Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 2434–2443

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems(KES 2022)

Comparative Study of Twoing and Entropy Criterion for Decision Tree Classification of Dispersed Data

Samuel Aning, Małgorzata Przybyła-Kasperek*

University of Silesia in Katowice, Institute of Computer Science,
Będzińska 39, 41-200 Sosnowiec, Poland

Abstract

In decision tree building, the choice of the splitting criteria highly affects the quality of model that is developed. In this paper, decision tree models are developed on dispersed data using entropy measure and twoing criterion as the splitting criteria. Dispersed data in this sense has multiple independent local tables on which decision tree models are built. Prediction vectors are generated based on the local models and a final prediction is made from aggregation using majority voting. In effort to improve model quality ensemble method technique (bagging) is applied to build multiple models for each local table. The main purpose of this paper is to make a comparative study on the classification quality of decision tree models built on dispersed data using entropy and twoing splitting measure. The main observation is that when knowledge is highly dispersed in a lot of local tables, using twoing criterion in building decision tree models is better than using entropy measure.

Peer-review under responsibility of the scientific committee of the 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

Keywords: Ensemble of classifiers; Entropy; Twoing ; Dispersed data; Decision trees; Stop criterion; Bagging method; Classification trees; Independent data sources

1. Introduction

Decision making process where optimized decision is made based on observed data is common in every aspect of human activity. However, through machine learning models are able to mimic human decision making process and even make more optimized decision than humans could make. Making decision using data from more than a single data sources is known to more effective than data from single data source. Utilizing multiple data sources allows to gain comprehensive understanding of the entire case as well avoid biasness. Also it is common that knowledge on a subject is not limited to one source but collected in fragment by independent units.

* Corresponding author. Tel.: +48-32-368-97-56 ; fax: +48-32-368-97-60.
E-mail address: malgorzata.przybyla-kasperek@us.edu.pl; samuel.aning@us.edu.pl

1877-0509
Peer-review under responsibility of the scientific committee of the 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)
10.1016/j.procs.2022.09.301

However, making decision from multiple data sources also come with some challenges including data security, poor data quality, inconsistency in data among others. Federated learning [9] makes use of data from multiple independent sources to build optimized models. The objective of federated learning is to ensure data privacy while making use of the independently collected data. A typical example of federated learning is mobile keyboard prediction [7] implemented by Google AI team. Here, local models are built independently on local data sets generated from individual devices. These models predict the next word based on context of the expression. Information on the developed model is sent to the server which aggregates from the several models to produce a global model. The global model could then be used by individual devices to predict the next word.

Learning from dispersed data involves building a global model from independent data sources. Here the independent data sources called local tables collect data on the same subject but on attributes of which some may be common. We generally assume that attributes are not identical between local tables. The same assumption applies to objects included in local tables. Since local data collect data on the same subject; it is possible to develop single models to predict classification of using each local data. However the goal is to be able to make use of all the data available from different units to produce a global model which would be more effective. For instance we can consider purchasing behavior classification problem. One local data may be able to access all personal data on the individual customers such as age bracket, gender, race, where another local data may have data on consumer behavior towards such as time spent in aisle, time spent in the mall, usage of cart and having loyalty card. In such case there is possibility that data on an object may be collected in both local tables or more. Also, attributes occurring in local tables are not strict, thus an attribute may be present in more than one local table. Having an effective model that considers multiple local data required novel methods of building the model. Learning from dispersed data becomes necessary because it allows the usage of data from multiple sources while ensuing data privacy because local tables only receive the final global models but not have access to other local tables.

By defining dispersed data sets; 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.

In [10, 11], the successful use of decision trees with the Gini index and the locally applied bagging method for the classification task based on dispersed data was presented. The use of the bagging method does not affect the data privacy assumption in any way, as it is performed locally at the client/participant devices. In [11] limiting the minimum number of objects in each leaf was used as threshold in the decision building process. It was established that this, prepruning methods only affects the classification quality when the minimum number of objects in leaf is very large – equal to 10 percent of the total number of objects. In this paper, it is proposed to use two splitting criteria that have not been used before in the mentioned above classification model for dispersed data. The main objective is to make a comparison of classification quality when the twoing criterion with bagging and entropy criterion with bagging are both applied on dispersed data. It was observed that, in dispersed data, generally using twoing criterion in building decisions trees produces models with better classification quality than using entropy measure. More precisely much better classification quality for twoing criterion than entropy measure is observed for dispersed data when knowledge is highly distributed in many local tables.

