Coalition-Based Rule Induction and Decision Template Matching for Distributed Tabular Data

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

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

Abstract

This paper presents a novel approach to the classification of distributed data, which integrates the cooperation of local decision tables within coalitions with rule induction and decision templates. The method aims to preserve model transparency while taking into account the diversity of data sources. Experiments were conducted on three datasets, comparing the performance of four rule induction algorithms: exhaustive search algorithm, genetic algorithm, covering algorithm, and LEM2. The best classification results were obtained for the exhaustive and genetic algorithms, while the covering and LEM2 methods performed significantly worse. The proposed approach achieves results comparable to the baseline method, which does not incorporate the coalition mechanism, while offering higher interpretability. In addition, the proposed solution was compared with the Authors' earlier approaches based on decision tree classifiers.

Keywords: Distributed data classification, Decision rules, Coalitions, Interpretability, Rule induction algorithms.

1. Introduction

In modern data-driven systems, the utilization of distributed/dispersed tabular data – where information is fragmented across independently managed local tables – presents both opportunities and challenges. While decentralized storage enables scalability and domain-specific customization, it introduces critical data integrity risks resulting from potential contradictions or inconsistencies between local datasets. Inconsistencies in attribute distributions or measurement protocols can create conflicting representations of reality, destroying the reliability of analyses. At the same time, the interpretability of machine learning models is crucial – particularly in fields such as business, medicine, and other high-risk domains, where understanding and trusting model decisions is essential. In this paper, the literature review is integrated into the introduction section to ensure narrative cohesion and to better contextualize the research problem within existing approaches.

Machine learning techniques for classification can be broadly divided into interpretable models, such as rule-based classifiers and decision trees [8], and black-box models like deep learning and neural networks, which, while highly accurate, often lack transparency and explainability [9]. The challenge of classifying dispersed data has been addressed in the literature through various strategies. Ensemble learning approaches, for example, train multiple classifiers on separate data partitions and combine their outputs using methods such as voting or weighted averaging, as seen in bagging, boosting, and stacking [19]. However, these strategies typically focus on boosting accuracy rather than integrating knowledge from disparate sources. Federated

learning has emerged as an alternative, enabling decentralized model training without direct data exchange to preserve privacy, but at the cost of interpretability [13]. Other studies have explored hierarchical classification frameworks that aggregate local models at different levels, though these often lack explicit mechanisms for resolving conflicts between local predictions.

Pawlak's conflict analysis model [12] has been influential in addressing inconsistencies in decision-making, particularly within rough set-based learning and three-way decision theory [22]. Data fusion techniques have also been proposed, using statistical and mathematical measures to reconcile inconsistencies in dispersed datasets [23], but these methods rarely produce human-readable decision rules. More recent research in distributed learning has highlighted the importance of considering dependencies and potential conflicts between local models, as ignoring these relationships can undermine the quality and reliability of the final prediction. Rough set theory, as introduced by Pawlak [11], provides a theoretical foundation for handling uncertainty and conflict in classification by partitioning data into lower and upper approximations, and has since been extended by numerous researchers [6], [16, 17], including the development of three-way decision theory [20, 21].

This paper introduces a novel framework that addresses both challenges through coalition-based analysis, interpretable rule induction and decision templates representing patterns in predictive vectors generated from rules. By grouping local tables into coalitions with aligned statistical profiles, we mitigate data integrity risks while preserving meaningful variations between subgroups. Each coalition generates transparent decision rules with using rough set theory. These rule-based outputs are then synthesized into decision templates – patterns that capture coalition-specific decision strategies while enabling robust final classification through template matching.

The main novelty of this paper lies in the integration of decision templates with coalition-based classification for distributed tabular data. While the concept of forming coalitions from local tables has been explored in previous papers [14, 15], this is the first time that decision templates – constructed from rule-based predictions – are used in conjunction with coalitions to provide both classification accuracy and interpretability.

The remainder of this paper is structured as follows. Section 2 details the proposed framework for classifying dispersed data, coalition formation, the process of rule induction and the decision templates method. Section 3 outlines the dataset, experimental procedures, and results, with a focus on evaluating the performance of various rule induction algorithms. Section 4 concludes the paper and discusses directions for future research.

2. Methods and models

In this study, we focus on a classification method that works with data stored by independent units. Our goal is to use easy-to-interpret decision rules. The method has a layered structure, where similar data from local tables are grouped together. When data in these groups are consistent, we can create reliable decision rules. Predictions for training objects obtained based on decision rules derived from groups of local tables are used to construct decision templates for individual decision classes. These templates serve as reference patterns, capturing subtle distinctions that reflect the unique capabilities, specializations of each group of local tables. The final decision is determined by assessing the similarity between the predictions for the test object generated by the rules and the corresponding decision class templates. Four main steps in the proposed method are as follows:

1. Grouping local tables into coalitions – local tables containing compatible data,

2. Aggregating data and inducing local decision rules for coalitions,

3. Generating prediction vectors for training objects and creating decision template for each

decision class,

4. Final classification, prediction vectors for the test object are compared to the decision templates using normalized Euclidean distance.

