Handling Class Imbalance in Coalition-Based Distributed Classification with Decision Rule Induction

Katarzyna Kusztal[1][0000−0002−9970−5339] and Małgorzata Przybyła-Kasperek[1,2][0000−0003−0616−9694]

[1] University of Silesia in Katowice, Institute of Computer Science, Będzińska 39, 41-200 Sosnowiec, Poland
{katarzyna.kusztal,malgorzata.przybyla-kasperek}@us.edu.pl
[2] Constantine the Philosopher University in Nitra, Department of Informatics, Tr. A. Hlinku 1, 949 01 Nitra, Slovakia

Abstract. Learning from distributed data, maintained independently and analyzed without full central integration, poses significant challenges for building coherent and reliable classification models. In such environments, local datasets may differ in both content and class distributions, affecting the quality of the resulting global model. This paper extends the authors' previously proposed distributed classification framework integrating conflict analysis, coalition formation, and decision rule induction. The main novelty lies in incorporating a class balancing stage applied independently to each local dataset prior to system construction. Six representative data-level balancing techniques are examined, along with four rough set-based rule induction algorithms and three decision-making strategies. Experiments were conducted on two datasets from the UCI Machine Learning Repository: Car Evaluation and Balance Scale. The proposed approach was compared with a baseline without class balancing. The results indicate that class distribution adjustment improves imbalance-sensitive metrics under severe class imbalance, with a moderate reduction in overall accuracy.

Keywords: Distributed classification · Class imbalance · Hierarchical framework · Conflict analysis · Coalition-based modeling · Rule induction

1 Introduction

Contemporary information systems increasingly operate in environments where data are generated and managed by multiple independent entities. Due to organizational, legal, and security constraints, these data are typically stored locally, making their centralization difficult or infeasible. In this context, constructing decision models requires effective integration of local information while ensuring global consistency, which presents a significant computational challenge and requires efficient and transparent algorithms.

An important issue in data analysis, including distributed settings, is class imbalance. In many real-world applications, class distributions can be highly

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_33

uneven, which disrupts the learning process and limits the model's ability to correctly distinguish all classes [14]. As a result, models trained on imbalanced data may become biased toward the majority class, overlooking rare but important instances. Although numerous balancing techniques have been proposed [12], they have been studied mainly in centralized settings [16], while their role in distributed environments remains largely unexplored.

This paper constitutes a significant extension of the distributed data classification approach presented in the authors' earlier study [11], which relies on conflict analysis, coalition formation (groups of local data sources cooperating in decision-making), and decision rule induction. The proposed approach incorporates a class balancing stage preceding the construction of the global model. From the perspective of multi-criteria decision-making, the framework enables the integration of multiple decision criteria, while supporting the analysis of decision trade-offs and the assessment of decision quality under heterogeneous conditions.

The aim of this study is formulated through the following research questions:

- How does the incorporation of class balancing techniques into local datasets influence the quality of the resulting global classification model in a distributed framework?
- How do different class balancing techniques affect classification performance under varying levels of class imbalance?

Six techniques are considered: Random Undersampling, NearMiss, Tomek Links, Random Oversampling, SMOTE, and SMOTE-Tomek. Local datasets are balanced independently, followed by conflict analysis and the induction of decision rules, which form the basis for the final classification process.

The main contributions of this paper are as follows:

- Integration of class balancing techniques into a distributed classification framework based on conflict analysis, coalition formation, and decision rule induction.
- Comparative analysis of multiple balancing strategies in distributed settings, including statistical evaluation.
- Assessment of the impact of class balancing on classification performance under different levels of class imbalance.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed framework. Section 4 describes the experimental setup and results. Section 5 concludes the paper and discusses future research directions.

2 Related Work

In the literature on imbalanced data, three main groups of methods are distinguished: data-level, algorithm-level, and hybrid approaches [6]. Data-level techniques modify class distributions prior to model training, including undersampling, oversampling, and their combinations, with widely used methods such as

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_33

SMOTE and its variants [2]. In contrast, algorithm-level approaches improve minority class recognition by modifying the learning process, for example through cost-sensitive learning or ensemble methods [5]. In addition, hybrid strategies combine both approaches to leverage their advantages [3].

