Study on the Twoing Criterion with Pre-pruning and Bagging Method for Dispersed Data

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

Samuel Aning
University of Silesia in Katowice
Katowice, Poland *samuel.aning@us.edu.pl*

Abstract

In the paper, the issues related to classification based on dispersed data are considered. Dispersed data's idea is to be able to effectively make use of data collected independently from different information systems/sources/units on a single topic in the development of a classification model that could classify a new object irrespective of the information systems/source/unit the object is from. As in Federated learning approaches, also here, data is protected and not shared between the owners. Local models are built using the bagging method and decision trees with the Twoing criterion. Only prediction vectors generated based on the local models are sent to central server. Final aggregation is done using majority voting. The main purpose of the paper is to study the quality of classification obtained with the proposed approach. Another goal is to investigate the impact of the pre-pruning tree process on the quality of classification. Moreover, the comparison of results obtained for the Twoing criterion and for the Gini index during the tree construction is presented. The experiments were performed on seventeen dispersed data sets, two of which reflect the natural dispersion that occurs in reality – dispersed medical data collected by different hospitals and dispersed medical data collected in different countries. The contribution of this paper is to observe the effectiveness of using Twoing criteria as a splitting criterion together with bagging method in development of classification model for data stored in independent dispersed sources.

Keywords: Dispersed Data, Bagging, Twoing Criterion, Data Privacy

1. Introduction

Centralized data storage is now becoming a questionable way of collecting data due to the size of such data and the data privacy issue. It is natural that the data is stored in many independent information systems/silos/sets rather than in one central location. An example of application of the dispersed data in real life can be found in many areas such as economics, finance and medicine. For example, we deal with dispersed data when we want to use data sets collected by various stock exchanges or banks. Let us say we want to decide on a stock purchase. We can use the data collected by many units. We may use data available from several regions, countries or any other kind of different perspective. Since the data is collected independently, we cannot expect that the local tables structures are uniform. These tables may have different conditional attributes, but there may be some common attributes. It could be realized that, an expect analyst who has knowledge on stock exchanges could predict the optimal stock to purchase. However, when the data size is too large it becomes impossible for an expect to use all knowledge available from different stock exchanges. The objective here is to able to develop a model that could substitute the expect and perform better in such situations. A similar situation applies to the sets of object. Our goal is to use dispersed data to make a joint decision. Also a

very good example of such a situation could be different hospitals collecting data independently. Also, many sales-related websites collect data in a fragmented and dispersed manner. Such decentralized action is natural and hence becoming more and more popular mainly due to two reasons. The first is the fear of data sharing and the desire to ensure data privacy. The second reason is the independence of the institutions/units collecting the data. In the current situation, the development of information systems towards decentralized information systems is obvious.

Federated learning [9] issues deal with the use of independently collected data while protecting the data privacy. The assumption is that the data should not leave the organization/website/system through which it was collected. At the same time, the use of multiple data sources allows for the construction of much more accurate models than it would be possible if we use any of these sources independently.

Understanding the difference between Distributed Machine Learning (DML) [5] and Federated Learning (FL) [14] is crucial. The difference is fundamental and concerns the form of the data and the approach to model training. In DML, we deal with one data set, which we then divide (in some way over which we have control) and split between processors to generate local models. The predictions of local models are aggregated at the end. This is done in order to improve accuracy, for example when we have access to small data set; or in order to speed up/enable calculations when the data is too large. In such an approach, the assumption of the independent and identically distributed (IID) is usually fulfilled, i.e. that the variables in the subsets are paired independently and come from the same distribution. This is extremely important when building an aggregated model. The data privacy and data protection issues are not addressed here, as the data is originally stored in one system/data set.

The application for FL is completely different. First, data is collected independently. This causes a lot of problems. One of the most important is that the data is rather non-IID [3]. As a result, many statistical methods, the operation of which is based on this assumption, cannot be used in this situation. Another problem is the lack of control over the set of objects and set of attributes in the independently collected data sets. Although in the literature there is a division into: Horizontal Federated Learning (object ID spaces are the same and label features are different); Vertical federated learning (object ID spaces are different and label features are the same) or Federated Transfer Learning (both object ID spaces and label features are different), nevertheless, in this paper a more complex approach is considered in which we cannot assume either of these cases. We require neither the separability nor the equality of any of the attributes sets or objects sets in data. We refer to this type of data as dispersed data. Another problem in FL is that the number of objects in data sets of different participants are also very varied. It depends on the individual possibilities and cannot be controlled. In addition, the model building process in FL is completely different than in DML. The models are built locally, and only the models (not raw data) are made available to the central unit. In addition, the models are iteratively changed based on the feedback received after checking the quality of the aggregated model.

Decision tree classifier has been used frequently in classical cases of classification problems because of its intuitive approach in its design. The papers [10, 11] present the successful use of decision trees with the Gini index and the locally applied bagging method for the classification task based on dispersed data. It is a combination of DML and FL techniques in a task for data available from independent data sources – i.e. data used in FL problems. Only information about the results of the predictions made based on local models is made available to the centralized computation. In the study presented in [11] the use of prepruning, consisting of limiting tree growth by using the minimum number of objects in leaves, was very important. It turned out that up to a certain threshold limiting the size of the tree does not significantly affect the quality of classification, but it significantly reduces the complexity of the model. In [12] comparative analysis of the Twoing and entropy criterion for decision tree classification of dispersed data is presented. The main observation shown is when knowledge is highly dispersed in a lot of local

tables, using Twoing criterion in building decision tree models is better in terms of classification quality than using entropy measure (prepruning was not analyzed in [12]).