The idea of joining local tables into coalitions has been used before in the papers [14, 15]. However, this approach has not been combined with decision templates and decision rules generated based on rough set theory until now. Below, we give a formal explanation of how this approach works for dispersed data.

We use dispersed tabular data. All these tables have the same set of condition attributes. Let's define each local table as $D_i = (U_i, A, d)$ for $i \in \{1, \ldots, n\}$. Here, U_i is the set of objects in the table, A is the set of condition attributes, and d is the decision attribute. These local tables are managed independently and may be stored in different locations. Some of them may also contain conflicting or inconsistent data.

To group local tables into coalitions, we compare the statistical properties of the data and use Pawlak's conflict analysis model [12]. In this model, an information system is described as $S = (LD, A)$, where LD is the set of all local decision tables: $LD = \{D_i : i \in \{1, \ldots, n\}\}$ and A is the set of attributes occurring in these tables.

In this system, we describe the significance of each attribute using three values: $\{-1, 0, 1\}$. For each attribute $a \in A$, we define a function $a : LD \to \{-1, 0, 1\}$, which assigns a value to each local table. These values mean the following: $a(D_i) = 0$ means that the values of attribute a in table D_i are typical compared to all other tables. $a(D_i) = 1$ means the values are higher than usual. $a(D_i) = -1$ means the values are lower than usual. The way we assign these values depends on the type of attribute – whether it's qualitative (like colors or categories) or quantitative (like numbers). This allows us to fairly compare different kinds of data across all the local tables.

To assess quantitative attributes $a_{quan} \in A$, we begin by determining the average value of each attribute within every local decision table D_i, denoted as $\overline{Val}^i_{a_{quan}}$. We then compute the overall average and standard deviation of that attribute across all tables in the system, labeled $\overline{Val}_{a_{quan}}$ and $SD_{a_{quan}}$, respectively. Based on these global statistics, we define a classification function $a_{quan} : LD \to \{-1, 0, 1\}$, which evaluates whether a local table's attribute average deviates significantly from the norm:

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

This categorization allows us to capture variations in quantitative attributes by assigning each table to one of three significance levels.

In the case of qualitative attributes $a_{qual} \in A$, we take a different route. For each local table D_i, we count the frequency of every possible value the attribute can assume. If the attribute has c possible values, we construct a frequency vector $Val^i_{a_{qual}} = (n^i_1, \ldots, n^i_c)$, where n^i_j denotes the count of value val_j in table D_i. To group local tables by similarity in value distributions, we apply a 3-means clustering algorithm on the set of frequency vectors using Euclidean distance. Each cluster represents a different pattern of attribute distribution. Based on the assigned cluster, we label each table:

- $a_{qual}(D_i) = 1$ if D_i belongs to the first cluster,
- $a_{qual}(D_i) = 0$ if D_i belongs to the second cluster,
- $a_{qual}(D_i) = -1$ if D_i belongs to the third cluster.

After each local table has been encoded using these trinary values for all attributes, we define a pairwise conflict function $\rho : LD \times LD \to [0, 1]$. This function quantifies the disagreement between two local tables by comparing their categorical representations:

$$\rho(D_i, D_j) = \frac{card\{a \in A : a(D_i) \neq a(D_j)\}}{card\{A\}}. \tag{2}$$

A higher value of $\rho(D_i, D_j)$ reflects a stronger divergence between the attribute profiles of the two tables. Using this measure, we form coalitions by grouping tables that show a high degree of alignment. Specifically, two tables may be part of the same coalition if $\rho(D_i, D_j) < 0.5$, meaning that they agree on at least half of the attributes.

Each coalition is then represented by an aggregated decision table that consolidates the data of its members. For coalition j, this table is expressed as: $D_j^{aggr} = (U_j^{aggr}, A, d)$ is the union of all objects from the local tables within the coalition. The conditional attribute set A and decision attribute d remain the same as in the original tables. For each object $x \in U_i$, its attribute values in the aggregated table are directly acquired from the source table D_i.

We assume that, within each coalition, the raw datasets are shared to aggregation. This introduces potential privacy concerns, but it also enables to derive coherent decision-rule representations and to construct decision templates in the second step that capture the distinctive specialization of each coalition of decision tables.

