733	0.738
11	Prec.	0.734	0.743	0.739	0.720	0.708	0.729	0.738	0.700	0.699	0.726	0.704
	Recall	0.778	0.779	0.766	0.747	0.735	0.746	0.755	0.725	0.724	0.752	0.730
	F-m.	0.741	0.745	0.734	0.716	0.705	0.715	0.727	0.699	0.692	0.721	0.701
	$bacc$	0.661	0.669	0.642	0.642	0.626	0.618	0.622	0.558	0.584	0.588	0.565
	acc	0.778	0.779	0.766	0.747	0.735	0.746	0.755	0.725	0.724	0.752	0.730

in Table 6. In the table, the best results are highlighted in blue. As can be seen, in the vast majority of cases, it is the proposed model that provides significantly better results (in terms of each measure). In the cases of some data sets and degrees of dispersion, the baseline approach using the decision tree also generates very good results (Vehicle Silhouettes with 7 and 11 local

Table 4. Results of precision (Prec.), recall, F-measure (F-m.), balanced accuracy (*bacc*) and classification accuracy (*acc*) for the proposed model **the Lymphography data set**.

No. of tables		No. of baggs										
		10	20	30	40	50	75	100	150	200	300	500
3	Prec.	0.798	0.786	0.668	0.685	0.705	0.611	0.618	0.564	0.614	0.614	0.603
	Recall	0.791	0.783	0.678	0.696	0.722	0.626	0.635	0.583	0.626	0.626	0.617
	F-m.	0.790	0.781	0.664	0.685	0.708	0.610	0.621	0.569	0.609	0.609	0.601
	bacc	0.855	0.788	0.533	0.606	0.564	0.436	0.503	0.406	0.436	0.436	0.430
	acc	0.791	0.783	0.678	0.696	0.722	0.626	0.635	0.583	0.626	0.626	0.617
5	Prec.	0.751	0.717	0.755	0.720	0.830	0.810	0.830	0.814	0.822	0.805	0.847
	Recall	0.730	0.704	0.696	0.713	0.765	0.757	0.774	0.765	0.757	0.757	0.774
	F-m.	0.719	0.690	0.673	0.702	0.752	0.742	0.764	0.753	0.743	0.747	0.760
	bacc	0.752	0.673	0.788	0.679	0.836	0.770	0.842	0.776	0.830	0.830	0.842
	acc	0.730	0.704	0.696	0.713	0.765	0.757	0.774	0.765	0.757	0.757	0.774
7	Prec.	0.586	0.687	0.644	0.641	0.708	0.669	0.703	0.653	0.721	0.662	0.623
	Recall	0.574	0.704	0.670	0.670	0.739	0.696	0.713	0.661	0.748	0.678	0.600
	F-m.	0.553	0.691	0.651	0.641	0.722	0.677	0.707	0.633	0.729	0.654	0.586
	bacc	0.521	0.552	0.467	0.648	0.515	0.485	0.679	0.461	0.521	0.473	0.479
	acc	0.574	0.704	0.670	0.670	0.739	0.696	0.713	0.661	0.748	0.678	0.600
9	Prec.	0.466	0.524	0.568	0.436	0.497	0.488	0.559	0.559	0.488	0.539	0.539
	Recall	0.557	0.565	0.591	0.548	0.565	0.557	0.583	0.583	0.557	0.565	0.565
	F-m.	0.496	0.530	0.572	0.472	0.519	0.509	0.560	0.560	0.509	0.545	0.545
	bacc	0.570	0.515	0.412	0.503	0.455	0.448	0.406	0.406	0.448	0.394	0.394
	acc	0.557	0.565	0.591	0.548	0.565	0.557	0.583	0.583	0.557	0.565	0.565
11	Prec.	0.769	0.769	0.785	0.785	0.785	0.785	0.785	0.785	0.785	0.785	0.785
	Recall	0.774	0.774	0.783	0.783	0.783	0.783	0.783	0.783	0.783	0.783	0.783
	F-m.	0.770	0.770	0.782	0.782	0.782	0.782	0.782	0.782	0.782	0.782	0.782
	bacc	0.782	0.782	0.848	0.848	0.848	0.848	0.848	0.848	0.848	0.848	0.848
	acc	0.774	0.774	0.783	0.783	0.783	0.783	0.783	0.783	0.783	0.783	0.783

Table 5. Results of precision (Prec.), recall, F-measure (F-m.), balanced accuracy (*bacc*) and classification accuracy (*acc*) for the proposed model **the Extrapulmonary Tuberculosis data set**.