Despite extensive research on imbalanced data, most studies assume a centralized setting. By comparison, class imbalance in distributed environments has received limited attention. In particular, [10] investigated balancing techniques applied to independent local datasets within a distributed framework employing coalition mechanisms. However, in that approach, coalitions are formed dynamically for each test object. In contrast, the framework presented in this paper integrates class balancing into a static construction process, in which coalitions are determined once and remain fixed during model operation, reducing computational complexity. Moreover, while [10] used the k-nearest neighbors algorithm, this study relies on decision rule induction, providing an explicit and interpretable representation of the decision process.

From a broader perspective, distributed data classification has mainly evolved along two lines: ensemble-based learning [18] and federated learning [13]. Ensemble methods typically rely on centrally available data and are therefore not suited to independently maintained data sources. In contrast, federated learning enables decentralized training while preserving data privacy. However, it does not explicitly model relationships between local datasets and often relies on complex, less interpretable models such as deep neural networks [9].

At the same time, distributed data environments have also been examined using conflict analysis. A fundamental approach in this domain is Pawlak's conflict analysis model, which identifies and describes agreement and disagreement between agents [8]. The model has been further developed within rough set theory and extended to broader decision-making contexts, including three-way decision theory [17].

Overall, the role of class balancing techniques in distributed data environments remains insufficiently explored. In particular, their integration with static coalition formation mechanisms grounded in conflict analysis and coupled with interpretable decision rule induction has not yet been systematically examined.

3 Distributed Framework with Class Balancing

In this paper, we present an extension of the distributed classification framework proposed in [11], incorporating class balancing at the initial stage of system construction. The proposed approach retains its hierarchical structure, in which successive stages are executed in a predefined order.

In the adopted formal setting, distributed data are modeled as a set of local decision tables $T = \{T_i : i \in \{1, \ldots, n\}\}$. Each local table T_i is defined as a triple (U_i, A, d), where U_i represents a set of objects, A denotes the set of conditional attributes, and d is the decision attribute. All tables correspond to the same decision problem and therefore share the same conditional attributes and decision variable.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_33

To address class imbalance within local tables, class balancing is applied independently to each table prior to further processing. The following six data-level techniques are considered: Random Undersampling (RUS), NearMiss [7], Tomek Links [15], Random Oversampling (ROS), SMOTE [2], and SMOTE-Tomek [1]. These methods were selected due to their distinct mechanisms and impact on class distribution.

After balancing, the subsequent steps of the distributed classification process follow the framework described in [11], with all operations now performed on balanced local tables. The main stages are summarized below:

1. Conflict analysis is formulated according to Pawlak's conflict analysis model [8]. Local tables are compared based on encoded descriptors of conditional attributes. For each attribute $a \in A$, a three-valued function $a : T \to \{-1, 0, 1\}$ is defined, representing the relative position of attribute values within the global distribution across all local tables. The degree of conflict between two tables T_i and T_j is measured by:
$$\rho(T_i, T_j) = \frac{\text{card}\{a \in A : a(T_i) \neq a(T_j)\}}{\text{card}\{A\}},$$
which quantifies the proportion of attributes on which the tables differ.
2. Coalitions are defined as subsets of local tables for which $\rho(T_i, T_j) < 0.5$ for every pair of elements. The threshold of 0.5 follows Pawlak's conflict analysis model and corresponds to agreement on more than half of the conditional attributes.
3. For the j-th coalition, an aggregated decision table $T_j^{aggr} = (U_j^{aggr}, A, d)$ is constructed by merging all objects originating from the local tables belonging to that coalition.
4. Decision rules are induced from each aggregated table using four rough set-based methods: exhaustive search algorithm, genetic algorithm, covering algorithm, and LEM2.
5. The induced rule sets are used to classify test objects according to three decision-making strategies:
   - First Rule Approach (FRA) – the decision is based on the first matching rule;
   - All Rules Approach (ARA) – the decision is based on the majority class among the matching rules;
   - Weighted Rule Approach (WRA) – the decision is based on weighted voting, where each rule contributes proportionally to its number of matches.