After constructing the aggregated decision tables, we perform rule induction on each table individually to extract its local decision rules. While brute-force techniques exhaustively enumerate all possible rules, they quickly become impractical as the number of attributes grows. To address this, the literature offers a variety of heuristic and metaheuristic methods-ranging from ant colony optimization and approximation schemes to other algorithms that reduce computational cost. In our experiments, we employ four rough set–based induction strategies:

- Exhaustive search algorithm (Exh) [3]
- Genetic algorithm (Gen) [1]
- Covering algorithm (Cov) [2]
- LEM2 algorithm (LEM2) [7]

For each coalition, the rule set induced from its aggregated decision table is applied to generate prediction vector for objects. Measurement level vectors are generated for tests and training objects in the same way. For the object, we first identify all rules that cover the object and assign each a weight equal to rule's coverage frequency (i.e., the number of matching objects from the training set) divided by the total number of rules in the corresponding local rule set for coalition. We then sum these weights within each decision category. The prediction vector is normalized. Thus, for each $j-$th coalition and object x, a prediction vector is created $[\mu_{j,1}(x), \cdots, \mu_{j,i}(x), \cdots, \mu_{j,c}(x)]$, where c is the number of decision classes.

In the last stage of the proposed model, a method based on decision templates is used [10]. In the decision template approach, we first build a template for each decision class by combining the rule-based predictions made on our training data. Then, for a new object, we compare its prediction results to these templates and pick the class whose template it most closely matches.

During the decision template method training phase, we aggregate rule-based outputs to form a prototype prediction pattern for each decision class. The decision template DT_i for class i is the average of the prediction vectors of the objects of the training set labelled in class i, thus

$$DT_i = \frac{1}{card\{X_i\}} \sum_{x \in X_i} \begin{bmatrix} \mu_{1,1}(x) & \cdots & \mu_{1,i}(x) & \cdots & \mu_{1,c}(x) \\ & & \cdots & & \\ \mu_{j,1}(x) & \cdots & \mu_{j,i}(x) & \cdots & \mu_{j,c}(x) \\ & & \cdots & & \\ \mu_{L,1}(x) & \cdots & \mu_{L,i}(x) & \cdots & \mu_{L,c}(x) \end{bmatrix}, \tag{3}$$

where X_i is the set of objects from the training set that belongs to the class i and L is the number of coalitions. When classifying objects $\bar{x}$, we generate their prediction vectors and store in a the decision profile

$$DP(\bar{x}) = \begin{bmatrix} \mu_{1,1}(\bar{x}) & \cdots & \mu_{1,i}(\bar{x}) & \cdots & \mu_{1,c}(\bar{x}) \\ & & \cdots & & \\ \mu_{j,1}(\bar{x}) & \cdots & \mu_{j,i}(\bar{x}) & \cdots & \mu_{j,c}(\bar{x}) \\ & & \cdots & & \\ \mu_{L,1}(\bar{x}) & \cdots & \mu_{L,i}(\bar{x}) & \cdots & \mu_{L,c}(\bar{x}) \end{bmatrix} \tag{4}$$

Then we match the decision profile against decision templates DT_i to identify the most similar decision class. To calculate the distance between the decision profile and the decision templates, we use the normalised Euclidean distance

$$s(DP(\bar{x}), DT_i) = \frac{1}{L \cdot c} \sum_{m=1}^{L} \sum_{l=1}^{c} \left(DP^{m,l}(\bar{x}) - DT_i^{m,l}\right)^2, \tag{5}$$

where $DP^{m,l}(\bar{x})$ and $DT_i^{m,l}$ is an element at the m-th row and the l-th column of the matrix $DP(\bar{x})$ or DT_i respectively.

3. Experiments

To experimentally evaluate the proposed approach, three datasets from the UCI Machine Learning Repository [5] were used: Vehicle Silhouettes [18], Car Evaluation [4], and Balance Scale. Each dataset was partitioned into two disjoint subsets through a stratified sampling procedure: a training set (70% of the instances) and a test set (the remaining 30%). The characteristics of the datasets are presented in Table 1.

Table 1. Dataset characteristics

Dataset	# Training set	# Test set	# Conditional attributes	Attributes type	# Decision classes
Vehicle Silhouettes	592	254	18	Integer	4
Car Evaluation	1209	519	6	Categorical	4
Balance Scale	437	188	4	Categorical	3

Next, to reflect environments in which data are collected and maintained by multiple independent sources, each training set was partitioned into several levels of data dispersion, comprising 5, 7, 9, and 11 local tables. This partitioning was performed in a stratified manner. Each local table retains the full set of attributes while containing only a subset of instances from the original training set. In total, 12 dispersed versions of the datasets were prepared.

The quality of classification was assessed based on the test set. To enable a thorough comparison of the results, several complementary performance metrics were applied: classification accuracy (Acc), balanced accuracy (BAcc), precision (Prec.), recall, F-measure (F-m.), and geometric mean (G-mean). Accuracy refers to the proportion of correctly classified instances. Precision indicates the percentage of predictions assigned to a given class that were correct, whereas recall measures the percentage of instances belonging to a given class that were correctly identified. F-measure is the harmonic mean of precision and recall, balancing these two aspects of evaluation:

$$\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{6}$$

In turn, balanced accuracy and G-mean are metrics that account for class imbalance – the former is based on the average recall across all classes, while the latter evaluates the overall balance of classification performance, emphasizing high recall in each class.

The experimental procedure consisted of the following steps:

- Formation of coalitions of local decision tables.
- Induction of decision rules for the aggregated tables corresponding to each coalition using one of four rough set-based methods: exhaustive search algorithm, genetic algorithm, covering algorithm, or LEM2 algorithm.
- Generation of prediction vectors for training and test objects based on the derived rule sets, using the weighted rules method.
- Construction of decision templates for each decision class by averaging the prediction vectors of training objects belonging to that class.
- Classification of test objects from the test set by comparing their prediction vectors with the decision templates using normalized Euclidean distance, and assigning the final decision based on the best match.

The obtained results will be analyzed from two perspectives. First, the proposed approach will be compared to a baseline approach, in which no conflict analysis or coalition formation is applied; instead, decision rules are induced independently for each local table. Second, the performance of the proposed approach will be compared with selected results presented in previous studies [14, 15].

Tables 2 and 3 present the classification results for all dispersed datasets, corresponding to the proposed and baseline approaches, respectively. The tables include abbreviations for the analyzed rule induction methods: exhaustive search algorithm (Exh), genetic algorithm (Gen), covering algorithm (Cov), and LEM2 algorithm (LEM2). For the genetic algorithm, the impact of the number of reducts was investigated by testing values of 10, 100, and 1000. Ultimately, the results for 100 reducts are reported, as using 10 yielded inferior performance, while increasing the number to 1000 did not lead to any noticeable improvement in classification quality.

The results show variation depending on the dataset characteristics, the number of local tables, and the chosen rule induction strategy. Among the analyzed methods, the exhaustive search (Exh) and genetic (Gen) algorithms consistently achieve the highest performance metrics, regardless of the dataset and the level of dispersion. Their results are marked by high stability and robustness to changes in data structure. For the Car Evaluation and Balance Scale datasets, the obtained results are nearly identical, which may suggest that the decision rules generated by both algorithms overlap to a large extent. In contrast, the Vehicle dataset exhibits the greatest variation in performance across induction methods, which may indicate a higher sensitivity of this dataset to the choice of algorithm.

Noticeably weaker results were observed for the covering (Cov) and LEM2 algorithms – especially Cov, which in many configurations produced the lowest values across most of the analyzed metrics. It is worth noting, however, that metrics such as precision and G-mean are often relatively higher for these methods than for the remaining ones, which may suggest that the models exhibit selective alignment with one of the classes at the expense of overall classification performance.

The choice of a specific rule induction method within the proposed framework should take into account the trade-offs between flexibility, interpretability, and implementation simplicity. While the exhaustive search and genetic algorithms offer high-quality rule sets and are effective across a variety of settings, they tend to be more demanding in terms of computational resources and execution time, which may be relevant in applications with limited hardware capacity or large-scale data. In contrast, covering and LEM2 algorithms are simpler and require less implementation effort, but they exhibit greater variability in prediction quality and lower robustness to changes in data structure. Therefore, the selection of the appropriate method should

Table 2. Results of classification accuracy (Acc), balanced accuracy (BAcc), precision (Prec.), recall, F-measure (F-m.), and geometric mean (G-mean) for dispersed datasets using the proposed approach.