No. of tables		No. of baggs										
		10	20	30	40	50	75	100	150	200	300	500
4	Prec.	0.515	0.501	0.520	0.509	0.520	0.510	0.522	0.518	0.520	0.503	0.521
	Recall	0.507	0.488	0.504	0.503	0.509	0.500	0.509	0.510	0.517	0.500	0.512
	F-m.	0.509	0.493	0.509	0.504	0.511	0.503	0.513	0.512	0.516	0.499	0.515
	bacc	0.477	0.462	0.476	0.471	0.483	0.472	0.483	0.481	0.482	0.466	0.481
	acc	0.507	0.488	0.504	0.503	0.509	0.500	0.509	0.510	0.517	0.500	0.512

tables, Soybean with 3 local tables and Lymphography with 7 local tables).

Statistical tests were performed in order to confirm significant differences in the obtained results. The received classification quality were divided into four dependent data samples, results from Table 6 and approaches AB, DT, NB, PM. The Friedman test was used to detect differences in multiple test samples. There was a statistically significant difference in the results obtained for the three different approaches being considered, $\chi^2(79, 3) = 114.03, p = 0.000001$. Additionally, comparative box-whiskers charts for the results with three approaches were created (Fig. 3). As can be observed, the values of the classification quality for the proposed approach are the best (much better than the other approaches). In the next step, the Wilcoxon each-pair test was used. This test confirmed that the differences in the classification quality were significant between all pairs tested (with $p < 0.0005$) except for one pair – the baseline approach using AdaBoost and Naive Bayes algorithms. This analysis definitively confirms that the proposed approach works well with dispersed data.

When discussing the limitations of the proposed method, it is crucial to address the scalability of the model. In the proposed approach, the highest computational complexity is observed during the creation of multiple decision trees. For instance, employing 11 local tables and 500 bags results in the generation of 5,500 trees. However, it is noted that such a vast number of bags is often unnecessary for achieving optimal classification quality. Furthermore, this tree construc-

Table 6. Comparison of precision (Prec.), recall, F-measure (F-m.), balanced accuracy ($bacc$) and classification accuracy (acc) obtained for the baseline approaches using AdaBoost (AB), Decision Tree (DT), NaiveBayes (NB) algorithms and the proposed model (PM).

No. of tables	Measure	Vehicle Silhouettes				Soybean			
		AB	DT	NB	PM	AB	DT	NB	PM
3	Prec.	0.694	0.700	0.599	0.727	0.110	0.846	0.786	0.758
	Recall	0.646	0.677	0.520	0.717	0.202	0.823	0.699	0.793
	F-m.	0.662	0.674	0.506	0.720	0.139	0.815	0.690	0.759
	$bacc$	0.628	0.669	0.513	0.696	0.158	0.878	0.864	0.668
	acc	0.646	0.677	0.520	0.717	0.202	0.823	0.699	0.793
5	Prec.	0.664	0.698	0.599	0.723	0.112	0.185	0.189	0.796
	Recall	0.630	0.693	0.504	0.721	0.177	0.086	0.083	0.830
	F-m.	0.635	0.692	0.484	0.709	0.134	0.106	0.108	0.801
	$bacc$	0.619	0.678	0.500	0.695	0.114	0.050	0.071	0.739
	acc	0.630	0.693	0.504	0.721	0.177	0.086	0.083	0.830
7	Prec.	0.656	0.711	0.633	0.681	0.155	0.099	0.050	0.775
	Recall	0.520	0.717	0.484	0.683	0.135	0.105	0.041	0.803
	F-m.	0.556	0.710	0.445	0.679	0.140	0.101	0.044	0.767
	$bacc$	0.518	0.699	0.484	0.655	0.064	0.056	0.056	0.664
	acc	0.520	0.717	0.484	0.683	0.135	0.105	0.041	0.803
9	Prec.	0.674	0.690	0.567	0.694	0.174	0.111	0.142	0.714
	Recall	0.441	0.681	0.457	0.682	0.110	0.124	0.135	0.745
	F-m.	0.471	0.680	0.422	0.684	0.133	0.117	0.136	0.709
	$bacc$	0.441	0.665	0.459	0.665	0.082	0.046	0.099	0.588
	acc	0.441	0.681	0.457	0.682	0.110	0.124	0.135	0.745
11	Prec.	0.601	0.689	0.551	0.669	0.181	0.089	0.092	0.743
	Recall	0.547	0.673	0.441	0.649	0.133	0.133	0.086	0.779
	F-m.	0.557	0.678	0.396	0.643	0.148	0.104	0.088	0.745
	$bacc$	0.551	0.656	0.450	0.636	0.077	0.104	0.066	0.669
	acc	0.547	0.673	0.441	0.649	0.133	0.133	0.086	0.779
No. of tables	Measure	Lymphography				Extrapulmonary Tuberculosis			
		AB	DT	NB	PM	AB	DT	NB	PM
3	Prec.	0.335	0.779	0.676	0.798	0.445	0.510	0.434	0.522
	Recall	0.386	0.773	0.682	0.791	0.355	0.492	0.559	0.517
	F-m.	0.322	0.764	0.675	0.790	0.338	0.496	0.488	0.516
	$bacc$	0.596	0.532	0.467	0.855	0.374	0.458	0.461	0.483
	acc	0.386	0.773	0.682	0.791	0.355	0.492	0.559	0.517
5	Prec.	0.576	0.826	0.685	0.847				
	Recall	0.545	0.818	0.659	0.774				
	F-m.	0.543	0.809	0.646	0.764				
	$bacc$	0.682	0.563	0.457	0.842				
	acc	0.545	0.818	0.659	0.774				
7	Prec.	0.426	0.786	0.645	0.721				
	Recall	0.386	0.750	0.614	0.748				
	F-m.	0.399	0.738	0.612	0.729				
	$bacc$	0.268	0.520	0.424	0.679				
	acc	0.386	0.750	0.614	0.748				
9	Prec.	0.541	0.790	0.750	0.568				
	Recall	0.545	0.682	0.682	0.591				
	F-m.	0.540	0.654	0.694	0.572				
	$bacc$	0.374	0.478	0.472	0.570				
	acc	0.545	0.682	0.682	0.591				
11	Prec.	0.560	0.790	0.898	0.785				
	Recall	0.523	0.682	0.523	0.783				
	F-m.	0.495	0.654	0.616	0.782				
	$bacc$	0.368	0.478	0.683	0.848				
	acc	0.523	0.682	0.523	0.783				