However, due to the still unexplored approach of combining DML and FL techniques for dispersed data, further research is needed. In this paper, for the first time decision trees with the Twoing criterion and bagging method was used for dispersed data. The Twoing criterion, like the Gini index, is used in CART model, but has a fundamentally different approach as it produces more balanced trees. While the Gini index only takes into account the creation of clean partitions, the Twoing criterion also takes into account the balance of the tree which is extremely important in the context of the quality of classification and the complexity of the tree itself. This influence is especially important when many different trees are built (as in the case discussed in this paper). Also, the idea in the Twoing criterion is to group all classes into two superclasses so to be considered as a two-class problem [2]. In this paper, a wide comparison of the two splitting criterion, for different versions of data dispersion, different stop criterion values, and different data has been provided. In total 2815 experiments were performed.

In addition, in this paper, for the first time, the bagging method with decision trees (the Twoing criterion and the Gini index) are investigated on two real data sets in the medical field, in which the division into dispersed data is generated by natural dispersion (due to the countries or hospitals from which the data are derived).

2. Methods and Concept

This section describes the classification model for dispersed data available in many local decision tables. The innovative approach used in this paper to dispersed data is a combination of techniques from Federated Learning and Distributed Machine Learning.

First, we assume that local data sets have been collected by independent units. Very often in such a situation the sets of attributes occurring in local data sets are limited (have very few elements). This results from the limited possibilities of individual units and we cannot control or change this state. A similar situation applies to sets of objects in local data sets. Sets of objects and attributes in local data are independent and we do not impose any conditions limiting their form. Moreover, we do not assume the existence of global object identifiers that would allow us to identify whether there are common objects between particular local sets. Such assumptions and the guarantee of privacy and data protection is a characteristic feature of the FL approach.

Second, our idea is to use bagging method with decision trees for each local set. This DML method is used to increase the quality of classification and is especially effective for data with small set of attributes or set of objects – as it can be the case with dispersed data. In this paper, the Classification and Regression Trees CART algorithm was used to build decision trees and for the first time with the Twoing criterion.

The classification process can be described in the following steps:

- For each local set, the bagging method with CART trees and the Twoing criterion is used. In this step, the different numbers of bags, namely: 10, 20, 30, 40 and 50 were considered in the bagging method. Also, various minimal numbers of objects in the leaf when the tree is pre-pruned were investigated, namely: 2, 4, 6, 8, 10 and $0.1\times$ the number of objects in the training set.

- Predictions results stored in the form of vectors determined based on local sets are aggregated with the use of majority voting.

2.1. Bagging Method with CART Trees and the Twoing Criterion

We assume that a set of decision tables is given. The tables were collected independently by separate units. Based on each table, a classifier is built locally, without sharing information with

the others participants. We assume that a set of decision tables $D_{ag} = (U_{ag}, A_{ag}, d)$, $ag \in Ag$ from one discipline is available, where U_{ag} is the universe, a set of objects; A_{ag} is a set of conditional attributes; d is a decision attribute. Ag is a set of agents – participants in classification process.

Bagging method is used separately for each local table. A given number of bags are drawn from decision table using the bootstrap sampling method. It means that the set of objects is drawn with returning from the original set of objects from local table. The cardinality of bag is equal to the original set. The set of conditional attributes in each bag is the same as the original set of attributes from a given local table. Based on each bag, a decision tree is built with the CART algorithm. In the CART algorithm the tree is constructed iteratively by searching for the optimal division in terms of the given splitting criterion. The CART algorithm creates a binary tree, which means that the division into two subsets is considered each time. There are different possible stopping criterion in the algorithm. First, the obvious, if we get a clean partition, which means that all objects in a given set belong to one decision class, then a leaf is created containing the given decision class. Further stopping criteria are related to pre-pruning and prevention of excessive tree growth. These can include: limiting the number of nodes in the tree, limiting the tree height or setting a minimum number of objects in a given node to split. The last mentioned approach is used in this paper.

The splitting criterion is very important and that affects the tree shape the most. The CART algorithm typically considers the Gini index or the Twoing criterion. The first criterion was considered in [10, 11]. Below we define both of these criteria assuming that we consider the division of set of objects X into two disjoint sets X_1, X_2 (the split is defined by the test on the conditional attribute).

The Gini index of division $X_1, X_2 \subseteq U_{ag}$, that is defined based on the attribute $a \in A_{ag}$ is calculated as follows

$$Gini_a(X|X_1, X_2) = \frac{|X_1|}{|X|} Gini(X_1) + \frac{|X_2|}{|X|} Gini(X_2),$$

where $Gini(X_j) = 1 - \sum_{i \text{ decision class}} (p_j^i)^2$, p_j^i is a fraction of objects from the $i-$th decision class in the set X_j and $|X|$ is the size of the set X. Gini splitting rule will minimize the above value.