Federated learning issues are developed intensively nowadays, as more and more often we deal with decentralized industries organization and dispersed data collection in many areas of life. In addition, data privacy becomes very essential. Various approaches have been used in this subject, for example decision trees [6], neural networks [20] or principal component analysis [4]. In [8], a proposed algorithm, SimFL Gradient Boosting Decision tree model, is applied in federated learning. Though, there was success with respect to time efficiency due to lower computational complexity in decision tree building compared to the baseline method, Gradient Boosting Decision. However, the model suffers low model accuracy. By designing a weighted gradient boosting method, the author utilizes the similarity information between local data sets without sharing sensitive information to build decision trees with bounded errors. In [8], the local data used had the same conditional attributes. However, in this paper dispersed local tables may or may not share same conditional attributes. AdaBoost [15], Gradient Boost [18] or Random Forest [1] are also very popular classification methods using decision trees on a set of local tables. However, these methods are completely different from the model considered in the paper, as these approaches assume independent and identically distribution of samples across local tables. This is due to the process of defining local tables based on one data set.

The structure of the paper is organised as follows. Section 2 is dedicated to the decision trees and the splitting measures compared; entropy and twoing. Section 3 addresses the data sets that are used. Section 4 presents the conducted experiments and the obtained results. Final section gives conclusions and future research plans.

2. Building decision trees

Decision trees are among the most commonly used approach in supervised learning. The best-known algorithm for building decision trees are ID3, CART , C4.5 and CHAID [12, 13, 3, 5]. The paper [16] presents comparative study of the ID3, CART and C4.5 decision tree algorithm. The author offers some advantages and disadvantage of each of the methods. In CART algorithm, one disadvantages is that the model is unstable. Thus for any small change in training data model may be completely different. However, the use data from multiple local tables in our dispersed data attempts to fix this issue. This is because data is spread among multiple local tables, an as such each local table do not have absolute power in the final model developed, a change in training data objects from one local table would not have big impact on the overall training model.

In this paper we make use of the ID3 algorithm which uses the entropy measure and the CART algorithm which uses the twoing criterion. The CART algorithm is able to handle both numerical and categorical data hence whereas the ID3 algorithm handles only categorical datasets. The CART algorithm [3] was first developed with the Gini Index being the splitting criterion. The twoing criterion is also used in the CART algorithm. According to [3] the twoing criterion is recommended when the domain of the target attribute is relatively wide. Also the twoing criterion produces much more equally balanced decision tress than the Gini index.

As decision trees models are non-parametric, one of the common disadvantage is the model being prone to overfitting. Overfitting [19] is a fundamental issue in supervised machine learning which prevents us from perfectly generalizing the models to well fit observed data on training data, as well as unseen data on testing set. The overfitting problem is mainly caused by overly complex classifiers. Another important issue is that decision trees are intrinsically built using the greedy algorithm. Because of the greedy algorithm, data set with unbalanced classes tend to have poorly developed models. Thus objects with the class having low representation tend to be mostly misclassified. As observed in Figure 1 the data sets being used in this paper is generally unbalanced. This is a very common problem in most real life situations. Ensemble methods is one of the ways to solving problem of overfitting and lower representation. An ensemble[14] is a collection of a finite number of models or predictors that are all trained for the same task. The idea used here is that multiple models are developed using subsets of the data sets and the result of the models are aggregated. In this article bagging method is used. Here, a bag of size equal to the number of objects in the original local table is created by randomly selecting with replacements objects from the local table. We experiment different number of bags 10, 20, 30, 40 and 50 to observe the optimal number of bags which gives best classification quality.