tion process occurs only once. The classification itself is quite fast since previously built trees are utilized. Moreover, tree construction time can be significantly reduced through pruning, and the introduction of a one-level decision tree (comprising a root and leaves) is planned for future optimization. Another potential limitation of the proposed method is the requirement for a validation set with values specified for all conditional attributes from local tables. In future research, we aim to develop a method for constructing a global tree that avoids the need for such a validation set.

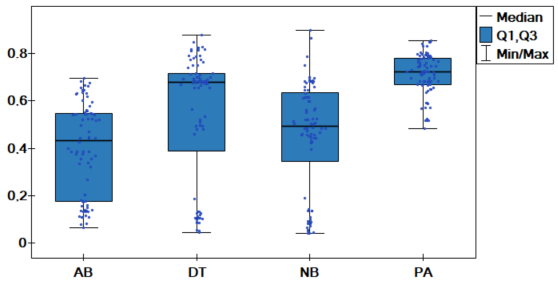

Fig. 3. Comparison of the results obtained for the four approaches: the baseline approach with algorithms AdaBoost, Decision Tree, Naive Bayes and the proposed approach.

5. Conclusions

In the paper, a hierarchical classification model with decision trees for dispersed data is proposed. The proposed model uses a bagging approach with decision trees at a lower level. Predictions obtained from first-level classifiers are used to build a global tree, which as a second-level model makes the final classification. The paper presents experimental results on sixteen dispersed data sets. It is shown that the number of bags in bagging approach about 100 or 200 is quite sufficient to achieve good results. In addition, the proposed approach was compared with the baseline approach, in which one model is generated based on each local table: AdaBoost, Decision Tree or Naive Bayes were used. It was shown that the proposed approach provides classifications with better qualities than the baseline approach.

The paper presents a model proposal and conducted experiments. Further research is needed to explore the model's performance under various conditions, such as different types of dispersed data and levels of data dispersion, to validate its effectiveness in practical applications. In the future work, it is planned to use the proposed hierarchical approach in combination with other machine learning models. Additionally, alternative hierarchical architectures of machine learning models for the two levels are planned to be explored.