Dataset	No. of tables	Method	Acc	BAcc	Prec.	Recall	F-m.	G-mean
Vehicle	5	Exh	0.713	0.700	0.711	0.713	0.706	0.803
		Gen	0.685	0.673	0.676	0.685	0.676	0.784
		Cov	0.551	0.555	0.551	0.551	0.544	0.687
		LEM2	0.579	0.590	0.686	0.579	0.592	0.716
	7	Exh	0.701	0.686	0.697	0.701	0.693	0.796
		Gen	0.677	0.669	0.683	0.677	0.666	0.780
		Cov	0.488	0.488	0.488	0.488	0.485	0.636
		LEM2	0.500	0.521	0.638	0.500	0.505	0.659
	9	Exh	0.701	0.695	0.707	0.701	0.693	0.797
		Gen	0.689	0.684	0.690	0.689	0.681	0.788
		Cov	0.469	0.464	0.473	0.469	0.467	0.622
		LEM2	0.559	0.574	0.679	0.559	0.565	0.702
	11	Exh	0.697	0.683	0.700	0.697	0.689	0.793
		Gen	0.701	0.690	0.706	0.701	0.696	0.796
		Cov	0.516	0.512	0.521	0.516	0.511	0.657
		LEM2	0.504	0.493	0.583	0.504	0.508	0.649
Car	5	Exh	0.740	0.671	0.758	0.740	0.746	0.742
		Gen	0.744	0.674	0.761	0.744	0.750	0.745
		Cov	0.408	0.322	0.762	0.408	0.509	0.617
		LEM2	0.647	0.613	0.766	0.647	0.689	0.741
	7	Exh	0.748	0.641	0.760	0.748	0.752	0.743
		Gen	0.748	0.641	0.760	0.748	0.752	0.743
		Cov	0.408	0.322	0.762	0.408	0.509	0.617
		LEM2	0.651	0.562	0.740	0.651	0.684	0.723
	9	Exh	0.765	0.726	0.786	0.765	0.772	0.774
		Gen	0.765	0.726	0.786	0.765	0.772	0.774
		Cov	0.408	0.322	0.762	0.408	0.509	0.617
		LEM2	0.661	0.622	0.750	0.661	0.690	0.737
	11	Exh	0.765	0.758	0.790	0.765	0.773	0.783
		Gen	0.765	0.758	0.790	0.765	0.773	0.783
		Cov	0.408	0.322	0.762	0.408	0.509	0.617
		LEM2	0.655	0.672	0.774	0.655	0.690	0.751
Balance Scale	7	Exh	0.745	0.742	0.890	0.745	0.795	0.841
		Gen	0.745	0.742	0.890	0.745	0.795	0.841
		Cov	0.644	0.613	0.783	0.644	0.690	0.749
		LEM2	0.628	0.583	0.772	0.628	0.681	0.741
	9	Exh	0.686	0.681	0.856	0.686	0.745	0.798
		Gen	0.686	0.681	0.856	0.686	0.745	0.798
		Cov	0.590	0.501	0.700	0.590	0.636	0.697
		LEM2	0.596	0.542	0.747	0.596	0.653	0.715
	11	Exh	0.697	0.670	0.863	0.697	0.756	0.806
		Gen	0.697	0.670	0.863	0.697	0.756	0.806
		Cov	0.511	0.462	0.628	0.511	0.554	0.629
		LEM2	0.559	0.497	0.754	0.559	0.632	0.696

be guided by the requirements of a particular use case: when high accuracy and transparent rule sets are critical, Exh or Gen are preferable; in scenarios where simplicity or computational efficiency is prioritized, Cov or LEM2 may be more suitable.

The results achieved with the proposed approach are comparable to those of the baseline method. However, it is worth emphasizing that the model developed within the proposed framework offers a high level of interpretability, which constitutes a significant advantage in practical applications. Considering this potential, future work will focus on relaxing the full coverage requirement, enabling classification based on partial rule matching. This modification is expected to improve the overall effectiveness of the proposed approach. Statistical tests were performed to confirm the there are no statistically significant differences in mean values, F-measure values were used for comparison. Two dependent samples each containing of 44 observations were created – results for the proposed approach and the baseline approach from Tables 2 and 3. The Wilcoxon test indicated that there is no statistically significant difference in mean F-measure for these two approaches, $p = 0.087$. A comparative box plot illustrating the F-measure results for the two methods is provided in Figure 1. As can be observed, the median F-measure value for the baseline approach is slightly lower than that of the proposed approach. On the other

Table 3. Results of classification accuracy (Acc), balanced accuracy (BAcc), precision (Prec.), recall, F-measure (F-m.), and geometric mean (G-mean) for dispersed datasets using the baseline approach.