Since class balancing modifies the internal structure of local tables, it may influence inter-table similarity, coalition configuration, and the resulting rule-based models. Thus, the preprocessing phase plays a structural role in the behavior of the entire system.

4 Experimental Evaluation

The experiments were conducted on two benchmark datasets from the UCI repository [4]: Car Evaluation and Balance Scale. The Car Evaluation dataset

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_33

contains 1,728 instances with six categorical attributes and four decision classes: unacc, acc, good, vgood. The Balance Scale dataset consists of 625 instances with four categorical attributes and three classes: R, L, B. For each dataset, 70% of objects were assigned to the training set and 30% to the test set using stratified sampling. The training data were further partitioned into 7, 9, and 11 local tables, preserving class distributions, with additional sampling from the remaining tables when necessary to ensure equal class counts.

The datasets exhibit different levels of class imbalance. In the Car Evaluation, the unacc class constitutes approximately 70% of instances, while the good and vgood classes each account for about 4%, indicating strong imbalance. In contrast, the Balance Scale is more balanced, with the R and L classes representing approximately 46% each and the B class about 8%. For undersampling and oversampling techniques, balancing was applied to enforce equal class proportions within each local table, while hybrid methods may introduce slight deviations. For SMOTE and NearMiss, the number of neighbors was set to $k = 3$, which is a commonly used configuration. In addition to the proposed approach, a baseline variant without class balancing was considered for comparative purposes. The performance was evaluated on the test sets using classification accuracy (Acc), balanced accuracy (BAcc), precision (Prec.), recall (Rec.), F-measure (F.-m.), and geometric mean (G-mean), with BAcc and G-mean particularly relevant for imbalanced data.

The experimental procedure included class balancing of local tables, coalition formation, rule induction from aggregated tables, and classification using FRA, ARA, or WRA. All experiments were conducted with a fixed random seed to ensure reproducibility.

Table 1 presents the comparative results of the proposed approach and the baseline approach for the Car Evaluation and Balance Scale datasets. As no consistent dominance of any decision-making strategy (FRA, ARA, WRA) was observed across datasets and configurations, the results are averaged across these strategies. For the genetic algorithm, preliminary experiments showed no significant performance differences for varying numbers of reducts; therefore, only the configuration with 100 reducts is considered.

For the Car Evaluation dataset, the baseline typically achieves higher overall accuracy, as expected under severe class imbalance, but at the cost of substantially lower balanced accuracy and G-mean. For example, with 9 local tables, the baseline reaches Acc = 0.744, while BAcc drops to 0.387 and G-mean to 0.619. Under the same configuration, the proposed approach improves class-balanced performance across all balancing techniques (BAcc: 0.506-0.538, G-mean: 0.691-0.712), with a moderate reduction in accuracy (0.660-0.675), leading to a more balanced predictive behavior across decision classes. A complementary view is provided in Fig. 1, where each point represents the average performance of a given balancing technique across all experimental configurations. For the Car Evaluation, most techniques form a cluster corresponding to lower overall accuracy and improved balanced accuracy. Notably, one technique (SMOTE) achieves higher values for both Acc and BAcc than the baseline. In contrast, for the Bal-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_33

Table 1. Results for the proposed and baseline approaches.