The Twoing criterion of division $X_1, X_2 \subseteq U_{ag}$, that is defined based on the attribute $a \in A_{ag}$ is calculated as follows

$$Twoing_a(X|X_1, X_2) = \frac{|X_1| \cdot |X_2|}{4} \Big(\sum_{i \text{ decision class}} |p_1^i - p_2^i| \Big)^2.$$

Twoing splitting rule will maximize the above value.

The paper [6] showed significant differences and characteristics concerning these two splitting criteria. The main conclusions are as follows. The Twoing criterion produces much more equally balanced tress than the Gini index. In addition, Twoing criterion detects independent attributes and makes them more important when building the tree. In [7] it was found that Twoing criterion is better for multi-class dependent variables than Gini index. In a theoretical study of Breiman [2] it was proved that if the number of dependent attributes is small, then all division criteria give a similar result. However, many studies [6, 7] have shown that Twoing criterion gives a better quality of classification than Gini index. Yet, such a comparison has not been made for dispersed data.

A test object that will be classified with the use of dispersed data should have specified values on all the attributes occurring in the local tables. Classification for the test object generated based on one local table is made by using the subset of attributes from a given local table and presented as a vector over the decision classes. Each tree that was built based on the bag

classifies the test object and gives a vote for one of the decision classes. Votes are counted – the vector's coordinate corresponding to a given decision class is equal to the number of votes cast by decision trees for a given decision class. The vector is sent for central computation.

This phase algorithm is performed locally and the pseudo-code is given in Algorithm 1. In the first loop, the model is trained in each of the local device/unit. It is quite a complex process, but it is only performed once. Depending on the approach, such local models can be sent to a central server for global decision making. The second approach allows not even sending this local model, but only the prediction vectors generated for the test objects with the use of the generated local models. The second loop of the algorithm presents the process of generating local prediction vectors with the use of decision trees. This process has much lower computational complexity as it uses one path in each of the generated local trees.

Algorithm 1 Pseudo-code of algorithm generating predictions based on local tables

Input: A set of local decision tables $D_{ag} = (U_{ag}, A_{ag}, d), ag \in Ag$; test set - a decision table $D_{test} = (U_{test}, A_{test}, d_{test})$, $A_{test} = \bigcup_{ag \in Ag} A_{ag}$
Output: A set of vectors over decision classes $\mu_{ag}(x), ag \in Ag, x \in U_{test}$.

foreach $ag \in Ag$ the following calculations are made locally
    foreach i from 1 to the number of bags
        Create the i-th bag by randomizing with returning objects from the set U_{ag}, define a decision table $D^i_{ag} = (U^i_{ag}, A_{ag}, d)$, where $|U^i_{ag}| = |U_{ag}|$.
        Build a decision tree $Tree^i_{ag}$ based on D^i_{ag}.
    end foreach
end foreach
foreach $x \in U_{test}$
    foreach $ag \in Ag$ the following calculations are made locally
        foreach i from 1 to the number of bags
            Classify the test object x based on the tree $Tree^i_{ag}$.
        end foreach
        $\mu_{ag,j}(x)$ is equal to the number of votes cast by decision trees $Tree^i_{ag}$, i from 1 to the number of bags, for the j-th decision class of the decision table D_{ag}.
    end foreach
end foreach

2.2. Central Computation

The prediction vectors' calculated locally based on each table are then made available to make decision for the test object. For this purpose, majority voting is used. Thus, the sum of these vectors is determined and all decision classes that received the highest number of votes constitute the set of global decisions made. Thus, we define the global decision set for the test object x as follows $\arg\max_j$ decision classes $\{\sum_{ag \in Ag} \mu_{ag,j}(x)\}$. As can be seen, the aggregation of the prediction results is very simple and has low computational complexity. In the formula above, draws may occur. In this situation, a decision set that contains more than one decision can be generated. However, as experiments have shown, when decision trees with a bagging approach are used, such a situation is extremely rare.

3. Data

Five different data sets were used in the experimental part. Three sets were taken from the UC Irvine Machine Learning Repository [1], one set from the Robert Koch Institut [4] and one from the Harvard Dataverse Support [8]. All data are originally available in one decision table. However, they were dispersed into local tables in the data preparation process. The data from the UCI repository: Vehicle Silhouettes, Lymphography and Soyabean (Large) was artificially dispersed in five different versions (3, 5, 7, 9 and 11 local tables). These data were also used in the papers [10, 11], therefore, we will not describe the data preparation process in detail here, as it is the same as before. It is only worth to mention that conditional attributes were dispersed between local tables, but some attributes were common between randomly selected tables. All the objects in the original data set were included in each of the local tables, but the object identifiers were not stored, so re-concatenation of the local tables is not possible. The two remaining data sets: Avian influenza A (H5N1) and Extrapulmonary Tuberculosis, are from the medical domain and dispersion in these cases simulate the natural process of collecting data independently by different entities. The Avian influenza is epidemiological data set established to track avian influenza infections in humans around the world. A natural dispersion here was defined by the countries in which the data were collected. In fact, the data were collected in 12 countries, but due to the very small number of objects from some countries, it is reasonable to divide the data set into 4 local tables, namely for countries: Egypt (99 objects), Vietnam (34 objects), Indonesia (126 objects) and other countries - China, Republic of Korea, Nigeria, Laos, Cambodia, Myanmar, Pakistan, Bangladesh, Thailand (35 objects). Local tables are created in such a way that each of them contained all conditional attributes, while the subsets of the objects correspond to the specific country (or countries as is the case of the last local table). Of course, the country attribute is not stored in the local tables. The decision attribute was created based on the combination of two attributes: the confirmed status (whether the disease has been confirmed) and the outcome (outcome of disease at the time of report). The confirmed status attribute has the following values: confirmed, probable, suspected; and the outcome attribute has the values: cured, fatal, under treatment and n.d. However, not all combinations of these values are present in the data set, so 9 decision classes were obtained as a result. The dispersion of the Avian influenza data set was inspired by the paper [13]. The last step was to divide the data set into the training set (70%) and the test set (30%). The stratified method has been used here where test data was obtained by randomly selecting 30% of each of the four local tables. 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 [13]. As before, the set was divided (using stratified method) into two separate subsets: training and test in the proportion of 70%, 30% respectively. The characteristics of the data sets are given in Table 1.