Dataset	No. of tables	Method	Acc	BAcc	Prec.	Recall	F-m.	G-mean
Vehicle	5	Exh	0.717	0.703	0.715	0.717	0.708	0.807
		Gen	0.693	0.690	0.705	0.693	0.687	0.791
		Cov	0.583	0.581	0.584	0.583	0.578	0.709
		LEM2	0.594	0.603	0.689	0.594	0.603	0.727
	7	Exh	0.685	0.674	0.683	0.685	0.676	0.784
		Gen	0.654	0.649	0.660	0.654	0.645	0.763
		Cov	0.583	0.580	0.588	0.583	0.580	0.709
		LEM2	0.547	0.563	0.613	0.547	0.546	0.691
	9	Exh	0.693	0.682	0.694	0.693	0.684	0.790
		Gen	0.693	0.687	0.701	0.693	0.688	0.791
		Cov	0.591	0.586	0.609	0.591	0.586	0.716
		LEM2	0.606	0.598	0.631	0.606	0.615	0.728
	11	Exh	0.669	0.655	0.668	0.669	0.657	0.772
		Gen	0.650	0.646	0.675	0.650	0.640	0.760
		Cov	0.598	0.595	0.616	0.598	0.596	0.722
		LEM2	0.571	0.565	0.608	0.571	0.580	0.705
Car	5	Exh	0.750	0.706	0.763	0.750	0.755	0.746
		Gen	0.750	0.706	0.763	0.750	0.755	0.746
		Cov	0.408	0.322	0.762	0.408	0.509	0.617
		LEM2	0.657	0.607	0.759	0.657	0.690	0.740
	7	Exh	0.753	0.736	0.777	0.753	0.761	0.769
		Gen	0.753	0.736	0.777	0.753	0.761	0.769
		Cov	0.408	0.322	0.762	0.408	0.509	0.617
		LEM2	0.663	0.566	0.754	0.663	0.693	0.740
	9	Exh	0.767	0.757	0.787	0.767	0.773	0.775
		Gen	0.765	0.755	0.784	0.765	0.771	0.772
		Cov	0.408	0.322	0.762	0.408	0.509	0.617
		LEM2	0.705	0.657	0.786	0.705	0.730	0.779
	11	Exh	0.742	0.712	0.774	0.742	0.752	0.767
		Gen	0.742	0.713	0.776	0.742	0.753	0.769
		Cov	0.410	0.416	0.763	0.410	0.505	0.621
		LEM2	0.642	0.639	0.769	0.642	0.675	0.745
Balance Scale	7	Exh	0.713	0.774	0.922	0.713	0.776	0.830
		Gen	0.707	0.770	0.922	0.707	0.772	0.827
		Cov	0.601	0.546	0.720	0.601	0.641	0.704
		LEM2	0.580	0.567	0.819	0.580	0.659	0.725
	9	Exh	0.718	0.759	0.901	0.718	0.776	0.828
		Gen	0.718	0.759	0.901	0.718	0.776	0.828
		Cov	0.612	0.535	0.731	0.612	0.659	0.719
		LEM2	0.644	0.577	0.793	0.644	0.701	0.756
	11	Exh	0.713	0.737	0.899	0.713	0.773	0.825
		Gen	0.713	0.737	0.899	0.713	0.773	0.825
		Cov	0.644	0.522	0.722	0.644	0.677	0.731
		LEM2	0.537	0.536	0.786	0.537	0.618	0.692

hand, the baseline approach exhibits less variation than the proposed approach.

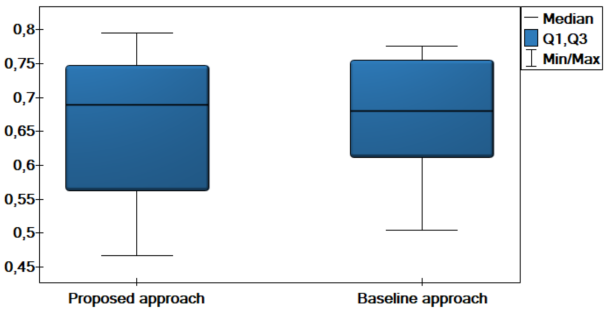

Fig. 1. Comparison of F-measure obtained for the proposed and the baseline approaches.

In previous papers [14, 15], alternative approaches to the classification of distributed data were proposed, based on the same mechanism for aggregating local decision tables into coalitions. Both studies applied a classification process based on decision trees. Study [14] analyzed

the impact of coalition formation on classification quality, while study [15] extended this approach by introducing a stopping criterion referring to the minimum number of instances in a decision tree node. The experiments considered three values of this parameter: 2, 7, and 12. Tables 4 and 5 present the selected results for the approaches described in [14] and [15], respectively, including both the coalition-based and baseline variants. The results for the Vehicle dataset are taken directly from the original papers, whereas the experiments for the Car Evaluation and Balance Scale datasets were conducted as part of the present study to enable comparison with the current rule-based approach on the same data. In addition to classification accuracy (Acc), the results include the Acc_{ONE} metric, which reflects the performance of unambiguous classification, and the $\bar{d}$ index, which indicates the average number of ties in the classification. Better classification accuracy results are highlighted in blue.

Table 4. Comparison of classification results for the coalition-based and baseline approaches using decision tree models, based on methods from [14].