2.1. Entropy Measure

In building decision trees, an initial node called the root node is first created by selecting among all the conditionals attribute the best attribute to be used as basis for the split of the objects. Entropy is the measure of the level of disorderliness in a subnode when objects are split based on the conditional attribute. The entropy value for the set of objects $X \subseteq U$ is defined using the formula

$$E(X) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

where p_i is a fraction of objects belonging to decision class i in the set X and c is the number of decision classes. When the number of decision classes is equal to two, the entropy value is between 0 and 1. Entropy equal to 0 shows that there is no disorderliness in the set X whereas entropy equal to 1 implies that the set X is highly disordered. At every level of tree construction various attributes that generate splits based on their values are considered. An attribute with the minimum entropy is selected to define a node. As entropy is used in the ID3 algorithm, models can be created for only categorical training data sets. As advantage the using the entropy measure is effective for multiclass training data sets, and such data are considered in this paper.

Entropy criterion has been used in decision tree models to classify objects on varying topics resulting in very good accuracy rates. For example, in [17] very good results were obtained when decision tree built using entropy measure as splitting criterion was used in customer classification in web marketing.

2.2. Twoing Measure

The twoing criterion is a measure used in the CART algorithm [3] to select the best division of objects into subsplits at every node. The twoing criterion of division $X_1, X_2 \subseteq U$, that is defined based on the attribute $a \in A$ is calculated as follows

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

where p_j^i is a fraction of objects from the i-th decision class in the set X_j, $j \in \{1, 2\}$. The CART algorithm selects the attributes that maximizes twoing splitting rule above.

The main differences between the above criteria can be summarized as follows. Twoing criterion gives a binary sub split at each node whereas entropy criterion may give multiple sub splits depending on the variables in the conditional attribute. Due to this, complexity in terms of height of tree produced when twoing algorithm is used in the decision tree building process is higher than using the entropy algorithm. Twoing criterion handles both categorical and numerical values whereas entropy handles only categorical value. In [16], some differences in characteristics of models built when twoing and entropy is used as splitting criteria can be found. First twoing criterion can handle outliers where as entropy is susceptible on outliers. Lastly twoing criterion handles missing data better than entropy measure.

2.3. Classification model for dispersed data

In this paper, bagging method is fused in the decision tree algorithm to develop the models used for dispersed data classification. Thus a double dispersion of data occurs. First, data is in dispersed form because of the way it is collected and then the generation of set of bootstrap samples based on each local decision table creates a second dispersion. The method used in this paper involved is outlined as;

- Consider dispersed data defined by decision tables $D_{ag} = (U_{ag}, A_{ag}, d)$, $ag \in Ag$ 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.
- First, K bootstrap samples $X_1^{ag}, \dots, X_K^{ag}$ are created based on the training set D_{ag}. Each X_i^{ag} is defined by drawing with replacement from the set U_{ag} to create diversity in data set. A decision tree is built based on each set X_i^{ag}.
- The model built is used on test objects which had conditional attributes from all the local tables.
- The results from each local table are aggregated into a vector. Each vector coordinate corresponds to one decision class and represents the number of decision trees that indicated such a decision for the test object.
- Vectors from each local table are then aggregated using the majority voting method with each local table having the same weight. Thus, for each object, decision classes with the maximum value of the coordinates are selected.

It can be observed from the procedure above that a two-step process of aggregation of results is also used since we had a two-level process of dispersion.

3. Data

In this paper, data sets used in the experimental part was taken from UC Irvine Machine Learning Repository [2]. Four data sets: Lymphography, Soyabean (Large), Primary Tumor and Audiology were not in the dispersed form. Thus a dispersed version was prepared by creating local tables of 3, 5, 7, 9 and 11 for each data set. In each case; when creating dispersed data of a smaller number of local tables more conditional attributes (knowledge of the objects) is stored in a single local table whereas when creating a dispersed data with high number of local tables, say 11, lesser number of attribute (knowledge of the objects) is available to each local table. Also, 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. It was important that the data used from the UCI repository has big number of conditional attributes to allow the dispersed versions to be created. For data sets Soyabean (Large) and Audiology the test sets were obtained directly from the UCI repository. In the case of Lymphography and Primary Tumor data sets from the original set 30% of objects were randomly selected for the test sets.

In Table 1, the characteristics of the data is shown. All the data sets are have categorical attributes thus the entropy measure and the twoing criterion could be used. Figure 1 explores the distribution of number of training objects with their respective classes. It is observed the data is not balanced. However very good results are obtained in terms of classification quality due to the dispersed nature and the use of bagging method.

Table 1. Data set characteristics

Data set	# The training set	# The test set	# Conditional attributes	# Decision classes
Lymphography	104	44	18	4
Soybean	307	376	35	19
Primary Tumor	237	102	17	22
Audiology	200	26	69	24

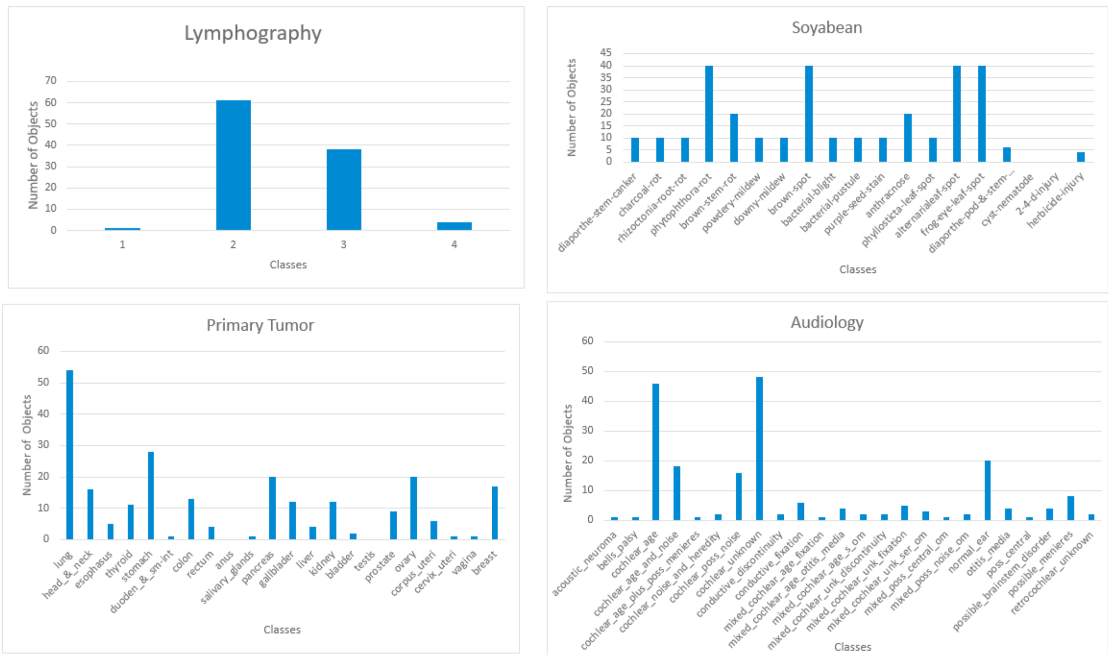

Fig. 1. Distribution of objects from training set to decision classes

4. Experiments and Results

In carrying out the experiments, for each of the data sets five versions of dispersed data had been created, as was described above. As it is of interest to know how the number of bags used in each local data set affect classification quality, different number of bags were tested for the bagging method: 10, 20, 30, 40, 50. First respective bags are created from local tables with the number of objects equal to number of training objects. Then local models are built for each bag. Based on these models initial classification of test objects are done. Next, prediction vectors over decision classes are created for each local table – the vector's coordinate corresponding to decision class i is equal to the number of votes cast by the models for this class. The sum of prediction vectors for all local tables is computed. The decision classes with the maximum value of the coefficient of the obtained vector are the final decisions. In such calculation, there is possibility of a tie between two or more classes for classification of an object.

Two measures were used to evaluate the classification quality: the error rate e and the ambiguity measure d. Ambiguity measure is the sum of the number of decision classes assigned to all test objects divided by the total number of test objects. Since an object may have a tie of two or more possible decision classes, a greater ambiguity measure d implies that more test object had a tie on the class to be assigned and vice versa.

