Bagging and single decision tree approaches to dispersed data

Małgorzata Przybyła-Kasperek[0000−0003−0616−9694] and Samuel Aning[0000−0002−3061−3081]

University of Silesia in Katowice, Institute of Computer Science,
Będzińska 39, 41-200 Sosnowiec, Poland
{malgorzata.przybyla-kasperek,samuel.aning}@us.edu.pl
https://us.edu.pl/

Abstract. The article is dedicated to the issue of classification based on independent data sources. A new approach proposed in the paper is a classification method for independent local decision tables that is based on the bagging method. For each local decision table, sub-tables are generated with the bagging method, based on which the decision trees are built. Such decision trees classify the test object, and a probability vector is defined over the decision classes for each local table. For each vector decision classes with the maximum value of the coordinates are selected and the final joint decision for all local tables is made by majority voting. The results were compared with the baseline method of generating one decision tree based on one local table. It cannot be clearly stated that more bootstrap replicates guarantee better classification quality. However, it was shown that the bagging classification trees produces more unambiguous results which are in many cases better than for the baseline method.

Keywords: Ensemble of classifiers · Dispersed data · Bagging method · Classification trees · Independent data sources

1 Introduction

Classification based on data provided by independent sources is a current and challenging issue. Data can be provided by different sensors, collected in separate data sets or provided by various devices/units. Methods for classification based on independent sources were used for example in the streaming domain [4], transfer learning [11], medicine [9], land cover identification [1] and others. In this paper, independent data sources are decision tables that are collected independently by various units/entities/agents. They must relate to the same domain and have a common decision attribute, besides these requirements, there are no other restrictions as to their form. Decision trees are one of the most popular methods used in classification problems. The best-known algorithm for building decision trees are ID3, C4.5, CART and CHAID [2]. Bagging, boosting and random forests [8] methods are the next stage in the development of decision trees.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_35

The novelty of the work is the use of the bagging method and classification trees for independent data sources stored in local decision tables. We use the bagging method as it is simple, often produces very good results, and works well with diverse data (so appropriate for independent data). We are dealing with a two-level division of the data in the method. The first level concerns the independent way of collecting data. The second level concerns the using of the bagging method. Based on each local table, we generate sub-tables from which decision trees are constructed to define prediction vectors. Based on these vectors, the votes for the decisions with the maximum vector's coordinates are calculated. At the end, the majority voting method combines the classification results for local tables.

The structure of the paper is organised as follows. Sect. 2 is dedicated to the classification tree and the bagging method. Sect. 3 contains a description of the proposed approach. Sect. 4 addresses the data sets that are used. Sect. 5 presents the conducted experiments and comments on the obtained results. Sect. 6 gives conclusions and future research plans.

2 Classification tree and bagging method

ID3 and CART, were the first algorithms to be proposed independently by Quinlan and by the team Breiman, Friedman, Stone and Olshen [2]. Both algorithms are greedy. At first, the full training set of objects is considered, and the division of the set defined by the conditional attributes is optimized to a specific measure. The two most popular measures are information gain and the Gini index. The Gini index measures the purity of the set $Gini(X) = 1 - \sum_{i=1}^{m} p_i^2$, where m is the number of decision classes, $p_i = \frac{|C_{i,X}|}{|X|}$, $|X|$ is the size of the training set X and $|C_{i,X}|$ is the number of objects from the i-th decision class. The Gini index of division X_1, X_2, that is defined based on the attribute a is calculated as follows $Gini_a(X) = \frac{|X_1|}{|X|}Gini(X_1) + \frac{|X_2|}{|X|}Gini(X_2)$. The tree induction algorithm selects a conditional attribute that minimizes the Gini index. The minimum number of objects in a given node or the maximum tree height is used as a stop condition - we stop splitting the nodes and defined leaves. In an ensemble of classifiers approach, each new test object is classified by a set of classifiers, and the final decision is made by majority voting. The most popular ensemble methods proposed are: bagging, boosting and random forests methods [8]. In the bagging method K bootstrap samples $X_1, \ldots, X_K$ are created based on the training set X. Each X_i is defined by drawing with replacement from the set X to create diversity in data set. A decision tree is built based on each set X_i. For the test object decisions trees make decisions and the most common decision is chosen. Usually this approach improves the quality of the classification and is more resistant to outliers and overfitting.

3 Classification tree applied for dispersed data

Dispersed data is a set of local decision tables that share the same decision attribute. $S_i = (U_i, A_i, d)$, $i \in \{1, \ldots, N\}$ is called a $i-$th decision table, where

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_35

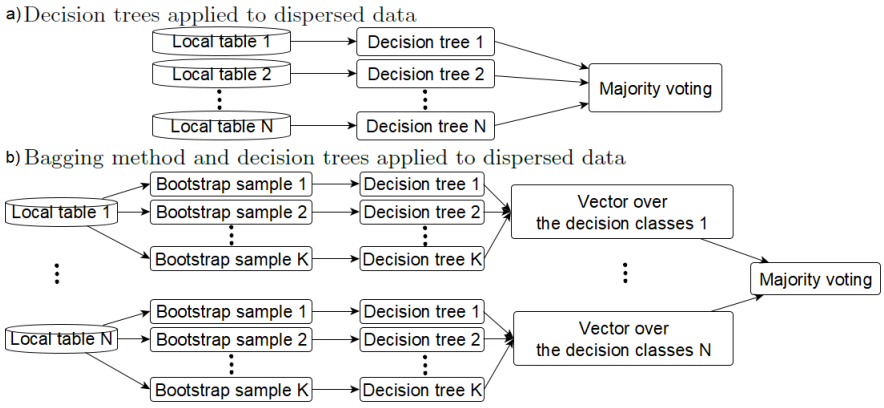

Fig. 1. Graphical representation of the proposed approaches

U_i is the universe, a set of objects of $i-$th decision table, A_i is a set of condition attributes of $i-$th decision table and $d \notin \bigcup_{i=1}^{N} A_i$ is a decision attribute. Both the conditional attribute sets and the sets of objects in such local tables have no restrictions or constrains.

Research on the use of dispersed data was conducted in earlier papers [5–7]. These articles mainly used the k-nearest neighbors classifier to generate decisions based on dispersed data. In this paper, for the first time, a decision tree-based classifier has been used for dispersed data. Two approaches are used in this article. The first, simpler, approach consist of building one decision tree based on each local decision table. For this purpose, the Gini index and the stop criterion determined by two objects in the node were used. When a test object is classified, each decision tree votes for one decision value, the final decision is made by majority voting. A graphic illustration of this approach is presented on Figure 1.

In the second approach, the bagging method was used for each local decision table. Based on the bootstrap samples the decision trees are built in analogous way to that described above. It should be noted that a double dispersion of data occurs in this approach. Firstly, data is in dispersed form because it is collected by independent units. Secondly, a set of bootstrap samples is generated based on each local decision table. Since we have a two-level process of dispersion in this approach, a two-step process of aggregation of results is also used. The results of classification obtained based on the bootstrap samples and decision trees is aggregated into a vector over the decision classes. Each vector coordinate corresponds to one decision class and represents the number of decision trees that indicated such a decision for the test object. In this study, the majority voting method is used to aggregate such vectors. Votes are calculated based on each vector, each decision class with the maximum value of the coordinates is given a voice. Then, the final decisions are the classes that received the maximum number of votes defined based on all vectors. This aggregation method can generate ties. A graphic illustration of this approach is presented on Figure 1.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_35

4 Description of the data and experiments

In the experimental part of the paper, three data sets were used: the Vehicle Silhouettes, the Soybean (Large) and the Lymphography data sets. All of these data are available in a non-dispersed version (single decision table) at the UC Irvine Machine Learning Repository. The quality of the classification for the proposed approaches was evaluated by the train and test method. The analysed data sets are multidimensional (from 18 to 35 conditional attributes) and have several decision classes (from 4 to 19 decision classes). Dimensionality is important because a set of local decision tables are created based on each training sets. The dispersion on 3, 5, 7, 9 and 11 local tables were created such that the lesser number of local tables, the greater the number of attributes in the tables. All local tables contain the full set of objects. The large number of decision classes in the analysed data sets is important because we allow that ties occur. The application of the proposed approaches to data with missing values is also analysed in this article. Missing values occur in the Soybean data set. Four different approaches to dealing with such data were analysed: the global dominant method (Gl) - for each attribute the dominant value based on all values in the table is selected and objects with missing values are supplemented with this dominant value; the dominant in relation to the decision classes method (Dec) - the dominant values are determined separately for each decision class and each attribute; objects with more than 50% of conditional attributes with missing values are removed (50%); all objects, no matter how many missing values they have, are used (all).

4.1 Approach with single decision tree created based on local table

For all data sets based on each local table, a decision tree was built using the Python language and the function sklearn.tree.DecisionTreeClassifier. The final decision was made using majority voting - ties may occur. Therefore, the following two measures are used: the classification error e – the fraction of the number of misclassified objects by the total number of objects in the test set; the average number of generated decisions sets $\bar{d}$. Results of classifications are considered unambiguous if $\bar{d} = 1$ and otherwise when $\bar{d} > 1$. The results are given in Table 1. Based on the presented results, it can be concluded that for the Lymphography data set, unambiguous results are mostly obtained. There is ambiguity for the Vehicle data set, however, it is small and acceptable. Greater ambiguity occurs for the Soybean data set. However, this data set has 19 decision classes, so such ambiguity is acceptable. A graph (Figure 2) was created. It can be seen that for a smaller number of local tables, rather better classification quality were achieved. This is due to the fact that in the case of greater dispersion, the number of conditional attributes in local tables is smaller. Decision trees that are built based on a very small number of conditional attributes do not provide good classification quality. When we consider different methods of completing the missing values in the Soybean data set, it is observed that the

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_35

Table 1. Classification error e and the average number of generated decisions $\bar{d}$ for decision trees created directly based on local tables.

No. of local tables	Data sets											
	Vehicle		Lympho-graphy		Soybean Gl all		Soybean Gl 50%		Soybean Dec all		Soybean Dec 50%	
	e	$\bar{d}$	e	$\bar{d}$	e	$\bar{d}$	e	$\bar{d}$	e	$\bar{d}$	e	$\bar{d}$
3	0.220	1.260	0.205	1	0.090	1.457	0.068	1.450	0.074	1.388	0.075	1.352
5	0.224	1.169	0.182	1.045	0.205	1.258	0.208	1.283	0.189	1.274	0.199	1.303
7	0.264	1.134	0.250	1.045	0.223	1.189	0.218	1.235	0.221	1.152	0.212	1.218
9	0.287	1.154	0.318	1	0.218	1.231	0.189	1.235	0.207	1.234	0.182	1.241
11	0.276	1.146	0.409	1	0.335	1.225	0.332	1.215	0.324	1.197	0.326	1.215

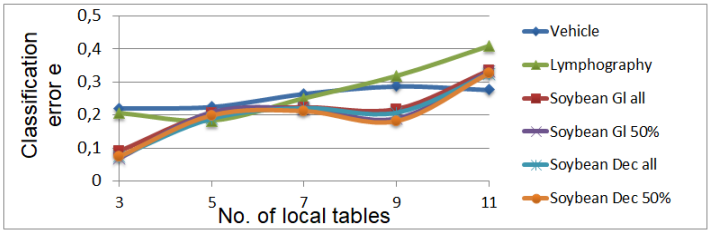

Fig. 2. Classification error e for decision trees created directly based on local tables

method with the global dominance (Gl) and leaving all objects (all) produces the worst results.

4.2 Approach with bagging method and decision trees

One of the experimental goals was to check the impact of the number of bootstrap samples on the quality of classification. Therefore different number of samples were tested for the bagging method: 10, 20, 30, 40, 50. All evaluations were performed five times due to the indeterminism of generating bootstrap samples. The results that are given in Table 2 are the average value of the results obtained from these five runs. In Table 2, the best results, in terms of classification error e, within each data set and the number of local tables are shown in blue. If for two different numbers of bootstrap samples the same values of classification error were obtained, the result with a smaller number of the $\bar{d}$ measure was selected. As before, better results are obtained for a smaller number of local tables.

From Table 2 it can be noticed that the higher the number of bootstrap samples is, the lesser the $\bar{d}$ measure occur. In order to compare the obtained results for two proposed approaches Table 3 was created. In this table, for each dispersed data set, the lowest values of the classification error obtained for both approaches are given. As can be seen for the Vehicle data set the approach (2) – the bagging method produces better results. Unlikely results of Lymphography is because decision trees built from bagging method is not able to learn well from its small number of objects. For the Soybean data set, in the case of 5

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_35

Table 2. Classification error e and the average number of generated decisions $\bar{d}$ for bagging method, decision trees and the two-step process of aggregation results.

No. of local tables	No. of bootstrap samples	Data sets											
		Vehicle		Lympho-graphy		Soybean Gl all		Soybean Gl 50%		Soybean Dec all		Soybean Dec 50%	
		e	$\bar{d}$	e	$\bar{d}$	e	$\bar{d}$	e	$\bar{d}$	e	$\bar{d}$	e	$\bar{d}$
3	10	0.226	1.030	0.214	1.014	0.121	1.021	0.113	1.015	0.122	1.026	0.110	1.016
	20	0.220	1.017	0.205	1.005	0.120	1.016	0.117	1.005	0.123	1.016	0.110	1.010
	30	0.225	1.010	0.205	1.009	0.132	1.009	0.114	1.003	0.131	1.003	0.116	1.005
	40	0.219	1.010	0.209	1.005	0.124	1.009	0.114	1.004	0.135	1.006	0.117	1.005
	50	0.227	1.010	0.205	1	0.131	1.005	0.117	1.006	0.132	1.005	0.114	1.004
5	10	0.214	1.021	0.232	1.009	0.180	1.043	0.183	1.029	0.200	1.047	0.170	1.026
	20	0.203	1.013	0.250	1	0.176	1.023	0.178	1.011	0.192	1.029	0.170	1.016
	30	0.216	1.006	0.245	1.014	0.188	1.015	0.175	1.006	0.188	1.014	0.186	1.013
	40	0.204	1.003	0.245	1	0.178	1.006	0.177	1.010	0.190	1.013	0.183	1.009
	50	0.205	1.005	0.241	1	0.175	1.004	0.177	1.045	0.192	1.015	0.180	1.007
7	10	0.251	1.017	0.323	1.023	0.220	1.014	0.207	1.023	0.216	1.023	0.213	1.016
	20	0.254	1.010	0.295	1	0.207	1.010	0.206	1.008	0.218	1.012	0.219	1.008
	30	0.255	1.008	0.318	1.023	0.214	1.008	0.208	1.005	0.215	1.007	0.207	1.005
	40	0.258	1.007	0.332	1.005	0.209	1.006	0.208	1.007	0.212	1.003	0.217	1.004
	50	0.255	1.008	0.323	1	0.216	1.004	0.215	1.006	0.212	1.004	0.218	1.005
9	10	0.277	1.017	0.354	1.018	0.279	1.013	0.267	1.020	0.278	1.018	0.256	1.020
	20	0.286	1.007	0.359	1.005	0.287	1.010	0.270	1.012	0.275	1.011	0.262	1.018
	30	0.288	1.006	0.355	1	0.279	1.003	0.273	1.005	0.282	1.007	0.256	1.004
	40	0.270	1.006	0.359	1.005	0.283	1.005	0.267	1.005	0.285	1.005	0.252	1.005
	50	0.277	1.009	0.350	1	0.284	1.002	0.263	1.007	0.283	1.002	0.266	1.005
11	10	0.280	1.019	0.373	1.009	0.352	1.015	0.349	1.018	0.360	1.014	0.339	1.012
	20	0.288	1.005	0.386	1.014	0.349	1.010	0.350	1.005	0.358	1.007	0.338	1.007
	30	0.291	1.002	0.368	1	0.348	1.006	0.345	1.004	0.361	1.004	0.341	1.007
	40	0.286	1.003	0.364	1.009	0.346	1.005	0.353	1.003	0.363	1.004	0.341	1.004
	50	0.294	1.006	0.386	1	0.349	1.003	0.347	1.006	0.361	1.001	0.343	1.005

and 7 local decision tables the approach (2) provides better results, for the remaining versions of dispersion the approach (1) gives better results. However, it should be noted that for the Soybean data set, better results for the approach (1) are generated with greater ambiguity. Therefore, in applications where the unambiguity of the generated decisions matters, the bagging method should be used. Based on Table 3 it can also be concluded that for almost all cases the best quality of classification is obtained with using the dominant value in relation to the decision class and removing objects with more than half of attributes with the missing values. Paper [5] presents the results obtained by direct aggregation of the predictions generated by the k-nearest neighbors classifier instead of the decision trees. It can be concluded that when using decision trees and bagging method, better results were obtained in most cases.

5 Conclusions

In this paper, two new approaches on applying decision trees to dispersed data were presented: the approach with decision trees directly generated based on local tables and the approach with the bagging method and decision trees. It was found that the bagging method gives more unambiguous results than the method based on the direct generation of decision trees based on local tables.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_35

Table 3. Comparison of classification error (e) for approaches: (1) single decision tree created based on one local table vs. (2) bagging method with decision trees

No. of local tables	Data sets											
	Vehicle		Lympho-graphy		Soybean Gl all		Soybean Gl 50%		Soybean Dec all		Soybean Dec 50%	
	(1) e	(2) e	(1) e	(2) e	(1) e	(2) e	(1) e	(2) e	(1) e	(2) e	(1) e	(2) e
3	0.220	0.219	0.205	0.205	0.090	0.120	0.068	0.113	0.074	0.122	0.075	0.110
5	0.224	0.203	0.182	0.232	0.205	0.175	0.208	0.175	0.189	0.188	0.199	0.170
7	0.264	0.251	0.250	0.295	0.223	0.207	0.218	0.206	0.221	0.212	0.212	0.207
9	0.287	0.270	0.318	0.350	0.218	0.279	0.189	0.263	0.207	0.275	0.182	0.252
11	0.276	0.280	0.409	0.364	0.335	0.346	0.332	0.345	0.324	0.358	0.326	0.338

Moreover, it was noticed that the higher the number of bootstrap samples is, the lesser the $\bar{d}$ measure occur. When dealing with missing data, it was found that the method with the dominant value in relation to the decision class and removing objects with more than half of attributes with the missing values provide the best results. In future work, it is planned to analyse various parameters when building decision trees (different stop conditions and applying information gain). It is also planned to use other fusion methods to combine the predictions of decision trees.

References

1. Elmannai, H., Salhi, A., Hamdi, M., Sliti, M., Algarni, A.D., Loghmari, M.A., Naceur, M.S.: Rule-based classification framework for remote sensing data, J. Appl. Remote Sens. 13(1), 014514 (2019)
2. Kotsiantis, S. B.: Decision trees: a recent overview. Artif. Intell. Rev., 39(4), 261–283 (2013)
3. Kuncheva, L. I.: Combining Pattern Classifiers: Methods and Algorithms. Hoboken, NJ, USA: Wiley (2004)
4. Li, Y., Li, H.: Online transferable representation with heterogeneous sources. Appl Intell 50, 1674–1686 (2020)
5. Przybyła-Kasperek, M.: Generalized objects in the system with dispersed knowledge. Expert Syst. Appl., 162, 113773 (2020)
6. Przybyła-Kasperek, M.: Coalitions' Weights in a Dispersed System with Pawlak Conflict Model. Group Decis. Negot., 1–43 (2020)
7. Przybyła-Kasperek, M., Wakulicz-Deja, A.: Dispersed decision-making system with fusion methods from the rank level and the measurement level – A comparative study. Inf. Syst., 69, 124–154 (2017)
8. Sagi, O., Rokach, L.: Ensemble learning: A survey. Wiley Interdiscip Rev Data Min Knowl Discov, 8(4), e1249 (2018)
9. Giger, M. L.: Machine learning in medical imaging. J. Am. Coll. Radiol., 15(3), 512–520 (2018)
10. Wang, L., Zhou, D., Zhang, H., Zhang, W., Chen, J.: Application of relative entropy and gradient boosting decision tree to fault prognosis in electronic circuits. Symmetry, 10(10), 495 (2018)
11. Wu, Q., Wu, H., Zhou, X., Tan, M., Xu, Y., Yan, Y., Hao, T.: Online transfer learning with multiple homogeneous or heterogeneous sources. IEEE Trans. Knowl. Data Eng., 29(7), 1494–1507 (2017)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_35