BT	Method	7 local tables Acc/BAcc/Prec./Rec./F.-m./G-mean	9 local tables Acc/BAcc/Prec./Rec./F.-m./G-mean	11 local tables Acc/BAcc/Prec./Rec./F.-m./G-mean
CAR EVALUATION				
Proposed approach				
RUS	Exh	0.592/0.682/0.732/0.592/0.631/0.705	0.671/0.515/0.721/0.671/0.692/0.712	0.681/0.515/0.712/0.681/0.694/0.701
	Gen	0.583/0.676/0.728/0.583/0.623/0.700	0.661/0.519/0.711/0.661/0.681/0.699	0.690/0.548/0.720/0.690/0.702/0.710
	Cov	0.585/0.672/0.729/0.585/0.625/0.702	0.676/0.535/0.723/0.676/0.694/0.711	0.674/0.517/0.704/0.674/0.686/0.692
	LEM2	0.577/0.673/0.719/0.577/0.616/0.693	0.661/0.508/0.711/0.661/0.681/0.699	0.668/0.464/0.699/0.668/0.681/0.690
NearMiss	Exh	0.583/0.682/0.726/0.583/0.622/0.697	0.670/0.533/0.720/0.670/0.690/0.709	0.673/0.493/0.704/0.673/0.686/0.691
	Gen	0.581/0.673/0.721/0.581/0.619/0.695	0.674/0.542/0.721/0.674/0.692/0.711	0.673/0.508/0.707/0.673/0.687/0.699
	Cov	0.573/0.664/0.725/0.573/0.614/0.693	0.665/0.518/0.711/0.665/0.683/0.701	0.690/0.533/0.716/0.690/0.701/0.708
	LEM2	0.586/0.674/0.730/0.586/0.626/0.701	0.673/0.552/0.724/0.673/0.693/0.712	0.686/0.527/0.713/0.686/0.696/0.702
Tomek Links	Exh	0.586/0.681/0.721/0.586/0.623/0.695	0.675/0.530/0.723/0.675/0.694/0.712	0.678/0.509/0.708/0.678/0.690/0.696
	Gen	0.591/0.685/0.735/0.591/0.630/0.706	0.662/0.533/0.715/0.662/0.683/0.702	0.683/0.535/0.719/0.683/0.698/0.711
	Cov	0.584/0.667/0.729/0.584/0.625/0.701	0.659/0.535/0.707/0.659/0.678/0.695	0.675/0.507/0.709/0.675/0.689/0.697
	LEM2	0.590/0.681/0.734/0.590/0.630/0.705	0.669/0.524/0.715/0.669/0.687/0.703	0.674/0.475/0.705/0.674/0.687/0.696
ROS	Exh	0.584/0.676/0.730/0.584/0.624/0.703	0.660/0.518/0.710/0.660/0.680/0.697	0.682/0.518/0.718/0.682/0.696/0.708
	Gen	0.573/0.662/0.720/0.573/0.613/0.691	0.658/0.515/0.707/0.658/0.677/0.695	0.676/0.490/0.708/0.676/0.689/0.698
	Cov	0.598/0.700/0.737/0.598/0.636/0.712	0.664/0.525/0.716/0.664/0.684/0.705	0.670/0.525/0.706/0.670/0.684/0.692
	LEM2	0.582/0.682/0.721/0.582/0.619/0.694	0.656/0.507/0.711/0.656/0.678/0.698	0.672/0.522/0.706/0.672/0.686/0.696
SMOTE	Exh	NO COALITIONS	0.664/0.538/0.705/0.664/0.680/0.691	0.677/0.512/0.710/0.677/0.691/0.698
	Gen		0.671/0.547/0.716/0.671/0.688/0.705	0.677/0.533/0.711/0.677/0.691/0.699
	Cov		0.678/0.568/0.725/0.678/0.696/0.713	0.675/0.486/0.705/0.675/0.687/0.697
	LEM2		0.686/0.549/0.733/0.686/0.705/0.726	0.681/0.545/0.712/0.681/0.693/0.700
SMOTE-Tomek	Exh	0.577/0.682/0.722/0.577/0.616/0.695	0.668/0.506/0.715/0.668/0.687/0.704	0.686/0.492/0.721/0.686/0.701/0.714
	Gen	0.577/0.657/0.715/0.577/0.615/0.689	0.669/0.511/0.715/0.669/0.688/0.704	0.671/0.508/0.701/0.671/0.684/0.691
	Cov	0.585/0.686/0.726/0.585/0.623/0.700	0.672/0.515/0.716/0.672/0.690/0.705	0.681/0.487/0.713/0.681/0.694/0.707
	LEM2	0.586/0.675/0.723/0.586/0.624/0.699	0.662/0.509/0.715/0.662/0.683/0.702	0.682/0.516/0.711/0.682/0.694/0.701
Baseline approach				
	Exh	0.732/0.433/0.717/0.732/0.700/0.629	0.744/0.387/0.726/0.744/0.697/0.619	0.747/0.409/0.726/0.747/0.707/0.626
	Gen	0.732/0.422/0.710/0.732/0.700/0.630	0.744/0.388/0.727/0.744/0.698/0.621	0.740/0.420/0.710/0.740/0.703/0.624
	Cov	0.488/0.344/0.683/0.488/0.561/0.627	0.492/0.363/0.668/0.492/0.557/0.621	0.492/0.365/0.670/0.492/0.558/0.621
	LEM2	0.674/0.457/0.697/0.674/0.684/0.687	0.701/0.489/0.716/0.701/0.707/0.705	0.679/0.485/0.716/0.679/0.694/0.707
BALANCE SCALE				
Proposed approach				
RUS	Exh	0.729/0.571/0.759/0.729/0.741/0.787	0.597/0.494/0.658/0.597/0.623/0.683	0.633/0.526/0.689/0.633/0.657/0.712
	Gen	0.732/0.555/0.754/0.732/0.742/0.789	0.615/0.513/0.652/0.615/0.631/0.687	0.624/0.501/0.672/0.624/0.645/0.700
	Cov	0.743/0.569/0.764/0.743/0.752/0.797	0.626/0.496/0.674/0.626/0.647/0.703	0.619/0.473/0.682/0.619/0.648/0.704
	LEM2	0.720/0.552/0.759/0.720/0.737/0.784	0.610/0.521/0.670/0.610/0.634/0.693	0.631/0.518/0.687/0.631/0.656/0.711
NearMiss	Exh	0.734/0.562/0.765/0.734/0.748/0.794	0.606/0.543/0.679/0.606/0.634/0.695	0.637/0.516/0.679/0.637/0.655/0.709
	Gen	0.727/0.557/0.744/0.727/0.735/0.781	0.635/0.546/0.677/0.635/0.652/0.707	0.638/0.517/0.688/0.638/0.660/0.714
	Cov	0.732/0.561/0.756/0.732/0.743/0.789	0.645/0.523/0.677/0.645/0.659/0.712	0.656/0.561/0.701/0.656/0.675/0.727
	LEM2	0.739/0.579/0.758/0.739/0.748/0.793	0.615/0.507/0.667/0.615/0.638/0.695	0.637/0.541/0.687/0.637/0.658/0.713
Tomek Links	Exh	0.723/0.555/0.753/0.723/0.736/0.783	0.633/0.538/0.680/0.633/0.653/0.708	0.621/0.511/0.683/0.621/0.647/0.703
	Gen	0.736/0.570/0.749/0.736/0.740/0.786	0.649/0.544/0.687/0.649/0.665/0.718	0.638/0.530/0.685/0.638/0.659/0.713
	Cov	0.723/0.549/0.751/0.723/0.736/0.784	0.622/0.506/0.669/0.622/0.643/0.698	0.610/0.485/0.663/0.610/0.633/0.690
	LEM2	0.718/0.557/0.746/0.718/0.731/0.778	0.619/0.522/0.670/0.619/0.640/0.696	0.597/0.494/0.662/0.597/0.625/0.684
ROS	Exh	NO COALITIONS	0.640/0.568/0.697/0.640/0.663/0.718	0.633/0.508/0.674/0.633/0.651/0.706
	Gen		0.615/0.550/0.677/0.615/0.640/0.699	0.619/0.510/0.666/0.619/0.640/0.696
	Cov		0.633/0.538/0.700/0.633/0.661/0.717	0.633/0.514/0.686/0.633/0.656/0.710
	LEM2		0.615/0.544/0.671/0.615/0.638/0.696	0.646/0.517/0.692/0.646/0.666/0.720
SMOTE	Exh	NO COALITIONS	0.640/0.555/0.686/0.640/0.659/0.713	0.631/0.513/0.687/0.631/0.656/0.710
	Gen		0.613/0.536/0.680/0.613/0.640/0.699	0.628/0.497/0.678/0.628/0.650/0.706
	Cov		0.574/0.471/0.644/0.574/0.604/0.666	0.631/0.549/0.693/0.631/0.656/0.712
	LEM2		0.636/0.553/0.686/0.636/0.657/0.712	0.638/0.511/0.691/0.638/0.662/0.716
SMOTE-Tomek	Exh_FRA	0.741/0.586/0.769/0.741/0.753/0.799	0.626/0.551/0.670/0.626/0.644/0.700	0.633/0.538/0.692/0.633/0.657/0.714
	Gen	0.727/0.564/0.764/0.727/0.743/0.789	0.627/0.528/0.670/0.627/0.645/0.701	0.630/0.505/0.681/0.630/0.652/0.707
	Cov	0.732/0.555/0.761/0.732/0.746/0.792	0.603/0.504/0.649/0.603/0.622/0.681	0.660/0.533/0.696/0.660/0.674/0.726
	LEM2	0.750/0.592/0.775/0.750/0.761/0.804	0.663/0.536/0.705/0.663/0.681/0.733	0.621/0.505/0.671/0.621/0.643/0.699
Baseline approach				
	Exh	0.798/0.578/0.738/0.798/0.765/0.813	0.802/0.581/0.739/0.802/0.768/0.816	0.789/0.572/0.734/0.789/0.760/0.808
	Gen	0.798/0.578/0.738/0.798/0.765/0.813	0.802/0.581/0.739/0.802/0.768/0.816	0.789/0.572/0.734/0.789/0.760/0.808
	Cov	0.674/0.537/0.679/0.674/0.675/0.724	0.603/0.474/0.618/0.603/0.607/0.664	0.536/0.425/0.567/0.536/0.547/0.611
	LEM2	0.608/0.502/0.659/0.608/0.630/0.688	0.620/0.486/0.642/0.620/0.630/0.685	0.647/0.512/0.646/0.647/0.646/0.697

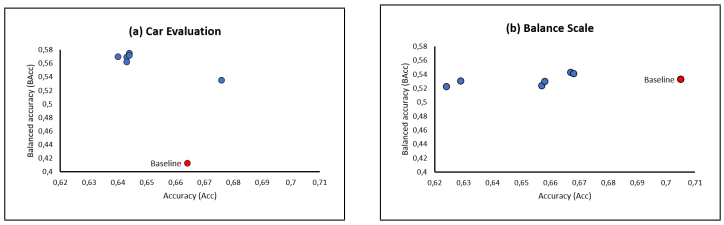

Fig. 1. Comparison of accuracy (Acc) and balanced accuracy (BAcc) for different balancing techniques: (a) Car Evaluation, (b) Balance Scale.

ance Scale dataset, differences between approaches are less pronounced, and performance remains comparable across configurations, reflecting the milder class imbalance. This is also demonstrated in Fig. 1, where the results show greater variability, particularly in accuracy, indicating less consistent behavior across techniques. Furthermore, in some configurations (Car Evaluation with 7 local tables for SMOTE, and Balance Scale for ROS and SMOTE), no compatible coalitions were identified, suggesting that balancing techniques may influence the structural relations between local tables.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_33

We compared seven approaches to handling class imbalance (six class balancing techniques and a baseline) using the F-measure as the performance metric. The statistical analysis was carried out on the original results before averaging across decision strategies, yielding 48 paired observations (one per dataset/dispersion instance; cases without coalitions were excluded). As normality could not be assumed, the nonparametric Friedman test was applied. The Friedman test showed no statistically significant differences among the seven approaches: $\chi_F^2(6, N = 48) = 7.443, \quad p = 0.282$. The associated effect size (Kendall's W) was negligible: $W = \frac{\chi_F^2}{N(k-1)} = \frac{7.443}{48\times 6} \approx 0.026$. Therefore, no post-hoc pairwise tests were conducted. The descriptive distribution of the F-measure values is shown in Fig. 2. The baseline method exhibits the highest median but also the

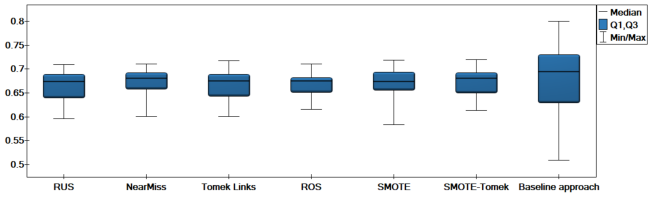

Fig. 2. Comparison of F-measure for all seven approaches (box = Q1–Q3, line = median, whiskers = min/max).

largest variability, while rebalancing methods form a compact group with similar central tendencies and narrower interquartile ranges.

5 Conclusions

This paper extends the authors' previously proposed distributed classification framework, based on conflict analysis, coalition formation, and decision rule induction, by incorporating a local class balancing stage. Six data-level balancing techniques were evaluated on two benchmark datasets, Car Evaluation and Balance Scale.

The results indicate that class balancing improves performance with respect to imbalance-sensitive measures under severe class imbalance, as observed for the Car Evaluation dataset. For moderately imbalanced data (Balance Scale), differences between configurations remain less pronounced. The computational cost of the framework is dominated by coalition formation and may be exponential in the worst case; however, no significant performance limitations were observed in practice.

From a practical perspective, the proposed approach may support decision-making in domains such as healthcare or finance, where data are distributed and imbalanced, and where reliable identification of rare cases and transparent reasoning are required. Future research will focus on extending the framework to

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_33

environments with partially different attribute sets and on incorporating feature selection mechanisms.

References

1. Batista, G. E. A. P. A., Prati, R. C., Monard, M. C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter **6**(1), 20–29 (2004).
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)
3. Chen, Z., Duan, J., Kang, L., Qiu, G.: A hybrid data-level ensemble to enable learning from highly imbalanced dataset. Information Sciences **554**, 157–176 (2021)
4. Dua, D., Graff, C.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA, USA (2019)
5. Grzyb, J., Woźniak, M.: SVM ensemble training for imbalanced data classification using multi-objective optimization techniques. Applied Intelligence **53**(12), 15424–15441 (2023)
6. Koziarski, M., Woźniak, M.: Local neighborhood encodings for imbalanced data classification. Machine Learning **113**(10), 7421–7449 (2024)
7. Mani, I., Zhang, I.: kNN approach to unbalanced data distributions: a case study involving information extraction. In: Proceedings of Workshop on Learning from Imbalanced Datasets, pp. 1–7. ICML, United States (2003)
8. Pawlak, Z.: An inquiry into anatomy of conflicts. Information Sciences **109**, 65–78 (1998)
9. Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M.P., Shyu, M.L., Chen, S.C., Iyengar, S.S.: A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys **51**(5), 1–36 (2018)
10. Przybyła-Kasperek, M.: Study of selected methods for balancing independent data sets in k-nearest neighbors classifiers with Pawlak conflict analysis. Applied Soft Computing **129**, 109612 (2022)
11. Przybyła-Kasperek, M., Kusztal, K.: Integrating Conflict Analysis and Rule-Based Systems for Dispersed Data Classification. In: Paszynski, M., Barnard, A.S., Zhang, Y.J. (eds.) Computational Science – ICCS 2025 Workshops. ICCS 2025. Lecture Notes in Computer Science, vol. 15910. Springer, Cham (2025).
12. Rezvani, S., Wang, X.: A broad review on class imbalance learning techniques. Applied Soft Computing **143**, 110415 (2023)
13. Shenoy, D., Bhat, R., Krishna Prakasha, K.: Exploring privacy mechanisms and metrics in federated learning. Artificial Intelligence Review **58**(8), 223 (2025)
14. Thabtah, F., Hammoud, S., Kamalov, F., Gonsalves, A.: Data imbalance in classification: Experimental evaluation. Information Sciences **513**, 429–441 (2020)
15. Tomek, I.: Two modifications of CNN. IEEE Transactions on Systems, Man and Cybernetics **6**(11), 769–772 (1976)
16. Widodo, A. O., Setiawan, B., Indraswari, R.: Machine learning-based intrusion detection on multi-class imbalanced dataset using SMOTE. Procedia Computer Science **234**, 578–583 (2024)
17. Yao, Y.: Three-way decision and granular computing. International Journal of Approximate Reasoning **103**, 107–123 (2018)
18. Zhou, Z. H.: Ensemble learning. In: Machine Learning, pp. 181–210. Springer, Singapore (2021)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_33