Table 2 shows the error rate e and the ambiguity measure d obtained when twoing criterion and entropy measure is used. In the table, for each number of bags, between the use of twoing criterion and the use entropy measure, we colored in blue which of the two approaches had better classification quality in terms of error rate. It could be observed that when the data is highly dispersed (data in 7, 9 and 11 local tables) twoing criterion gives much better classification quality compared to the used of entropy measure.

In order to investigate the significance in differences of error rate obtained for twoing criteria and entropy measure all results from Table 2 were used. Two dependent samples were created – one containing the results with twoing criteria and one containing the results with entropy measure. Each sample had the cardinality equal to 100 observations – results obtained for different data sets and number of local tables. The Wilcoxon test confirmed that differences in error rate in these two groups are significant, with a level of $p = 0.0001$.

Additionally, comparative box-plot chart for the values of error rate was created (Figure 2). We can observed an increase in error rate when entropy measure is used.

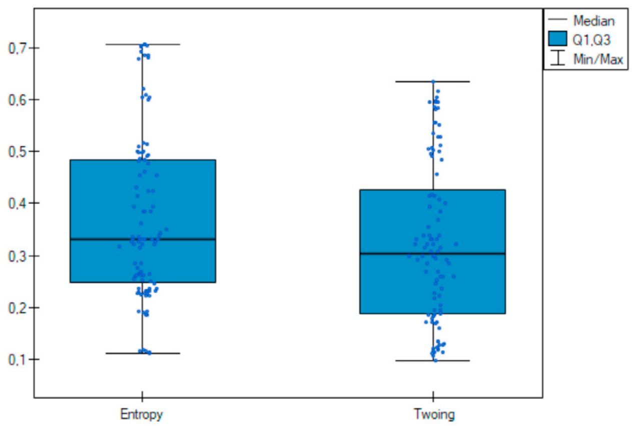

Fig. 2. Box-plot chart with (Median, the first quartile—Q1, the third quartile—Q3) the value of error rate e for the decision tress with different splitting criterion.

Figure 4 and Figure 5 shows color graph results of classification quality measured by errors rate. Horizontal axis for each plot represents the number of bags used, whereas the vertical axis shows the number of local tables the dispersed data has. It could be observed that in most cases both in the use of twoing and entropy, optimal number of bags is between 20 to 40 bags. Moreover, in both approaches, better results are obtained when the data is less dispersed (the number of local tables is equal to 3 or 5).

Additionally, in Figure 3, comparison results of error rate e for bags of 30 is shown. As it could be confirmed that number of bags of 30 is the optimal for decision trees in Figure 4 and Figure 5. It could be confirmed that, generally twoing measure performs better in almost all the data sets and varying number of local tables.

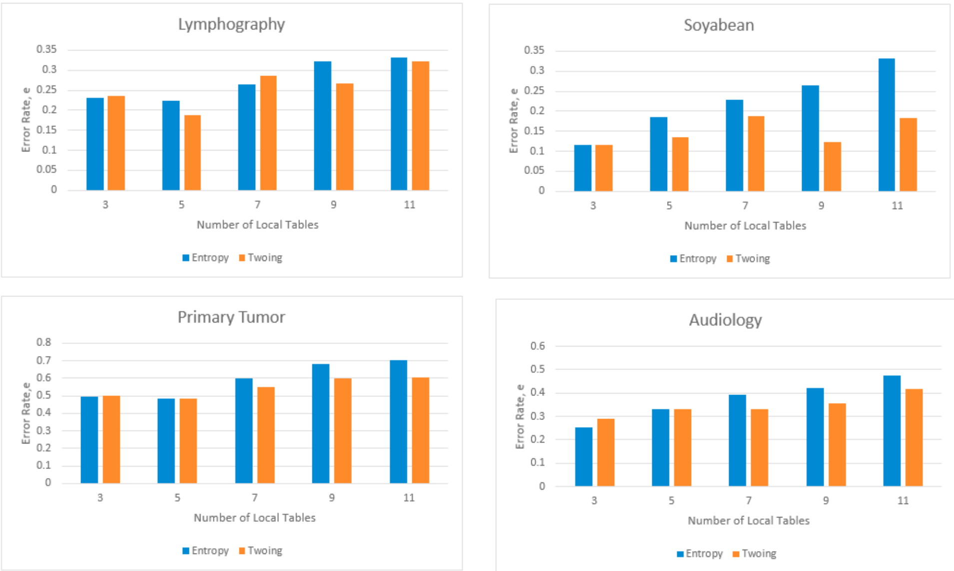

Fig. 3. Comparison of error rate between using entropy and twoing criterion for all data sets when number of bags equal to 30

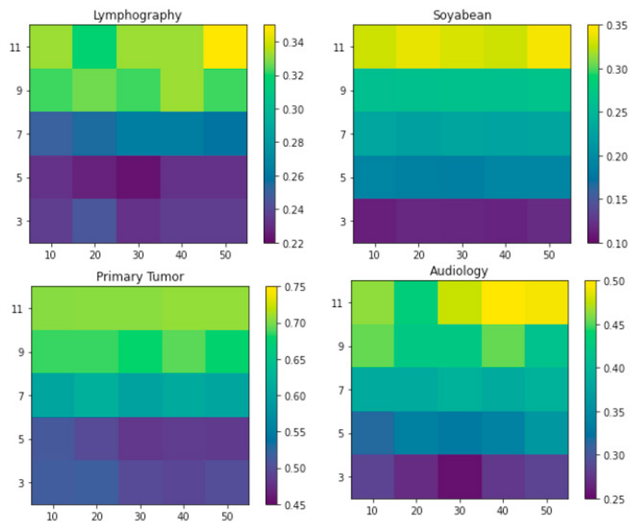

Fig. 4. Color bar plot of error rate for each data sets when entropy measure is used in decision tress with varying bags

5. Conclusion

In this paper, decision tree classification models combined with bagging method is used to classify dispersed data sets. The use of dispersed data is especially important in today's world, when virtually every website, shop, bank or

Table 2. Error rate e and ambiguity measure d for respective number of bags in bagging method, decision trees with entropy and twoing criterion for dispersed data set

No. of local tables	No. of boostrap sample	Lymphpgraphy				Soyabean				Primary Tumor				Audiology			
		Entropy		Twoing		Entropy		Twoing		Entropy		Twoing		Entropy		Twoing	
		e	**d**	**e**	**d**	**e**	**d**	**e**	**d**	**e**	**d**	**e**	**d**	**e**	**d**	**e**	**d**
3	10	0.236	1.014	0.218	1.027	0.112	1.022	0.113	1.034	0.514	1.073	0.500	1.100	0.285	1.046	0.300	1.054
	20	0.245	1.005	0.227	1.005	0.117	1.008	0.119	1.014	0.516	1.037	0.504	1.035	0.269	1.023	0.285	1.015
	30	0.232	1.000	0.236	1.000	0.116	1.003	0.115	1.009	0.498	1.024	0.502	1.041	0.254	1.023	0.292	1.015
	40	0.236	1.009	0.223	1.000	0.115	1.002	0.098	1.009	0.494	1.031	0.512	1.016	0.277	1.015	0.300	1.023
	50	0.236	1.000	0.245	1.005	0.119	1.007	0.107	1.004	0.500	1.016	0.508	1.016	0.285	1.000	0.292	1.023
5	10	0.232	1.018	0.196	1.014	0.190	1.029	0.122	1.034	0.510	1.047	0.529	1.033	0.315	1.054	0.315	1.077
	20	0.227	1.014	0.205	1.005	0.186	1.011	0.127	1.016	0.498	1.019	0.457	1.014	0.336	1.008	0.315	1.031
	30	0.223	1.005	0.187	1.00	0.185	1.007	0.135	1.009	0.482	1.014	0.484	1.023	0.331	1.015	0.331	1.015
	40	0.232	1.005	0.187	1.009	0.193	1.003	0.135	1.025	0.486	1.014	0.490	1.027	0.338	1.008	0.331	1.000
	50	0.232	1.000	0.196	1.000	0.193	1.010	0.131	1.007	0.484	1.008	0.495	1.016	0.362	1.015	0.338	1.000
7	10	0.250	1.014	0.268	1.009	0.229	1.016	0.173	1.025	0.604	1.018	0.555	1.075	0.385	1.000	0.338	1.038
	20	0.255	1.000	0.259	1.009	0.223	1.008	0.161	1.017	0.620	1.008	0.529	1.039	0.385	1.000	0.338	1.030
	30	0.264	1.000	0.286	1.000	0.228	1.005	0.188	1.007	0.600	1.004	0.551	1.012	0.393	1.008	0.331	1.000
	40	0.264	1.000	0.259	1.000	0.225	1.005	0.170	1.005	0.610	1.004	0.555	1.024	0.385	1.015	0.308	1.015
	50	0.259	1.005	0.259	1.000	0.228	1.005	0.171	1.003	0.604	1.010	0.535	1.012	0.393	1.008	0.331	1.000
9	10	0.323	1.014	0.323	1.009	0.261	1.020	0.127	1.018	0.686	1.010	0.635	1.041	0.454	1.015	0.408	1.038
	20	0.327	1.014	0.254	1.000	0.263	1.010	0.115	1.009	0.686	1.000	0.594	1.031	0.423	1.000	0.393	1.038
	30	0.323	1.000	0.268	1.000	0.264	1.005	0.122	1.002	0.680	1.000	0.598	1.010	0.423	1.008	0.354	1.015
	40	0.332	1.000	0.309	1.000	0.263	1.006	0.124	1.008	0.692	1.002	0.594	1.008	0.454	1.008	0.385	1.008
	50	0.323	1.000	0.300	1.005	0.266	1.005	0.116	1.000	0.678	1.000	0.594	1.010	0.415	1.008	0.393	1.015
11	10	0.332	1.005	0.300	1.009	0.328	1.014	0.190	1.026	0.702	1.002	0.584	1.027	0.462	1.023	0.415	1.023
	20	0.318	1.005	0.295	1.009	0.337	1.009	0.179	1.007	0.704	1.000	0.582	1.012	0.431	1.008	0.415	1.023
	30	0.332	1.009	0.323	1.000	0.332	1.005	0.183	1.010	0.704	1.000	0.604	1.007	0.477	1.015	0.416	1.000
	40	0.332	1.000	0.323	1.000	0.328	1.005	0.171	1.009	0.706	1.000	0.616	1.006	0.500	1.000	0.401	1.000
	50	0.350	1.005	0.323	1.000	0.342	1.005	0.183	1.007	0.706	1.000	0.586	1.018	0.492	1.008	0.369	1.000

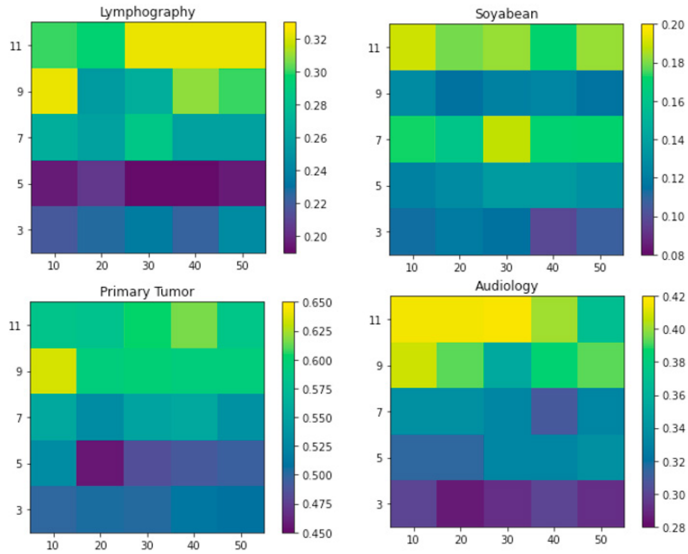

Fig. 5. Color bar plot of error rate for each data sets when twoing criterion is used in decision trees with varying bags

stock exchange has its own data. In real data, there is often an imbalance of decision classes. The occurrence of this problem is difficult to predict in the case of dispersed data and it is difficult to actually apply techniques to balance the data in the case of local tables. Therefore, in this paper, the ensemble method technique (bagging) is used, which allows to mitigate this imbalance to some extent and obtain better results. The proposed approach uses decision trees with different splitting criteria. Both criterions are very different and have different characteristics that were discussed in the paper.