References

1. Abdulla, N., Demirci, M., Ozdemir, S. (2024). Smart meter-based energy consumption forecasting for smart cities using adaptive federated learning. Sustainable Energy, Grids and Networks, 101342.
2. Asuncion, A., Newman, D. (2007). UCI Machine Learning Repository. Technical Report.
3. Breiman, L. (2017). Classification and regression trees. Routledge.
4. Czarnowski, I. (2022). Weighted Ensemble with one-class Classification and Oversampling and Instance selection (WECOI): An approach for learning from imbalanced data streams, Journal of Computational Science, 61, 101614, ISSN 1877–7503.
5. Grammenos, A., Mendoza Smith, R., Crowcroft, J., Mascolo, C. (2020). Federated principal component analysis. Advances in Neural Information Processing Systems, 33, 6453–6464.
6. Kanhegaonkar, P., Prakash, S. (2024). Federated learning in healthcare applications. In Data Fusion Techniques and Applications for Smart Healthcare (pp. 157–196). Academic Press.
7. Kashinath, S. A., Mostafa, S. A., Mustapha, A., Mahdin, H., Lim, D., Mahmoud, M. A., Mohammed, M.A., Al-Rimy, B.A.S., Fudzee M. F., Yang, T. J. (2021). Review of data fusion methods for real-time and multi-sensor traffic flow analysis. IEEE Access,

9, 51258–51276.
8. Kuncheva, L. I. (2014). Combining pattern classifiers: methods and algorithms. John Wiley & Sons.
9. Kwatra, S., Torra, V. (2021). A k-Anonymised Federated Learning Framework with Decision Trees. In Data Privacy Management, Cryptocurrencies and Blockchain Technology, 106–120, Springer, Cham.
10. 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.
11. Nam, G., Yoon, J., Lee, Y., Lee, J. (2021). Diversity matters when learning from ensembles. Advances in Neural Information Processing Systems, 34, 8367–8377.
12. Ng, W. W., Zhang, J., Lai, C. S., Pedrycz, W., Lai, L. L., Wang, X. (2018). Cost-sensitive weighting and imbalance-reversed bagging for streaming imbalanced and concept drifting in electricity pricing classification. IEEE Transactions on Industrial Informatics, 15(3), 1588–1597.
13. Ortega, L. A., Cabañas, R., Masegosa, A. (2022). Diversity and Generalization in Neural Network Ensembles. In International Conference on Artificial Intelligence and Statistics (11720–11743). PMLR.
14. Pławiak, P., Abdar, M., Pławiak, J., Makarenkov, V., Acharya, U. R. (2020). DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring. Information Sciences, 516, 401–418.
15. Przybyła-Kasperek, M., Kusztal, K. (2023). Rules' Quality Generated by the Classification Method for Independent Data Sources Using Pawlak Conflict Analysis Model. W J. Mikyška, C. de Mulatier, V. V. Krzhizhanovskaya, P. M. A. Sloot, P. Maciej, J. J. Dongarra (Red.), Computational Science - ICCS 2023 : 23rd International Conference, Prague, Czech Republic, July 3–5, 2023 : proceedings. Pt. 4 (T. 10476, s. 390–405).
16. Przybyła-Kasperek, M., Aning, S. (2022). Study on the Twoing Criterion with Prepruning and Bagging Method for Dispersed Data. W R. A. Buchmann (Red.), ISD2022 - Information Systems Development: Artificial Intelligence for Information Systems Development and Operations : proceedings (s. 1–12). Risoprint.
17. Przybyła-Kasperek, M., Marfo, K. F. (2022). Influence of Noise and Data Characteristics on Classification Quality of Dispersed Data Using Neural Networks on the Fusion of Predictions. W R. A. Buchmann (Red.), ISD2022 - Information Systems Development: Artificial Intelligence for Information Systems Development and Operations : proceedings (s. 1–12). Risoprint.
18. Sadilek, A., Liu, L., Nguyen, D., Kamruzzaman, M., Serghiou, S., Rader, B., Ingerman, A., Mellem, S., Kairouz, P., Nsoesie, E.O., MacFarlane, J., Vullikanti, A., Marathe, M., Eastham, P., Brownstein, J.S., Arcas, B.A., Howell, M.D., Hernandez, J. (2021). Privacy-first health research with federated learning. NPJ digital medicine, 4(1), 1-8, Nature Publishing Group.
19. Siebert, J. P. (1987). Vehicle recognition using rule based methods. Turing Institute Research Memorandum TIRM-87-018, London, UK.
20. Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., Khazaeni, Y. (2019). Bayesian nonparametric federated learning of neural networks. In International Conference on Machine Learning, 7252–7261, PMLR.
21. Zhang, Y., Qiu, M., Tsai, C. W., Hassan, M. M., Alamri, A. (2015). Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. IEEE Systems Journal, 11(1), 88–95.
22. Zwitter, M., Soklic, M. (1988). Lymphography domain. University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia.