Rules' Quality Generated by the Classification Method for Independent Data Sources Using Pawlak Conflict Analysis Model

Małgorzata Przybyła-Kasperek[1][0000−0003−0616−9694] and Katarzyna Kusztal[1][0000−0002−9970−5339]

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

Abstract. The study concerns classification based on dispersed data, more specifically data collected independently in many local decision tables. The paper proposes an approach in which coalitions of local tables are generated using Pawlak's conflict analysis model. Decision trees are built based on tables aggregated within the coalitions The paper examines the impact of the stop criterion (determined by the number of objects in a node) on the quality classification and on the rules' quality generated by the model. The results are compared with the baseline approach, in which decision trees are built independently based on each of the decision tables. The paper shows that using the proposed model, the generated decision rules have much greater confidence than the rules generated by the baseline method. In addition, the proposed model gives better classification quality regardless of the stop criterion compared to the non-coalitions approach. Moreover, the use of higher values of the stop criterion for the proposed model significantly reduces the length of the rules while maintaining the classification accuracy and the rules' confidence at a high level.

Keywords: Pawlak conflict analysis model · independent data sources · coalitions · decision trees · stop criterion · dispersed data.

1 Introduction

Nowadays, huge volumes of data are often dealt with, most dispersed and available from various independent sources. Processing such fragmented data is a big challenge both in terms of achieving satisfactory classification quality and understanding the basis for generated predictions. In particular, clear presentation of the extracted knowledge is difficult when we deal with dispersed data. It is assumed that fragmented data is available as a set of local decision tables collected by independent units. For real dispersed data the simple aggregation is not possible – it leads to inconsistencies and conflicts. In this study, we deal with an approach that generates coalitions and makes it possible to aggregate local tables for coalitions. Another important issue addressed in this paper is the

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

analysis of the quality of the patterns generated based on dispersed data. Studies are presented on how valuable the knowledge generated by the proposed model is.Among the many different machine learning models, those that are characterized by high interpretability and transparency of the generated knowledge can be distinguished. Approaches such as rule–based models [29, 18] or decision trees [7] can be mentioned here. There are also models that give a very good quality of classification, but without the simple interpretation and understanding the basis of the decisions made. Neural networks and deep learning can be mentioned here [8, 6]. In many real–world applications, the ability to justify the prediction is very important. Sometimes the knowledge generated by the model itself is more of a value than the prediction of new unknown objects.In the case when data is available in the form of multiple decision tables, the situation is even more difficult. For example, let's consider a medical application. Each clinic has a separate decision table. Models can be generated separately based on the local tables, but the knowledge generated in this way will be fragmented and will not allow to make simple interpretations. Aggregating all local tables into one table is not possible due to inconsistencies that occur. The proposed model can be applied in such situation.

This paper proposes an approach that gives a more condensed representation of knowledge, as it allows for partial aggregation of local tables. The proposed approach is designed for set of local tables in which the conditional attributes are identical. Coalitions of decision tables that store compatible values are determined. Identification of such coalitions is performed using Pawlak's conflict model. Aggregation of tables from one coalition is performed for a coalition. This allows, on the one hand, to combine data and reduce fragmentation and on the other hand to provide for the model a larger training set based on witch certain patterns/knowledge are generated. A decision tree is built based on partially aggregated data. Decision tree model was chosen because of the transparency and easy interpretability of the generated result. By using the proposed approach and doing data aggregation still, a few decision trees for dispersed data is received. However, as research shows, the number of obtained trees is smaller than if we generate trees separately based on individual local tables. Coalitions are not disjoint, which means that one local table can be included in several coalitions at the same time. Generated coalitions of local tables represent consistent data on common concepts. Decision trees describing a single concept are generated based on coalition – complete information about coherent knowledge, which also contributes to trees' quality. The paper [22] proposed the first study on this subject, in which for the basic approach (without the model's parameter optimizing) the classification quality was analyzed. The paper focused not only on the classification accuracy but also on the knowledge that the proposed model generates. Based on decision trees, decision rules are determined. Various parameters evaluating the quality of generated rules are analyzed, i.e. support, confidence and rule length. The main contribution and the novelty of this paper is a comprehensive analysis of the quality of rules generated by the model as well as the analysis of the impact of the stop criterion used during the construction of the tree on

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

the accuracy of classification and the quality of decision rules generated based on decision trees. The size of the constructed trees has a very significant impact on the time and memory complexity of the method. The presented study shows that using the proposed model, the generated decision rules have much greater confidence than rules generated by using decision trees for each decision table separately. In addition, the proposed model gives better classification accuracy regardless of the stop criterion compared to the non-coalitions approach. The use of higher values of the stop criterion for the proposed model significantly reduces the length of the rules while maintaining the classification accuracy and the rules' confidence at a high level. The study presented in the paper proves that the proposed classification model not only provides high classification accuracy but also generates more interesting knowledge – rules with large confidence – than is the case when coalitions are not used.

The paper is organized as follows. Section 2 contains a literature review. In Section 3, the proposed classification dispersed system and Pawlak's conflict model are described. Section 4 addresses the data sets that were used, presents the conducted experiments and discussion on obtained results. Section 5 is on conclusions and future research plans.

2 Related Works

Distributed data is used in the concept of ensemble of classifiers [3, 9] and federated learning [19]. In the classifier ensembles approach, the interpretation and the justification of the results obtained is not obvious and clear. There is no aggregation of data, the models are built locally in a hierarchical [34] or parallel manner [10], and the final decision is generated by the chosen fusion method [13]. There are many different fusion methods, some are based on fuzzy set theory [26] or mathematical measures [11]. Others fusion methods are based on issues related to evidence theory [27] and voting power [21]. However, a simple and concise justification with using some pattern like decision rule or decision tree is not possible. In the classifier ensembles approach, there is no aggregation of data and no effort is given to generate a concise representation of knowledge. On the other hand, in federated learning a common model is built, but without local data exchange, as data protection is one of the primary issues in this approach [12]. In this approach, only models' parameters are exchanged (models are generated locally) to a central space/server. The models are then aggregated and sent back to the local destinations. This iterative algorithms leads to convergence and agreement on a common central model [15]. The approach proposed in this paper can be seen as an intermediate solution. No single common model is generated as in federated learning. Instead, several common models are generated, one for each coalition. A coalition is a set of local tables containing compatible data – data on a single concept. Another difference between the proposed approach and the federated learning approach is the exchange of data within the coalition. Here, the model for a coalition is generated based on aggregated table. The proposed approach also differs from the ensemble of classifiers approach, as

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

it is based on coalitions formation and data aggregation. The second important difference is the form of local tables. In the classifier ensembles, it is assumed that initially the data is accumulated in a single decision table. The process of fragmentation/dispersion is realized in some controlled and planned way so as to increase the accuracy of the ensemble classification [5]. In contrast, in the proposed approach, we have no control over the form of local tables. They are provided by independent units. They do not satisfy the constraints of separability or equality of sets of objects. Nor is it possible to guarantee diversity or focus on the most difficult objects as it is in the case with many approaches known from classifier ensembles [14].

The method of conflict analysis that is used in the study was proposed in [17] by Pawlak. This approach was chosen for its simplicity and very good capabilities in identifying natural coalitions [25, 24]. In the approach one of only three values is assigned to conflicting issues by the conflicting sides. These three values correspond to three possible views: support for the issue, neutrality towards the issue and being against the issue. This relatively simple approach provides tools for efficient conflict analysis and to determine the sets of coalitions and the strength of the conflict intensity. Pawlak's conflict analysis approach is very popular and widely used and developed, for example, in the three-way decisions [33] or in approach proposed by Skowron and Deja in [28]. In papers [25, 24] Pawlak's conflict analysis was also used for dispersed data. However, the approach proposed there is quite different from the one considered in this paper. The main differences are the form of the local tables that are being considered and the basis for recognizing the conflict situation. In papers [25, 24] it was assumed that the sets of conditional attributes appearing in the local tables are of any form, no restrictions are imposed. In contrast, in the present study, we assume that the conditional attributes are the same in all local tables. In papers [25, 24] conflicts were considered in terms of decisions made by local classifiers defined based on local tables – the $k-$nearest neighbor classifiers were used there. In the present study, in contrast, no pre-determined classifiers were used. The basis for calculating the strength of conflicts is the values that are present in the local tables. We form coalitions in terms of a common concept – compatible values that are present in these tables.

In this study, rule generation and knowledge representation are very important concepts. In the literature, there are two main approaches to generate rules: directly from data in tables or from decision trees. A decision rule consists of two parts: a premise and a conclusion. We assume that rules are in the form of Horn clauses, i.e., the premise is a conjunction of conditions on conditional attributes, while the conclusion is a single value of a decision attribute. There are many algorithms for generating decision rules: based on the rough set theory [30], based on covering [4] or associative approaches [32]. We distinguish between exhaustive [16] and approximate [31] methods of rule generation. Decision rules may also be generated based on decision trees. The division criterion is crucial when building decision trees. The most popular are the Gini index, entropy and statistical measures approach [20]. A method of limiting tree's growth and thus

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

overfitting of the tree is using the stop criterion [23]. In this way we can generate approximate rules. The best situation would be to generate high quality and short rules (short means with a minimum number of conditions in the premise). Generating minimal rules is an NP-hard problem, so it is possible to apply the algorithm only to small decision tables. In the literature, we can find various measures to determine the quality of rules, among others, we distinguish confidence and support, gain, variance and chi-squared value and others [2]. In this study, confidence, support and rule length are used to determine rules quality.

3 Model and methods

In this section, we discuss the proposed hierarchical classification model for dispersed data. This model was first considered in the paper [22], where a detailed description, discussion of computational complexity and an illustrative example can be found. In the model, we assume that a set of local decision tables (collected independently, containing inconsistencies) with the same conditional attributes are given $D_i = (U_i, A, d), i \in \{1, \ldots, n\}$, where U_i is the universe, a set of objects; A is a set of conditional attributes; d is a decision attribute. Based on the values of conditional attributes stored in local tables, coalitions of tables containing compatible data are defined. For this purpose, Pawlak's conflict model is used [17, 22]. In this model, each agent (in our case a local table) determines its view of a conflict issue by assigning one of three values $\{-1, 0, 1\}$. The conflict issues will be conditional attributes, and the views of local tables will be assigned with respect to the values stored in the tables. For each quantitative attribute $a_{quan} \in A$, we determine the average of all attribute's values occurring in local table D_i, for each $i \in \{1, \ldots, n\}$. Let us denote this value as $\overline{Val}^i_{a_{quan}}$. We also calculate the global average and the global standard deviation. Let us denote them as $\overline{Val}_{a_{quan}}$ and $SD_{a_{quan}}$. For each qualitative attribute $a_{qual} \in A$, we determine a vector over the values of that attribute. Suppose attribute a_{qual} has c values $val_1, \ldots, val_c$. The vector $Val^i_{a_{qual}} = (n^i_1, \ldots, n^i_c)$ represents the number of occurrences of each of these values in the decision table D_i.

Then an information system is defined $S = (U, A)$, where U is a set of local decision tables and A is a set of conditional attributes (qualitative and quantitative) occurring in local tables. For the quantitative attribute $a_{quan} \in A$ a function $a_{quan} : U \to \{-1, 0, 1\}$ is defined

$$a_{quan}(D_i) = \begin{cases} 1 & \text{if } \overline{Val}_{a_{quan}} + SD_{a_{quan}} < \overline{Val}^i_{a_{quan}} \\ 0 & \text{if } \overline{Val}_{a_{quan}} - SD_{a_{quan}} \leq \overline{Val}^i_{a_{quan}} \leq \overline{Val}_{a_{quan}} + SD_{a_{quan}} \\ -1 & \text{if } \overline{Val}^i_{a_{quan}} < \overline{Val}_{a_{quan}} - SD_{a_{quan}} \end{cases} \tag{1}$$

For the quantitative attribute a_{quan} and tables, which have lower average values than typical, we assign -1. For tables with higher average values than typical, we assign 1; and for tables with typical values, we assign 0. Whereas, for the qualitative attribute a_{qual} we use the 3−means clustering algorithm for vectors

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

$Val^i_{a_{qual}}, i \in \{1, \ldots, n\}$. This is done in order to define three groups of tables with similar distribution of the attribute's a_{qual} values. Then for the attribute a_{qual} and the tables in the first group are assigned 1, in the second group 0, in the third group -1.

After defining the information system that determines the conflict situation, the conflict intensity between pairs of tables are calculated using the function $\rho(D_i, D_j) = \frac{card\{a \in A: a(D_i) \neq a(D_j)\}}{card\{A\}}$. Then coalitions are designated, a coalition is a set of tables that for every two tables D_i, D_j, $\rho(D_i, D_j) < 0.5$ is satisfied. An aggregated decision table is defined for each coalition. This is done by summing objects from local tables in the coalition. Based on the aggregated table the classification and regression tree algorithm is used with Gini index. In this way we obtain k models $M_1, \ldots, M_k$, where k is the number of coalitions. The final result $\hat{d}(x)$ is the set of decisions that were most frequently indicated by models $M_1, \ldots, M_k$. This means that there may be a tie, we do not resolve it in any way. In the experimental part the relevant measures for evaluating the quality of classification, which takes into account the possibility of occurring draws, were used. The results obtained using the proposed method are compared with the results generated by an approach without any conflict analysis. In the baseline approach, based on each local table the classification and regression tree algorithm is used. The final result is the set of decisions that were most frequently indicated by trees. Ties can arise, but analogously as before, we do not resolve them in any way.

4 Data sets, Results and Discussion

The experiments were carried out on the data available in the UC Irvine Machine Learning Repository [1]. Three data sets were selected for the analysis – the Vehicle Silhouettes, the Landsat Satellite and the Soybean (Large) data sets. In the case of Landsat Satellite and Soybean data sets, training and test sets are in the repository. The Vehicle data set was randomly split into two disjoint subsets, the training set (70% of objects) and the test set (30% of objects). The Vehicle Silhouettes data set has eighteen quantitative conditional attributes, four decision classes and 846 objects – 592 training, 254 test set. The Landsat Satellite data set has thirty-six quantitative conditional attributes, six decision classes and 6435 objects – 4435 training, 1000 test set. The Soybean data set has thirty-five quantitative conditional attributes, nineteen decision classes and 683 objects – 307 training, 376 test set. The training sets of the above data sets were dispersed. Only objects are dispersed, whereas the full set of conditional attributes is included in each local table. We use a stratified mode for dispersion. Five different dispersed versions with 3, 5, 7, 9 and 11 local tables were prepared to check for different degrees of dispersion for each data set.

The quality of classification was evaluated based on the test set. Three measures were used. The classification accuracy is the ratio of correctly classified objects from the test set to their total number in this set. When the correct decision class of an object is contained in the generated decision set, the object

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

is considered to be correctly classified. The classification ambiguity accuracy is also the ratio of correctly classified objects from the test set to their total number in this set. With the difference that this time when only one correct decision class was generated, the object is considered to be correctly classified. The third measure allows to assess the frequency and number of draws generated by the classification model. A very important aspect discussed in the paper is the rules' quality. Rules are generated based on decision trees that are built by the model. The quality of decision rules was evaluated based on the test set using the following measures. The rule confidence is the ratio of objects from the test set that matching the rule's conditions and its decision to the number of objects that satisfy the rule's conditions. The confidence is a measure of the strength of the relation between conditions and decision of rule. This is a very important measure in the context of classification because it indicates the quality of the knowledge represented by the rule – how strongly is it justified to make a certain decision based on given conditions. The rule support is the ratio of objects from the test set that matching the rule's conditions and its decision to their total number in this set. The support is a measure of the frequency of a rule. Support proves the popularity of rules. But common rules do not always constitute new and relevant knowledge. That is why confidence becomes so important. Another rules' related measure analyzed was the length of the rule indicated by the number of conditions occurring in the rule and the total number of rules generated by the model. The experiments were carried out according to the following scheme. For both the proposed and the baseline methods for five degrees of dispersion (3, 5, 7, 9, 11 local tables) different stop criterion were analyzed. The initial stop value was 2, and the step was 5. For smaller step values, non-significant differences in results were noted. The following stop values were tested: 2, 7, 12. The classification quality was evaluated using decision trees. Then, rules were generated based on decision trees and the rules' quality was estimated. For the Landsat Satellite and the Soybean data sets with 3 local tables, the proposed model did not generate coalitions, so the results obtained using the model are the same as for the baseline model. These results were omitted in the rest of the paper.

4.1 Classification quality

Table 1 shows the values of classification quality measures obtained for the proposed model and the baseline model. The higher value of classification accuracy is shown in bold. As can be seen, in the vast majority of cases, the proposed model generates better results. Improvements in classification accuracy were obtained regardless of the degree of dispersion, the used stop criterion or the analyzed data set. Statistical test was performed in order to confirm the importance in the differences in the obtained results *acc*. The received classification accuracy values were divided into two dependent data samples, each consisting of 39 observations. It was confirmed by the Wilcoxon test that the difference between the classification accuracy for both groups is significant with the level $p = 0.0001$.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

Table 1. Classification accuracy acc, classification ambiguity accuracy acc_{ONE} and the average number of generated decisions set $\bar{d}$ for the proposed method with coalitions and the baseline approach without coalitions. SC – Stop criteria, T – No. of tables

T	SC	Proposed method $acc/acc_{ONE}/\bar{d}$	Baseline method $acc/acc_{ONE}/\bar{d}$	Proposed method $acc/acc_{ONE}/\bar{d}$	Baseline method $acc/acc_{ONE}/\bar{d}$
		Landsat Satellite		Soybean	
5	2	**0.890**/0.816/1.104	0.875/0.838/1.049	**0.889**/0.780/1.223	0.865/0.777/1.152
	7	**0.891**/0.821/1.103	0.877/0.842/1.046	**0.814**/0.682/1.274	0.764/0.699/1.125
	12	**0.885**/0.812/1.094	0.862/0.831/1.045	**0.733**/0.615/1.291	0.672/0.601/1.145
7	2	**0.896**/0.824/0.824	0.877/0.844/1.038	**0.902**/0.814/1.111	0.814/0.726/1.139
	7	**0.880**/0.813/1.090	0.867/0.833/1.043	**0.858**/0.807/1.105	0.659/0.601/1.122
	12	**0.885**/0.808/1.097	0.865/0.834/1.043	**0.804**/0.699/1.115	0.601/0.574/1.115
9	2	**0.885**/0.847/1.055	0.875/0.853/1.032	**0.905**/0.851/1.064	0.807/0.713/1.145
	7	0.868/0.832/1.052	**0.872**/0.842/1.041	**0.861**/0.828/1.044	0.635/0.584/1.135
	12	**0.876**/0.831/1.064	0.867/0.840/1.039	**0.774**/0.747/1.034	0.601/0.557/1.118
11	2	**0.895**/0.861/1.040	0.872/0.851/1.031	**0.868**/0.841/1.057	0.791/0.740/1.074
	7	**0.887**/0.856/1.036	0.870/0.848/1.029	**0.872**/0.845/1.044	0.615/0.571/1.122
	12	**0.884**/0.853/1.037	0.867/0.847/1.025	**0.801**/0.794/1.010	0.341/0.304/1.071
		Vehicle Silhouettes			
3	2	**0.819**/0.516/1.382	0.780/0.665/1.236		
	7	**0.839**/0.488/1.417	0.795/0.673/1.252		
	12	**0.827**/0.480/1.413	0.776/0.657/1.236		
5	2	**0.768**/0.701/1.142	0.756/0.669/1.098		
	7	**0.760**/0.673/1.197	0.736/0.634/1.110		
	12	**0.732**/0.642/1.189	**0.732**/0.622/1.110		
7	2	0.776/0.697/1.181	**0.783**/0.705/1.114		
	7	**0.772**/0.638/1.268	0.748/0.677/1.118		
	12	**0.780**/0.657/1.252	0.736/0.661/1.114		
9	2	0.740/0.689/1.067	**0.752**/0.669/1.118		
	7	0.720/0.701/1.047	**0.732**/0.685/1.063		
	12	0.717/0.677/1.055	**0.732**/0.665/1.114		
11	2	**0.791**/0.744/1.087	0.717/0.665/1.083		
	7	**0.783**/0.732/1.063	0.736/0.677/1.071		
	12	**0.756**/0.720/1.047	0.713/0.650/1.075		

4.2 Rules' quality

Tables 2, 3 and 4 show the values of measures determining the rules' quality obtained for the data for the proposed and baseline approach. For confidence, support and length, the minimum and the maximum values obtained for the generated rules, as well as the average of values obtained for all rules with the standard deviation, are given. Results are given for different degrees of dispersion (number of tables) and different values of the stop criteria. For each degree of dispersion, the best average value (highest for confidence and support, lowest for length) obtained and the smallest number of rules generated are shown in bold. As can be seen, for lower values of the stop criterion, we get rules with higher confidence but smaller support. This can be concluded that rules

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

generated based on more expanded trees, better justify the connection between conditions and decision. The next conclusion is quite natural, for larger values of the stop criterion (by limiting the trees' growth) we get shorter rules and their number is smaller.

Table 2. Confidence, support and length of the rules generated by the proposed and the baseline method (Vehicle Silhouettes). SC – Stop criteria, T – No. of tables

T	SC	Rules' confidence Min/Max/AVG/SD	Rules' support Min/Max/AVG/SD	Rules' length Min/Max/AVG/SD	No. rules
Proposed method					
3	2	0.200/1/**0.660**/0.273	0.004/0.173/0.021/0.033	3/12/6.938/2.358	64
	7	0.111/1/0.625/0.258	0.004/0.177/0.023/0.035	3/11/6.525/2.227	59
	12	0.200/1/0.639/0.241	0.004/0.173/**0.025**/0.035	3/11/**6.173**/2.101	**52**
5	2	0.143/1/**0.655**/0.281	0.004/0.185/0.020/0.033	3/14/6.907/2.529	97
	7	0.111/1/0.605/0.268	0.004/0.185/0.024/0.035	3/14/6.578/2.475	83
	12	0.111/1/0.593/0.270	0.004/0.185/**0.027**/0.037	3/12/**5.986**/2.223	**71**
7	2	0.125/1/**0.644**/0.270	0.004/0.161/0.026/0.037	1/13/6.781/2.587	73
	7	0.091/1/0.584/0.276	0.004/0.161/0.028/0.038	1/12/6.212/2.390	66
	12	0.091/1/0.596/0.263	0.004/0.154/**0.034**/0.040	1/11/**5.804**/2.271	**56**
9	2	0.111/1/**0.649**/0.276	0.004/0.177/0.022/0.034	2/11/6.333/1.868	150
	7	0.100/1/0.605/0.288	0.004/0.177/0.024/0.036	2/10/5.852/1.774	135
	12	0.091/1/0.589/0.285	0.004/0.177/**0.027**/0.036	2/10/**5.593**/1.714	**118**
11	2	0.071/1/**0.630**/0.273	0.004/0.193/0.027/0.038	2/9/5.878/1.756	164
	7	0.091/1/0.611/0.263	0.004/0.193/0.033/0.040	2/9/5.348/1.636	138
	12	0.091/1/0.592/0.260	0.004/0.193/**0.037**/0.043	1/8/**4.992**/1.497	**120**
Baseline method					
3	2	0.111/1/**0.673**/0.273	0.004/0.173/0.027/0.037	2/12/6.167/1.979	72
	7	0.167/1/0.641/0.269	0.004/0.177/0.03/0.038	2/12/5.848/2.091	66
	12	0.167/1/0.605/0.261	0.004/0.177/**0.032**/0.039	2/10/**5.441**/1.844	**59**
5	2	0.111/1/**0.633**/0.633	0.004/0.185/0.030/0.038	2/9/5.388/1.470	103
	7	0.067/1/0.58/0.268	0.004/0.185/0.035/0.042	2/8/4.874/1.413	87
	12	0.067/1/0.581/0.271	0.004/0.185/**0.044**/0.044	2/6/**4.29**/1.105	**69**
7	2	0.071/1/**0.582**/0.271	0.004/0.181/0.033/0.040	1/10/5.165/1.697	127
	7	0.091/1/0.574/0.259	0.004/0.181/0.04/0.043	1/8/4.608/1.509	102
	12	0.111/1/0.566/0.252	0.004/0.181/**0.052**/0.045	1/7/**4**/1.271	**78**
9	2	0.050/1/**0.553**/0.263	0.004/0.217/0.038/0.044	2/9/4.674/1.449	132
	7	0.05/1/0.537/0.273	0.004/0.232/0.048/0.047	2/8/4.151/1.257	106
	12	0.048/1/0.551/0.264	0.004/0.217/**0.061**/0.050	1/7/**3.675**/1.163	**83**
11	2	0.036/1/0.523/0.252	0.004/0.181/0.041/0.046	2/8/4.437/1.369	151
	7	0.036/1/0.518/0.243	0.004/0.185/0.052/0.049	2/8/3.866/1.173	119
	12	0.095/1/**0.525**/0.242	0.004/0.185/**0.067**/0.055	1/6/**3.44**/1.061	**91**

In order to confirm the importance in the differences of the obtained measures (confidence, support and length) in relation to different stop criterion, statistical tests were performed. The results were grouped depending on the stop criterion, three dependent samples, were created. The Friedman test was used. There

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

Table 3. Confidence, support and length of the rules generated by the proposed and the baseline method (Landsat Satellite). SC – Stop criteria, T – No. of tables

T	SC	Rules' confidence Min/Max/AVG/SD	Rules' support Min/Max/AVG/SD	Rules' length Min/Max/AVG/SD	No. rules
Proposed method					
5	2	0.043/1/**0.646**/0.288	0.001/0.187/0.010/0.028	3/18/8.645/2.763	313
	7	0.043/1/0.637/0.283	0.001/0.188/0.012/0.030	3/17/8.209/2.838	268
	12	0.043/1/0.623/0.281	0.001/0.187/**0.014**/0.032	3/17/**7.927**/2.899	**233**
7	2	0.050/1/**0.665**/0.280	0.001/0.186/0.010/0.026	3/19/8.907/2.891	332
	7	0.050/1/0.634/0.276	0.001/0.186/0.011/0.027	3/19/8.524/2.971	290
	12	0.037/1/0.626/0.276	0.001/0.186/**0.013**/0.029	3/19/**8.313**/3.060	**252**
9	2	0.067/1/**0.663**/0.272	0.001/0.185/0.011/0.029	3/19/8.959/3.361	365
	7	0.053/1/0.654/0.270	0.001/0.184/0.013/0.031	2/19/8.477/3.314	302
	12	0.053/1/0.634/0.266	0.001/0.184/**0.015**/0.032	2/19/**8.348**/3.410	**273**
11	2	0.045/1/**0.656**/0.276	0.001/0.196/0.010/0.027	3/17/8.728/2.752	643
	7	0.048/1/0.630/0.277	0.001/0.195/0.012/0.029	3/17/8.336/2.761	560
	12	0.048/1/0.632/0.270	0.001/0.195/**0.013**/0.031	2/17/**8.109**/2.814	**494**
Baseline method					
5	2	0.083/1/**0.633**/0.270	0.001/0.187/0.012/0.030	3/15/8.161/2.410	348
	7	0.056/1/0.597/0.277	0.001/0.188/0.013/0.032	2/14/7.679/2.312	296
	12	0.056/1/0.603/0.277	0.001/0.187/**0.016**/0.034	2/14/**7.325**/2.411	**252**
7	2	0.063/1/**0.608**/0.279	0.001/0.187/0.014/0.033	3/15/7.731/2.383	401
	7	0.063/1/0.604/0.266	0.001/0.187/0.018/0.036	2/14/7.201/2.396	318
	12	0.080/1/0.606/0.261	0.001/0.187/**0.020**/0.038	2/14/**7.080**/2.441	**275**
9	2	0.045/1/**0.599**/0.286	0.001/0.196/0.017/0.038	3/14/7.005/2.144	419
	7	0.040/1/0.583/0.280	0.001/0.195/0.022/0.042	2/14/6.571/2.188	324
	12	0.045/1/0.575/0.281	0.001/0.194/**0.025**/0.045	2/13/**6.279**/2.177	**276**
11	2	0.053/1/**0.594**/0.280	0.001/0.199/0.020/0.041	3/14/6.668/1.985	440
	7	0.043/1/0.573/0.286	0.001/0.199/0.024/0.044	2/13/6.184/1.926	358
	12	0.043/1/0.581/0.276	0.001/0.199/**0.029**/0.047	2/13/**5.950**/2.004	**300**

were statistically significant differences in the results of confidence, support and length obtained for different stop criterion being considered. The following results were obtained: for confidence $\chi^2(26,2) = 36.538, p = 0.000001$; for support $\chi^2(26,2) = 46.231, p = 0.000001$; for length $\chi^2(26,2) = 52, p = 0.000001$. Additionally, we confirmed (by using the Wilcoxon test) that the differences in confidence, support and length are significant between each-pair of stop criterion values. Comparative box-whiskers charts for the results with three values of stop criterion were created (Fig. 1). Based on the charts, we can see that the greatest differences in results divided into stop criterion were obtained for support and length (here the boxes are located farthest from each other). Thus, using a larger stop criterion reduces the length of the rules and increases support to a greater extent than it reduces confidence.

Table 5 presents a comparison of the average values of rules' confidence, support and length generated by the proposed and baseline method. The higher values of confidence, support and lower values of rules' length are shown in bold.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

Table 4. Confidence, support and length of the rules generated by the proposed and the baseline method (Soybean). SC – Stop criteria, T – No. of tables

T	SC	Rules' confidence Min/Max/AVG/SD	Rules' support Min/Max/AVG/SD	Rules' length Min/Max/AVG/SD	No. rules
Proposed method					
5	2	0.091/1/**0.734**/0.282	0.003/0.142/0.032/0.034	3/12/5.337/1.872	89
	7	0.100/1/0.668/0.286	0.003/0.142/0.041/0.037	2/9/4.603/1.453	63
	12	0.100/1/0.621/0.286	0.003/0.169/**0.053**/0.043	2/7/**3.886**/1.112	**44**
7	2	0.074/1/**0.798**/0.284	0.003/0.142/0.033/0.035	2/12/5.521/1.785	94
	7	0.125/1/0.749/0.287	0.003/0.142/0.041/0.038	2/11/4.944/1.786	72
	12	0.093/1/0.692/0.277	0.003/0.145/**0.050**/0.044	2/9/**4.364**/1.577	**55**
9	2	0.125/1/**0.758**/0.302	0.003/0.142/0.033/0.036	3/10/5.457/1.724	138
	7	0.075/1/0.733/0.285	0.003/0.142/0.040/0.036	2/9/4.850/1.440	113
	12	0.083/1/0.627/0.277	0.003/0.149/**0.049**/0.040	2/8/**4.300**/1.364	**80**
11	2	0.077/1/**0.770**/0.286	0.003/0.132/0.031/0.036	2/12/5.917/1.950	132
	7	0.071/1/0.739/0.291	0.003/0.155/0.038/0.040	2/11/5.272/1.855	103
	12	0.083/1/0.747/0.278	0.003/0.172/**0.047**/0.043	1/9/**4.910**/1.834	**78**
Baseline method					
5	2	0.071/1/**0.723**/0.297	0.003/0.142/0.034/0.038	2/10/5.225/1.650	102
	7	0.083/1/0.640/0.281	0.003/0.142/0.045/0.041	2/7/4.362/1.285	69
	12	0.100/1/0.557/0.297	0.003/0.169/**0.062**/0.049	2/6/**3.444**/0.858	**45**
7	2	0.056/1/**0.685**/0.281	0.003/0.132/0.044/0.031	2/9/5.282/1.657	103
	7	0.016/1/0.576/0.286	0.003/0.135/0.046/0.038	1/8/4.067/1.352	60
	12	0.014/1/0.499/0.305	0.007/0.169/**0.070**/0.046	1/5/**2.824**/0.785	**34**
9	2	0.016/1/**0.584**/0.330	0.003/0.169/0.033/0.035	1/10/4.817/1.798	109
	7	0.056/1/0.514/0.279	0.003/0.169/0.057/0.045	2/6/3.525/1.168	61
	12	0.016/1/0.416/0.241	0.010/0.172/**0.078**/0.055	1/5/**2.611**/0.980	**36**
11	2	0.006/1/**0.566**/0.308	0.003/0.149/0.034/0.036	1/8/4.552/1.445	125
	7	0.014/1/0.492/0.292	0.007/0.169/0.063/0.051	1/6/3.034/0.920	59
	12	0.015/0.9/0.352/0.240	0.007/0.172/**0.083**/0.061	1/3/**1.970**/0.577	**33**

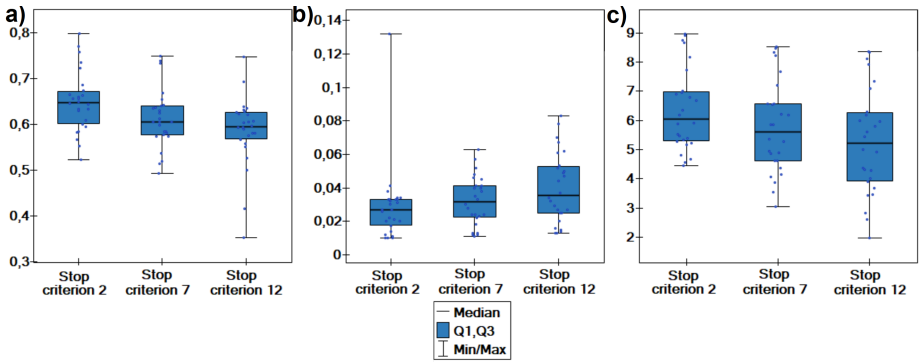

Fig. 1. Comparison of (a) the average rules' confidence (b) the average rules' support (c) the average rules' length obtained for different stop criterion.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

As can be seen in the vast majority of cases better confidence was obtained for the proposed method. The creation of coalitions and aggregated tables for the coalitions made it possible to generate rules that are better representation of knowledge hidden in data. The confidence is the most important measure, it shows how much the rule's conditions actually indicates the decision. Higher values of support were noted for the baseline approach. However, this measure only shows the fraction of objects supporting the rule's antecedent and decision – it does not indicate the actual connection between conditions and decision. The baseline method produces shorter rules than the proposed method. In most cases, the average number of conditions in rules is greater by one condition for the proposed method than for the baseline method. However, this measure is also less important than confidence when evaluating the quality of generated rules. Statistical tests were performed in order to confirm the importance in the differences in the obtained results of rules' confidence, support and length. At first, the average values of rules' confidence in two dependent groups were analysed – the proposed and the baseline methods. Both groups contained 39 observations each – all results for dispersed data sets. It was confirmed by the Wilcoxon test that the difference between the averages of rules' confidence for both groups is significant with the level $p = 0.0001$. In an analogous way – using the Wilcoxon test – the statistical significance of the differences between the averages of rules' support and rules' length were confirmed, with the level $p = 0.0001$ in both cases. Additionally, comparative box-whiskers chart for the values of rules' confidence, support and length was created (Fig. 2). The biggest difference can be noticed in the case of rules' confidence – the boxes are located in different places, do not overlap in any part. The difference in rules' confidence is the most significant for us, it shows that the knowledge generated when using the proposed method is of much better quality than for the non-coalitions approach.

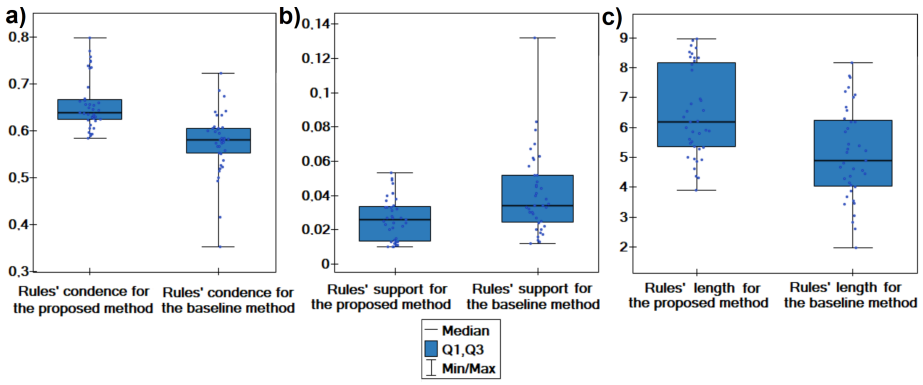

Fig. 2. Comparison of (a) the average rules' confidence (b) the average rules' support (c) the average rules' length obtained for the proposed method and the baseline method.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

Table 5. Comparison of the average confidence, support and length of rules obtained for the proposed and the baseline methods. SC – Stop criteria, T – No. of tables

T	SC	Rules' confidence		Rules' support		Rules' length	
		Proposed	Baseline	Proposed	Baseline	Proposed	Baseline
Vehicle Silhouettes							
3	2	0.660	**0.673**	0.021	**0.027**	6.938	**6.167**
	7	0.625	**0.641**	0.023	**0.03**	6.525	**5.848**
	12	**0.639**	0.605	0.025	**0.032**	6.173	**5.441**
5	2	**0.655**	0.633	0.020	**0.030**	6.907	**5.388**
	7	**0.605**	0.58	0.024	**0.035**	6.578	**4.874**
	12	**0.593**	0.581	0.027	**0.044**	5.986	**4.290**
7	2	**0.644**	0.582	0.026	**0.033**	6.781	**5.165**
	7	**0.584**	0.574	0.028	**0.04**	6.212	**4.608**
	12	**0.596**	0.566	0.034	**0.052**	5.804	**4**
9	2	**0.649**	0.553	0.022	**0.038**	6.333	**4.674**
	7	**0.605**	0.537	0.024	**0.048**	5.852	**4.151**
	12	**0.589**	0.551	0.027	**0.061**	5.593	**3.675**
11	2	**0.630**	0.523	0.027	**0.041**	5.878	**4.437**
	7	**0.611**	0.518	0.033	**0.052**	5.348	**3.866**
	12	**0.592**	0.525	0.037	**0.067**	4.992	**3.44**
Landsat Satellite							
5	2	**0.646**	0.633	0.010	**0.012**	8.645	**8.161**
	7	**0.637**	0.597	0.012	**0.013**	8.209	**7.679**
	12	**0.623**	0.603	0.014	**0.016**	7.927	**7.325**
7	2	**0.665**	0.608	0.010	**0.014**	8.907	**7.731**
	7	**0.634**	0.604	0.011	**0.018**	8.524	**7.201**
	12	**0.626**	0.606	0.013	**0.020**	8.313	**7.080**
9	2	**0.663**	0.599	0.011	**0.017**	8.959	**7.005**
	7	**0.654**	0.583	0.013	**0.022**	8.477	**6.571**
	12	**0.634**	0.575	0.015	**0.025**	8.348	**6.279**
11	2	**0.656**	0.594	0.010	**0.020**	8.728	**6.668**
	7	**0.630**	0.573	0.012	**0.024**	8.336	**6.184**
	12	**0.632**	0.581	0.013	**0.029**	8.109	**5.950**
Soybean							
5	2	**0.734**	0.723	0.032	**0.034**	5.337	**5.225**
	7	**0.668**	0.640	0.041	**0.045**	4.603	**4.362**
	12	**0.621**	0.557	0.053	**0.062**	3.886	**3.444**
7	2	**0.798**	0.685	0.033	**0.132**	5.521	**5.282**
	7	**0.749**	0.576	0.041	**0.046**	4.944	**4.067**
	12	**0.692**	0.499	0.050	**0.070**	4.364	**2.824**
9	2	**0.758**	0.584	0.033	0.033	5.457	**4.817**
	7	**0.733**	0.514	0.040	**0.057**	4.850	**3.525**
	12	**0.627**	0.416	0.049	**0.078**	4.300	**2.611**
11	2	**0.770**	0.566	0.031	**0.034**	5.917	**4.552**
	7	**0.739**	0.492	0.038	**0.063**	5.272	**3.034**
	12	**0.747**	0.352	0.047	**0.083**	4.910	**1.970**

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

5 Conclusion

The paper presents a classification model for data stored independently in several decision tables. We assume that the sets of conditional attributes in all tables are equal. The proposed model creates coalitions of tables containing similar data – more precisely, similar attributes' values. For the coalitions aggregated tables are created. Decision trees are generated based on these tables. The study compared the proposed model with a model in which coalitions are not used. An analysis of the quality of rules generated by the model and the effect of the stop criterion on the results was also made. It was shown that the proposed model generates significantly better classification accuracy than the model without coalitions. Also, the rules generated by the proposed model have significantly higher confidence than the rules generated by the baseline model. The use of larger values of stop criterion has less effect on reducing rules' confidence, while it has greater effect on increasing support and reducing rules' length. In the future work, it is planned to consider the variation of conditional attributes' values within each decision class for generating coalitions of local tables.

References

1. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, University of Massachusetts Amherst, USA. 2007.
2. Bayardo Jr, R. J., Agrawal, R.: Mining the most interesting rules. In Proceedings of the 5th ACM SIGKDD, 145–154, 1999.
3. Czarnowski, I., Jędrzejowicz, P.: Ensemble online classifier based on the one-class base classifiers for mining data streams. Cybern. Syst., 46(1-2), 51–68, 2015.
4. Dembczyński, K., Kotłowski, W., Słowiński, R.: Ender: a statistical framework for boosting decision rules. Data Mining and Knowledge Discovery, 21(1), 52–90, 2010.
5. Freund, Y., Schapire, R. E.: A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci., 55(1), 119–139, 1997.
6. Gao, J., Lanchantin, J., Soffa, M. L., Qi, Y.: Black-box generation of adversarial text sequences to evade deep learning classifiers. In 2018 IEEE Security and Privacy Workshops (SPW), 50–56, IEEE, 2018.
7. Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA), 80–89, IEEE, 2018.
8. Gut, D., Tabor, Z., Szymkowski, M., Rozynek, M., Kucybała, I., Wojciechowski, W.: Benchmarking of deep architectures for segmentation of medical images. IEEE Transactions on Medical Imaging, 41(11), 3231–3241, 2022.
9. Kozak, J.: Decision tree and ensemble learning based on ant colony optimization. Springer International Publishing, 2019.
10. Krawczyk, B., Woźniak, M., Cyganek, B.: Clustering-based ensembles for one-class classification. Inf. Sci., 264, 182–195, 2014.
11. Kuncheva, L. I. Combining pattern classifiers: methods and algorithms. 2014. John Wiley & Sons.
12. Li, Z., Sharma, V., Mohanty, S. P.: Preserving data privacy via federated learning: Challenges and solutions. IEEE Consumer Electronics Magazine, 9(3), 8–16, 2020.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30

13. Meng, T., Jing, X., Yan, Z., Pedrycz, W.: A survey on machine learning for data fusion. Information Fusion, 57, 115–129, 2020.
14. Nam, G., Yoon, J., Lee, Y., Lee, J.: Diversity matters when learning from ensembles. Advances in Neural Information Processing Systems, 34, 8367–8377, 2021.
15. Nguyen, H. T., Sehwag, V., Hosseinalipour, S., Brinton, C. G., Chiang, M., Poor, H. V.: Fast-convergent federated learning. IEEE J. Sel. Areas Commun., 39(1), 201–218, 2020.
16. Pawlak, Z., Skowron, A.: Rough sets and Boolean reasoning. Inf. Sci., 177(1), 41–73, 2007.
17. Pawlak, Z. Conflict analysis. In Proceedings of the Fifth European Congress on Intelligent Techniques and Soft Computing (EUFIT'97) 1589–1591, 1997.
18. Pięta, P., Szmuc, T.: Applications of rough sets in big data analysis: An overview. Int. J. Appl. Math. Comput. Sci., 31(4), 659–683, 2021.
19. Połap, D., Woźniak, M.: Meta-heuristic as manager in federated learning approaches for image processing purposes. Appl. Soft Comput., 113, 107872, 2021.
20. Priyanka, Kumar, D.: Decision tree classifier: A detailed survey. Int. J. Inf. Decis. Sci., 12(3), 246–269, 2020.
21. Przybyła-Kasperek, M., Smyczek, F.: Comparison of Shapley-Shubik and Banzhaf-Coleman power indices applied to aggregation of predictions obtained based on dispersed data by k-nearest neighbors classifiers. Procedia Comput. Sci., 207, 2134–2143, 2022.
22. Przybyła-Kasperek, M, Kusztal, K.: New Classification Method for Independent Data Sources Using Pawlak Conflict Model and Decision Trees. Entropy. 2022; 24(11):1604. https://doi.org/10.3390/e24111604
23. Przybyła-Kasperek, M., Aning, S.: Stop Criterion in Building Decision Trees with Bagging Method for Dispersed Data. Procedia Comput. Sci., 192, 3560–3569, 2021.
24. Przybyła-Kasperek, M.: Coalitions' Weights in a Dispersed System with Pawlak Conflict Model. Group Decis. Negot., 29(3), 549–591, 2020.
25. Przybyła-Kasperek, M.: Three conflict methods in multiple classifiers that use dispersed knowledge. Int J Inf Technol Decis Mak., 18(02), 555–599, 2019.
26. Ren, P., Xu, Z., Kacprzyk, J.: Group Decisions with Intuitionistic Fuzzy Sets. In: Kilgour, D.M., Eden, C. (eds) Group Decis. Negot. Springer, Cham, 2021.
27. Skokowski, P., Łopatka, J., Malon, K.: Evidence Theory Based Data Fusion for Centralized Cooperative Spectrum Sensing in Mobile Ad-hoc Networks. In 2020 Baltic URSI Symposium (URSI) 24–27, IEEE, 2020.
28. Skowron, A., Deja, R.: On some conflict models and conflict resolutions. Rom. J. Inf. Sci. Technol., 3(1–2), 69–82, 2002.
29. Słowiński, R., Greco, S., Matarazzo, B.: Rough set analysis of preference-ordered data. International Conference, RSCTC 44–59. Springer, Berlin, Heidelberg, 2002.
30. Stefanowski, J.: On rough set based approaches to induction of decision rules. Rough sets in knowledge discovery, 1(1), 500–529, 1998.
31. Ślęzak, D., Wróblewski, J.: Order based genetic algorithms for the search of approximate entropy reducts. Int. Workshop, RSFDGrC, 308–311, Springer, 2003.
32. Wieczorek, A., Słowiński, R. Generating a set of association and decision rules with statistically representative support and anti-support. Inf. Sci., 277, 56–70, 2014.
33. Yao, Y.: Rough sets and three-way decisions. International Conference, RSKT, 62–73, Springer, Cham, 2015.
34. Zou, X., Zhong, S., Yan, L., Zhao, X., Zhou, J., Wu, Y.: Learning robust facial landmark detection via hierarchical structured ensemble. Proc. IEEE Int. Conf. Comput. Vis. 141–150, 2019.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_30