The quality of classification was evaluated based on the test set. The measure that is used is the estimator of classification error e. It is a fraction of the total number of objects in the test set that were classified incorrectly. An object is considered to be correctly classified when its correct decision class belongs to the generated decision set. In the considered approach, the ties are possible, but they happen very rarely. The average number of generated decisions sets $\bar{d}$, is also considered. This measure allows an assessment of how often and how numerous are the draws generated by the dispersed classification model.

Table 1. Data set characteristics

Data set	# The training set	# The test set	# Conditional attributes	# Decision classes
Vehicle Silhouettes	592	254	18	4
Soybean	307	376	35	19
Lymphography	104	44	18	4
Avian influenza	205	89	6	9
TB patients	2338	1004	6	3

4. Results and Comparison

The experiments were carried out according to the following scheme:

- For 17 dispersed data sets (Vehicle Silhouettes, Lymphography and Soyabean with version 3, 5, 7, 9, 11 local tables; Avian influenza and Extrapulmonary Tuberculosis with 4 local tables) the bagging method was used in local destinations with a different number of bags (10, 20, 30, 40, 50 values were tested). Then, local models were build with using CART decision trees with the Twoing criterion and various minimal numbers of objects in the leaf when the tree is pre-pruned, namely: $2, 4, 6, 8, 10$ and $0.1\times$ the number of objects in the training set. Thus, for each dispersed set, local models were built using 30 different combinations of parameters. Moreover, the experiment was repeated 5 times (because the bagging method is non-deterministic) for each setting. Thus, a total of 2525 experiments were performed using the Twoing criterion.

- Prediction vectors for test objects were generated with the use of local models, which were then aggregated using majority voting.

- The above procedure was performed using the Gini index with the CART algorithm only for two dispersed data sets: Avian influenza and Extrapulmonary Tuberculosis with 4 local tables. Here, 290 experiments were performed. The results with the Gini index for the remaining dispersed data sets were taken from the paper [11].

Comparison of experimental results was made in terms of:

- The quality of classification for Twoing criterion with different numbers of objects in a leaf when the tree is pre-pruned;

- The trees complexity obtained for Twoing criterion with different numbers of objects in a leaf when the tree is pre-pruned;

- The quality of classification of the proposed model using Twoing criterion vs. Gini index.

The results obtained for the Twoing criterion and all dispersed data sets are presented in Table 2. The results obtained for Gini index and dispersed data sets Avian influenza and Extrapulmonary Tuberculosis are presented in Table 3 (for the remaining data sets, the results with Gini index will be taken from the paper [11]). The tables shows the average results obtained from five experiments executions. The value of the $\bar{d}$ measure are not included in the tables due to the limited page number, but in most cases the analyzed approach makes unambiguous decisions. The highest obtained $\bar{d}$ value was equal to 1.05, but in most cases it was equal to 1. As one of the research goals is to compare the quality of classification in relation to the minimum number of objects in the leaf, in each considered case (for each dispersed set and each number of bags in the bagging method), the best result obtained for different minimum number of objects in the leaf is marked in blue.

Table 2. Results of classification error e for respective minimum samples split when bagging method and Twoing criterion are applied. Designation $\#U$ is used for the number of objects in the training set.

No. of local tables	Minimum samples split	Data sets: Vehicle 10	Vehicle 20	Vehicle 30	Vehicle 40	Vehicle 50	Lymphography 10	Lymphography 20	Lymphography 30	Lymphography 40	Lymphography 50	Soyabean 10	Soyabean 20	Soyabean 30	Soyabean 40	Soyabean 50
3	2	0.229	0.227	0.238	0.229	0.225	0.227	0.209	0.227	0.236	0.245	0.102	0.104	0.124	0.103	0.107
	4	0.235	0.25	0.246	0.238	0.235	0.223	0.227	0.205	0.218	0.223	0.117	0.107	0.122	0.106	0.098
	6	0.247	0.227	0.237	0.236	0.219	0.205	0.205	0.205	0.223	0.223	0.112	0.099	0.104	0.101	0.094
	8	0.239	0.229	0.242	0.23	0.209	0.187	0.209	0.200	0.200	0.200	0.102	0.113	0.111	0.105	0.114
	10	0.228	0.237	0.232	0.238	0.217	0.205	0.191	0.200	0.200	0.196	0.120	0.126	0.112	0.111	0.113
	$0.1 \times \#U$	0.257	0.247	0.258	0.24	0.251						0.193	0.196	0.195	0.182	0.175
5	2	0.209	0.205	0.209	0.207	0.216	0.214	0.2	0.196	0.187	0.182	0.175	0.146	0.116	0.148	0.138
	4	0.235	0.202	0.207	0.213	0.206	0.205	0.177	0.205	0.216	0.209	0.145	0.126	0.129	0.144	0.151
	6	0.229	0.206	0.210	0.216	0.226	0.25	0.205	0.232	0.232	0.196	0.167	0.124	0.137	0.124	0.132
	8	0.213	0.222	0.220	0.218	0.217	0.209	0.196	0.214	0.196	0.218	0.127	0.144	0.123	0.119	0.133
	10	0.220	0.215	0.217	0.215	0.233	0.205	0.218	0.200	0.200	0.227	0.162	0.151	0.149	0.138	0.131
	$0.1 \times \#U$	0.250	0.244	0.236	0.238	0.236						0.234	0.255	0.23	0.23	0.22
7	2	0.262	0.261	0.268	0.267	0.262	0.309	0.259	0.277	0.236	0.264	0.174	0.156	0.165	0.164	0.173
	4	0.270	0.272	0.262	0.266	0.266	0.259	0.241	0.268	0.259	0.246	0.171	0.175	0.18	0.172	0.173
	6	0.272	0.258	0.266	0.258	0.258	0.282	0.291	0.273	0.259	0.264	0.156	0.165	0.175	0.193	0.180
	8	0.267	0.272	0.255	0.260	0.264	0.250	0.259	0.250	0.241	0.273	0.203	0.198	0.178	0.189	0.179
	10	0.262	0.259	0.257	0.264	0.267	0.282	0.264	0.250	0.291	0.291	0.193	0.186	0.206	0.188	0.187
	$0.1 \times \#U$	0.299	0.279	0.277	0.277	0.267						0.274	0.270	0.269	0.255	0.264
9	2	0.277	0.268	0.279	0.265	0.257	0.277	0.273	0.259	0.273	0.268	0.143	0.136	0.123	0.122	0.119
	4	0.276	0.272	0.277	0.264	0.258	0.282	0.300	0.286	0.277	0.273	0.161	0.175	0.177	0.163	0.166
	6	0.288	0.298	0.283	0.271	0.265	0.282	0.295	0.264	0.273	0.277	0.150	0.121	0.128	0.114	0.115
	8	0.281	0.292	0.283	0.254	0.262	0.286	0.291	0.291	0.273	0.286	0.146	0.143	0.127	0.131	0.128
	10	0.286	0.294	0.278	0.264	0.258	0.273	0.285	0.259	0.286	0.273	0.157	0.149	0.130	0.130	0.127
	$0.1 \times \#U$	0.301	0.3	0.299	0.302	0.301						0.251	0.240	0.234	0.236	0.232
11	2	0.295	0.3	0.292	0.281	0.291	0.341	0.318	0.309	0.309	0.309	0.198	0.178	0.186	0.177	0.172
	4	0.298	0.296	0.298	0.281	0.289	0.286	0.318	0.314	0.295	0.309	0.206	0.188	0.168	0.173	0.171
	6	0.298	0.31	0.29	0.263	0.284	0.327	0.304	0.304	0.327	0.318	0.190	0.179	0.166	0.172	0.169
	8	0.287	0.293	0.291	0.284	0.282	0.286	0.327	0.323	0.336	0.318	0.208	0.184	0.177	0.179	0.175
	10	0.31	0.303	0.298	0.29	0.3	0.332	0.286	0.318	0.309	0.314	0.201	0.184	0.184	0.181	0.172
	$0.1 \times \#U$	0.328	0.321	0.315	0.312	0.312						0.296	0.289	0.283	0.298	0.285

		Avian influenza 10	Avian influenza 20	Avian influenza 30	Avian influenza 40	Avian influenza 50	Extrapulmonary Tuberculosis 10	Extrapulmonary Tuberculosis 20	Extrapulmonary Tuberculosis 30	Extrapulmonary Tuberculosis 40	Extrapulmonary Tuberculosis 50
4	2	0.326	0.283	0.310	0.301	0.287	0.424	0.417	0.425	0.417	0.417
	4	0.303	0.283	0.312	0.301	0.312	0.421	0.417	0.422	0.424	0.420
	6	0.317	0.306	0.312	0.308	0.306	0.432	0.423	0.423	0.414	0.420
	8	0.315	0.310	0.306	0.321	0.301	0.426	0.416	0.420	0.418	0.414
	10	0.308	0.303	0.308	0.299	0.303	0.438	0.423	0.421	0.416	0.419
	$0.1 \times \#U$						0.409	0.414	0.418	0.416	0.414

Table 3. Results of classification error e for respective minimum samples split when bagging method and Gini index are applied. Designation $\#U$ is used for the number of objects in the training set.

No. of local tables	Minimum samples split	Data sets: Avian influenza 10	Avian influenza 20	Avian influenza 30	Avian influenza 40	Avian influenza 50	Extrapulmonary Tuberculosis 10	Extrapulmonary Tuberculosis 20	Extrapulmonary Tuberculosis 30	Extrapulmonary Tuberculosis 40	Extrapulmonary Tuberculosis 50
4	2	0.254	0.256	0.254	0.243	0.263	0.411	0.410	0.418	0.416	0.414
	4	0.281	0.211	0.254	0.267	0.245	0.416	0.420	0.416	0.417	0.421
	6	0.288	0.270	0.268	0.272	0.251	0.411	0.412	0.414	0.421	0.419
	8	0.288	0.312	0.297	0.315	0.283	0.416	0.416	0.421	0.418	0.414
	10	0.312	0.281	0.308	0.290	0.297	0.420	0.420	0.422	0.420	0.426
	$0.1 \times \#U$						0.425	0.425	0.426	0.428	0.424

4.1. Quality of Classification for Twoing Criterion with Different Numbers of Objects in a Leaf when the Tree Is Pre-pruned

Different minimum number of objects in a leaf were analyzed: $2, 4, 6, 8, 10, 0.1 \times \#U$, where $\#U$ is the number of objects in the training set. For the Lymphography data set the value $0.1\#U$ was not tested as it is equal to 10. For the Avian influenza data set the number of objects in local tables are varied but is relatively small (69, 24, 88, 24), hence the value $0.1\#U$ was also not tested because it is equal to the other analyzed stop criterion. Based on the results presented

in Table 2 and the optimal values marked in blue, it can be observed that generally, optimal results are obtained when the minimum number of objects in leaf used are 2 and 4. This implies that for dispersed data, the issue of overfitting is not significant as test data performs better on the model when trees are built extensively. This is due to the fact that each of the local tables has a limited set of conditional attributes, as here local information systems only have local knowledge. Also, when we apply the Twoing criterion, optimal results are obtained for a different minimum number of objects in a leaf. Thus, a hypothesis arises that with different values of the stop criterion we obtain comparable values of the classification error. To verify the hypothesis, statistical tests were performed. The results were divided into 6 groups, each for a different minimum number of objects in the leaf: Group 1 – stop criterion equal to 2; Group 2 – stop criterion equal to 4; Group 3 – stop criterion equal to 6; Group 4 – stop criterion equal to 8; Group 5 – stop criterion equal to 10 and Group 6 – stop criterion equal to $0.1\#U$. A set of six dependent variables (one for each group) with 85 observations in each was obtained (results from Table 2). The Friedman's test was performed at first. The test confirmed that differences among the classification error in these six groups are significant, with a level of $p = 0.000001$. But if the last group was omitted (results for stop criterion $0.1\#U$) then the $p-$value was equal to 0.1. Thus, only a drastic increase in the minimum number of objects in a leaf to $0.1 \times \#U$ generates a statistically significant difference in the quality of classification. The values of the minimum number of objects in a leaf in the range from 2 to 10 do not significantly change the quality of classification. In order to determine the pairs of groups between which statistically significant differences occur, the Wilcoxon each pair test for dependent groups were performed. The test showed that there is significant difference between Group 6 (stop criterion $0.1\#U$) and all other groups; also between Group 1 and Group 5 (stop criterion 2 and 10). Additionally, comparative box-plot chart for the values of classification error was created (Figure 1). As can be observed, distributions of the classification error values in groups are not different except for the last group. Thus, the stop criterion for pre-pruning up to 10 does not affect the quality of classification.

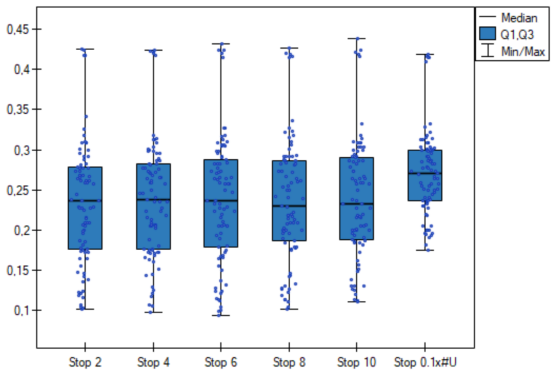

Fig. 1. Box-plot chart with (Median, the first quartile - Q1, the third quartile - Q3) the value of classification error e for different minimum number of objects in the leaf (2, 4, 6, 8, 10, $0.1\#U$).

4.2. Trees Complexity Obtained for Twoing Criterion with Different Numbers of Objects in a Leaf when the Tree is Pre-pruned

In Table 4, results of average number of nodes (denoted as **n**) and height (denoted as **h**) of decision trees from the experiments with 50 bags in the bagging method and the Twoing criterion are presented. For other versions of the bagging method, the results are similar, we do not include them due to the space limitation. There is a clear relation between the minimum number of objects in a leaf and the tree's complexity (expressed by the number of nodes and tree's

height). There is a negative correlation here, the complexity of trees decreases significantly with the increase of the stop criterion. In conjunction with the conclusion from the previous section, it follows that it is most optimal to use a stop criterion of around 8 or 10. This does not significantly affect the quality of classification, but considerably simplifies the model. That is, there is high complexity cost for when we use stop criterion for 2, 4 and 6.

Table 4. Average number of nodes (**n**) and average decision tree height (**h**) for respective minimum samples split when bagging method with 50 bags and the Twoing criterion are applied. Designation $\#U$ is used for the number of objects in the training set.

Minimum samples split	No. of local tables 3		5		7		9		11	
	n	**h**	**n**	**h**	**n**	**h**	**n**	**h**	**n**	**h**
	Vehicle									
2	200.176	14.32	249.862	15.552	288.621	16.115	324.976	16.275	333.616	16.166
4	176.792	13.998	216.334	14.909	243.706	15.491	279.203	15.703	285.556	15.784
6	158.833	13.329	188.342	14.537	207.357	14.942	233.991	15.226	239.741	15.239
8	143.648	13.329	167.028	14.012	179.858	14.232	198.350	14.593	203.062	14.662
10	130.965	12.841	149.31	13.534	158.415	13.979	171.247	14.139	174.581	14.172
$0.1\#U$	37.512	8.043	38.344	8.221	38.877	8.372	38.923	8.590	38.391	8.576
	Lymphography									
2	34.531	7.252	30.664	5.971	23.576	5.313	16.590	4.368	12.078	3.482
4	30.502	6.784	28.580	5.893	22.525	5.237	16.243	4.187	11.847	3.324
6	26.905	6.599	26.315	6.209	21.343	5.234	15.797	4.231	11.311	3.295
8	23.672	5.733	23.955	5.682	20.505	5.117	15.298	4.251	10.907	3.350
10	21.440	5.733	21.650	5.553	18.737	4.973	14.688	4.108	10.527	3.302
	Soyabean									
2	73.916	9.001	49.931	7.187	57.199	6.211	53.756	6.171	44.597	5.615
4	66.861	8.689	45.410	6.999	66.786	6.185	51.479	6.134	42.678	5.583
6	60.779	8.483	41.465	6.716	45.450	6.039	45.491	6.183	40.026	5.589
8	55.477	8.158	38.282	6.604	41.522	5.882	41.418	6.065	37.081	5.497
10	51.308	7.931	35.091	6.448	37.638	5.789	37.467	5.932	34.483	5.453
$0.1\#U$	27.374	6.836	19.867	5.120	18.990	4.834	19.580	4.971	18.860	4.757
	Avian influenza - 4 local tables					Extrapulmonary - 4 local tables				
	n	**h**				**n**	**h**			
2	25.787	5.894				209.112	10.183			
4	22.221	5.447				201.031	10.238			
6	18.605	5.208				190.100	10.168			
8	15.210	4.743				177.583	10.109			
10	15.084	4.614				164.580	10.013			
$0.1\#U$						73.412	8.659			

4.3. Quality of Classification of the Proposed Model Using Twoing Criterion vs. Gini Index

For all analyzed dispersed data sets and all considered stop criterion values, the the best obtained classification error values are marked on Figure 2. As can be clearly seen, the Twoing criterion provides better results than the Gini index. For all analyzed values of stop criterion, a lower classification error was observed using the Twoing criterion compared to the Gini index. In order to confirm the hypothesis that there is a significant difference in the average classification errors generated with the proposed approach using the Twoing criterion against Gini index, statistical tests were performed. The results were divided into 2 groups: Group 1 – results with Twoing criterion and Group 2 – results with Gini index. In each group 96 observations was obtained. The Wilcoxon test for dependent groups confirmed that differences among the classification error in these two groups are significant, with a level of $p = 0.0001$. The t-Student test also confirmed this hypothesis, with a level of $p = 0.000001$. Additionally, comparative box-plot chart for the values of classification error was created (Figure 3). As can be observed, obtained values of the classification error with using the Twoing criterion are much lower than values obtained with using the Gini index.

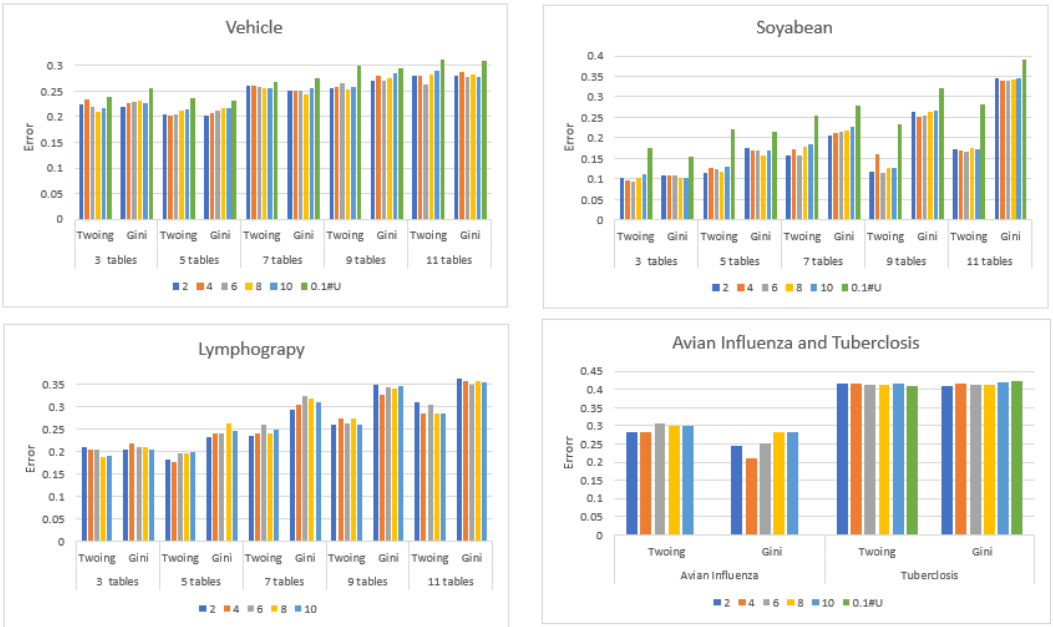

Fig. 2. Comparison of classification error (e) obtained for Twoing criterion and Gini measure.