Dataset	No. of tables	Proposed approach			Baseline approach		
		Acc	Acc_{ONE}	$\bar{d}$	Acc	Acc_{ONE}	$\bar{d}$
Vehicle	5	0.791	0.709	1.173	0.756	0.677	1.094
	7	0.780	0.669	1.228	0.752	0.681	1.114
	9	0.740	0.685	1.075	0.760	0.693	1.098
	11	0.776	0.728	1.051	0.740	0.673	1.087
Car	5	0.842	0.690	1.158	0.802	0.778	1.031
	7	0.836	0.721	1.123	0.769	0.741	1.037
	9	0.786	0.753	1.035	0.825	0.769	1.064
	11	0.811	0.734	1.083	0.803	0.780	1.025
Balance Scale	7	0.851	0.676	1.202	0.846	0.798	1.059
	9	0.872	0.777	1.165	0.862	0.846	1.037
	11	0.856	0.771	1.154	0.824	0.771	1.074

Table 5. Comparison of classification results for the coalition-based and baseline approaches using decision tree models with a stopping criterion (minimum number of instances in a node), based on methods from [15].

Dataset	No. of tables	Stop criteria	Proposed approach			Baseline approach		
			Acc	Acc_{ONE}	$\bar{d}$	Acc	Acc_{ONE}	$\bar{d}$
Vehicle	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
Car	5	2	0.850	0.692	1.168	0.792	0.763	1.035
		7	0.819	0.701	1.121	0.790	0.757	1.044
		12	0.844	0.709	1.152	0.798	0.755	1.060
	7	2	0.834	0.719	1.127	0.751	0.724	1.031
		7	0.807	0.736	1.075	0.800	0.771	1.042
		12	0.823	0.751	1.075	0.798	0.782	1.015
	9	2	0.778	0.736	1.105	0.825	0.773	1.064
		7	0.778	0.759	1.023	0.800	0.775	1.033
		12	0.800	0.751	1.039	0.771	0.757	1.023
	11	2	0.809	0.748	1.066	0.798	0.773	1.031
		7	0.815	0.753	1.073	0.792	0.763	1.037
		12	0.834	0.748	1.094	0.798	0.773	1.031
Balance Scale	7	2	0.856	0.686	1.197	0.840	0.782	1.064
		7	0.830	0.713	1.138	0.824	0.745	1.085
		12	0.824	0.702	1.144	0.803	0.734	1.074
	9	2	0.856	0.766	1.144	0.851	0.824	1.059
		7	0.846	0.702	1.223	0.804	0.777	1.074
		12	0.782	0.702	1.090	0.782	0.707	1.074
	11	2	0.846	0.766	1.144	0.830	0.793	1.069
		7	0.851	0.782	1.128	0.883	0.793	1.112
		12	0.814	0.729	1.112	0.883	0.755	1.128

As can be seen, the results indicate a clear advantage of the coalition-based approach over the baseline one – in almost all cases, it delivers better classification performance. The differences between the compared methods are more pronounced than in the case of the currently proposed approach based on decision rules. This may be due to the fact that tree-based models offer greater flexibility in fitting the data and more effective use of the structural information in the dataset.

On the other hand, the current rule-based method offers greater transparency and interpretability, which is a significant advantage in applications that require explainable decision–making. Despite less pronounced differences compared to the baseline approach, the results remain competitive, indicating potential for further development.

4. Conclusion

This study introduced a novel method for the classification of distributed data, combining the mechanism of forming coalitions of local decision tables with interpretable decision rule induction and decision templates. The method enables the integration of knowledge from multiple sources while maintaining model transparency and robustness to data heterogeneity.

The effectiveness of the proposed approach was evaluated using three datasets: Vehicle Silhouettes, Car Evaluation, and Balance Scale. Four rule induction algorithms were analyzed: exhaustive search, genetic, covering, and LEM2. The best classification results were achieved using the exhaustive and genetic algorithms, which demonstrated high stability. In contrast, the covering and LEM2 algorithms showed significantly weaker overall classification performance. The results indicate that the proposed approach achieves classification quality comparable to the baseline method, which does not incorporate conflict analysis or the coalition mechanism. At the same time, the proposed approach offers significantly greater model interpretability.

In addition, the performance of the currently proposed solution was compared with the Authors' previous approaches based on decision tree classifiers. The results suggest that tree-based models provide greater flexibility but at the expense of reduced transparency. The current method, due to the simplicity of the decision rule structure, is particularly suitable for scenarios requiring explainable decisions.

The applicability of the method can be illustrated across several critical real-world contexts. In medical decision support, for example, rule-based models can assist clinicians by providing clear explanations for diagnostic classification or patient risk assessment, thereby increasing trust in automated systems. In the banking sector, interpretable rules are valuable for credit scoring and fraud detection, where decisions must be auditable and regulatory compliance is essential. Similarly, in educational analytics, simple decision rules can help implement personalized learning strategies in a way that is both actionable and understandable for educators.

In future work, the proposed approach will be extended, including the introduction of classification based on partial rule coverage. Additionally, the effectiveness of the method is planned to be evaluated in combination with a decision tree classifier.

References

[1] Bazan, J.G., Nguyen, H.S., Nguyen, S.H., Synak, P., Wróblewski, J.: Rough set algorithms in classification problem. Rough set methods and applications: new developments in knowledge discovery in information systems pp. 49–88 (2000)

[2] Bazan, J.G., Szczuka, M.: Rses and rseslib-a collection of tools for rough set computations. In: Rough Sets and Current Trends in Computing: Second International Conference, RSCTC 2000 Banff, Canada, October 16–19, 2000 Revised Papers 2. pp. 106–113. Springer (2001)

[3] Bazan, J.G., Szczuka, M.: The rough set exploration system. In: Transactions on rough sets III, pp. 37–56. Springer (2005)

[4] Bohanec, M., Rajkovič, V.: Knowledge acquisition and explanation for multi-attribute decision making. 8th Intl Workshop on Expert Systems and their Applications, Avignon, France (1988)

[5] Dua, D., Graff, C.: Uci machine learning repository. irvine, ca: University of california, school of information and computer science. IEEE transactions on pattern analysis and machine intelligence 1(1), pp. 1–29 (2019)

[6] Gillani, Z., Bashir, Z., Aquil, S.: A game theoretic conflict analysis model with linguistic assessments and two levels of game play. Information Sciences 677, pp. 120840 (2024)

[7] Grzymala-Busse, J.W.: A new version of the rule induction system lers. Fundamenta Informaticae 31(1), pp. 27–39 (1997)

[8] Hu, X., Zhu, X., Yang, L., Pedrycz, W., Li, Z.: A design of fuzzy rule-based classifier for multi-class classification and its realization in horizontal federated learning. IEEE Transactions on Fuzzy Systems (2024)

[9] Komorniczak, J., Ksieniewicz, P.: Distance profile layer for binary classification and density estimation. Neurocomputing 579, pp. 127436 (2024)

[10] Kuncheva, L.I., Bezdek, J.C., Duin, R.P.: Decision templates for multiple classifier fusion: an experimental comparison. Pattern recognition 34(2), pp. 299–314 (2001)

[11] Pawlak, Z.: Rough sets. International journal of computer & information sciences 11, pp. 341–356 (1982)

[12] Pawlak, Z.: Conflict analysis. In: Proceedings of the Fifth European Congress on Intelligent Techniques and Soft Computing (EUFIT'97). vol. 1589, p. 1591 (1997)

[13] Pękala, B., Wilbik, A., Szkoła, J., Dyczkowski, K., Żywica, P.: Federated learning with the choquet integral as aggregation method. In: 2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). pp. 1–8. IEEE (2024)

[14] Przybyła-Kasperek, M., Kusztal, K.: New classification method for independent data sources using pawlak conflict model and decision trees. Entropy 24(11), pp. 1604 (2022)

[15] Przybyła-Kasperek, M., Kusztal, K.: Rules' quality generated by the classification method for independent data sources using pawlak conflict analysis model. In: International Conference on Computational Science. pp. 390–405. Springer (2023)

[16] Przybyła-Kasperek, M., Sacewicz, J.: Ensembles of random trees with coalitions-a classification model for dispersed data. Procedia Computer Science 246, pp. 1599–1608 (2024)

[17] Przybyła-Kasperek, M., Wakulicz-Deja, A.: A dispersed decision-making system–the use of negotiations during the dynamic generation of a system's structure. Information Sciences 288, pp. 194–219 (2014)

[18] Siebert, J.P.: Vehicle recognition using rule based methods. Turing Institute (1987)

[19] Trajdos, P., Burduk, R.: Ensemble of classifiers based on score function defined by clusters and decision boundary of linear base learners. Knowledge-Based Systems 303, pp. 112411 (2024)

[20] Yan, Y., Li, X.: Three-way conflict analysis with negative feedback. In: International Joint Conference on Rough Sets. pp. 196–209. Springer (2024)

[21] Yao, Y.: Set-theoretic models of three-way decision. Granular Computing 6(1), pp. 133–148 (2021)

[22] Yao, Y.: The dao of three-way decision and three-world thinking. International Journal of Approximate Reasoning 162, pp. 109032 (2023)

[23] Zyblewski, P., Sabourin, R., Woźniak, M.: Preprocessed dynamic classifier ensemble selection for highly imbalanced drifted data streams. Information Fusion 66, pp. 138–154 (2021)