The comparison is made on the classification quality in terms of error rate e between the use of twoing criterion and entropy measure when building the decision trees. Varying number of bags: 10, 20, 30, 40 and 50 are considered when building the decision trees. It can be concluded that generally using bags of 30 gives the optimal classification quality. It is also worth noting that from Table 2, classification quality is better when twoing criterion is used compared of the use of entropy measure especially when data is highly dispersed (data in 7, 9 and 11 local tables). This implies that when data is highly dispersed in a lot of local tables it is best to use twoing criterion than entropy measure to build decision trees.

In a future work, it is planned to use bagging method with weights to reward decision classes that are less well represented in the data set.

References

[1] Ali J., Khan R., Ahmad N., Maqsood I.: Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5), 272, 2012.

[2] 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 April 2022)

[3] Breiman L., Friedman J., Stone C., Olshen R.: Classification and regression trees. Wadsworth International Group 1984.

[4] Grammenos A., Mendoza Smith R., Crowcroft J., Mascolo, C.: Federated principal component analysis. Advances in Neural Information Processing Systems, 33, 6453-6464, 2020.

[5] Kass G. V.: An Exploratory Technique for Investigating Large Quantities of Categorical Data. Journal of the Royal Statistical Society, Series C (Applied Statistics), 29(2), 119–127, 1980.

[6] Kwatra S., Torra V.: A k-Anonymised Federated Learning Framework with Decision Trees. In Data Privacy Management, Cryptocurrencies and Blockchain Technology, 106–120, Springer, Cham 2021.

[7] Li L., Fan Y., Tse M., Lin K.: A review of applications in federated learning, Computers & Industrial Engineering, Volume 149, 2020, 106854, ISSN 0360-8352.

[8] Li Q., Wen Z., He B.: Practical Federated Gradient Boosting Decision Trees. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), 4642–4649. https://doi.org/10.1609/aaai.v34i04.5895, 2020.
[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] Quinlan J. R.: Improved Use of Continuous Attributes in C4. 5. Journal of artificial intelligence research, 4, 77–90, 1996.
[13] Quinlan J. R.: Induction of Decision Trees. Machine Learning, 1(1), 81–106, 1986.
[14] Sollich P., Krogh A.,: Learning with ensembles: How over-fitting can be useful. In Proceedings: Advances in Neural Information Processing Systems 8, NIPS, Denver, CO, November 27-30, 1995
[15] Schapire R. E.: Explaining adaboost. In Empirical inference, 37–52, Springer, Berlin, Heidelberg, 2013.
[16] Singh S., Gupta P.: Comparative Study ID3, CART and C4.5 Decision Tree Algorithm: A survey, International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682 Vol.27, No.27, 2014.
[17] Xiaowei L.: Application of decision tree classification method based on information entropy to web marketing. In 2014 Sixth International Conference on Measuring Technology and Mechatronics Automation, 121–127, IEEE. doi: 10.1109/ICMTMA.2014.34. 2014.
[18] Ye J., Chow J. H., Chen J., Zheng Z.: Stochastic gradient boosted distributed decision trees. In Proceedings of the 18th ACM conference on Information and knowledge management, 2061–2064, 2009.
[19] Ying X.: An Overview of Overfitting and its Solutions, Journal of Physics: Conference Series, IOP Publishing, 1742-6588, 2019.
[20] Yurochkin M., Agarwal M., Ghosh S., Greenewald K., Hoang N., Khazaeni Y.: Bayesian nonparametric federated learning of neural networks. In International Conference on Machine Learning, 7252–7261, PMLR, 2019.