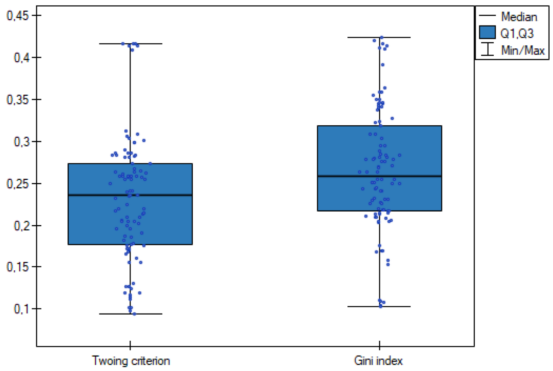

Fig. 3. Box-plot chart with (Median, the first quartile - Q1, the third quartile - Q3) the value of classification error e obtained with using Twoing criterion and Gini index.

5. Conclusions

In this study, the classification method for dispersed data was analyzed. This model can be applied to decentralized information systems. The proposed model complies with federated learning principles, it preserves the privacy of data and consists of building models locally based on the bagging method with CART trees and the Twoing criterion. The central system only aggregates prediction vectors using majority voting. The main aim of the paper was to examine the quality of classification obtained with the use of the Twoing criterion and to analyze impact of various values of the stop criterion in the pre-pruning process on the quality of classification. The paper also compares the obtained results with those obtained using the Gini index.

In total 17 dispersed data sets were analyzed. Two of the analyzed sets reflects the real dispersion of the data. The main conclusions of the research is that; limiting the growth of trees by specifying the minimum number of objects in a leaf up to 10 does not significantly change the quality of classification, while it guarantees a significant simplification of the model resulting from the reduction of the number of nodes and the height of trees. This implies in

dispersed data, the issue of overfitting model is not significant. Furthermore, much better results are generated by the use of the Twoing criterion than the Gini index in the proposed approach. In future research, it is planned to compare the obtained results with other algorithms for decision tree construction and the entropy used as the splitting criterion. Also in further research, instead of using bagging method, random forest would be used in combination with twoing criterion to observed the effectiveness of classification model that could be built.

References

1. Asuncion, A., Newman, D.J. UCI Machine Learning Repository; University of Massachusetts Amherst: Amherst, MA, USA, (2007) Available online: https://archive.ics.uci.edu (accessed on 15 February 2022)
2. Breiman, L.: Technical Note: Some Properties of Splitting Criteria, Machine Learning, vol. 24, 41–47, (1996)
3. Briggs, C., Fan, Z., Andras, P.: Federated learning with hierarchical clustering of local updates to improve training on non-IID data. In 2020 IJCNN, pp. 1–9. IEEE, (2020)
4. Fiebig, L., Soyka, J., Buda, S., Buchholz, U., Dehnert, M., Haas, W., 2011, Avian influenza A(H5N1) in humans - line list, http://dx.doi.org/10.25646/7661 (accessed on 15 February 2022)
5. Janusz, B. J., Wołk, K.: Implementing contextual neural networks in distributed machine learning framework. In ACIIDS, pp. 212-223. Springer, Cham, (2018)
6. Kayri, M., Kayri, I.: The comparison of Gini and Twoing algorithms in terms of predictive ability and misclassification cost in data mining: an empirical study. International Journal of Computer Trends and Technology (IJCTT), vol. 27(1), (2015)
7. Martens, G., De Meyer, H., De Baets, B., Leman, M., Lesaffre, M., Martens, J. P.: Tree-based versus distance-based key recognition in musical audio. Soft Computing, 9(8), 565–574, (2005)
8. Ohene, S.A., 2018, Replication Data for Extra-pulmonary tuberculosis: a retrospective study of patients in Accra, Ghana, https://doi.org/10.7910/DVN/TA1OII, Harvard Dataverse (accessed on 15 February 2022)
9. Połap, D., Woźniak, M.: Meta-heuristic as manager in federated learning approaches for image processing purposes. Applied Soft Computing, 113, 107872, (2021)
10. Przybyła-Kasperek, M., Aning, S.: Bagging and Single Decision Tree Approaches to Dispersed Data. In: ICCS, pp. 420–427. Springer, Cham, (2021)
11. Przybyła-Kasperek, M., Aning, S.: Stop Criterion in Building Decision Trees with Bagging Method for Dispersed Data. Procedia Computer Science, 192, 3560–3569, (2021)
12. Aning, S., Przybyła-Kasperek, M.: Comparative Study of Twoing and Entropy Criterion for Decision Tree Classification of Dispersed Data. Procedia Computer Science, (2022)
13. 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.: Privacy-first health research with federated learning. NPJ digital medicine, 4(1), 1-8, Nature Publishing Group, (2021)
14. Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., Yu, H.: Federated learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 13(3), 1-207, (2019)