Knowl Inf Syst
DOI 10.1007/s10115-017-1126-1

REGULAR PAPER

Dynamic affinity-based classification of multi-class imbalanced data with one-versus-one decomposition: a fuzzy rough set approach

Sarah Vluymans[1,2,3] · **Alberto Fernández**[3] · **Yvan Saeys**[1,2] · **Chris Cornelis**[1,3] · **Francisco Herrera**[3,4]

Received: 11 October 2016 / Revised: 8 September 2017 / Accepted: 10 October 2017

Abstract Class imbalance occurs when data elements are unevenly distributed among classes, which poses a challenge for classifiers. The core focus of the research community has been on binary-class imbalance, although there is a recent trend toward the general case of multi-class imbalanced data. The IFROWANN method, a classifier based on fuzzy rough set theory, stands out for its performance in two-class imbalanced problems. In this paper, we consider its extension to multi-class data by combining it with one-versus-one decomposition. The latter transforms a multi-class problem into two-class sub-problems. Binary classifiers are applied to these sub-problems, after which their outcomes are aggregated into one prediction. We enhance the integration of IFROWANN in the decomposition scheme in two steps. Firstly, we propose an adaptive weight setting for the binary classifier, addressing the varying characteristics of the sub-problems. We call this modified classifier IFROWANN-$\mathcal{W}_{\mathrm{IR}}$. Second, we develop a new dynamic aggregation method called WV–FROST that combines the predictions of the binary classifiers with the global class affinity before making a final decision. In a meticulous experimental study, we show that our complete proposal outperforms the state-of-the-art on a wide range of multi-class imbalanced datasets.

Keywords Imbalanced data · Multi-class classification · One-versus-one · Fuzzy rough set theory

✉ Sarah Vluymans
sarah.vluymans@ugent.be

1 Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

2 Data Mining and Modeling for Biomedicine, VIB Inflammation Research Center, Ghent, Belgium

3 Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

4 Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Published online: 22 October 2017

1 Introduction

This paper focuses on the challenge of class imbalance for classification problems, which occurs when the elements of a dataset are unevenly distributed among the classes. Such class imbalance poses a challenge to traditional classifiers, such that specific methods able to deal with the imbalance need to be employed instead [31,55]. In a two-class scenario, the imbalance ratio (IR), the ratio between the majority and minority class examples [48], is used to identify this type of datasets. The most straightforward cause of the performance degradation is the misclassification of minority class examples. The recognition of these examples can be ignored in favor of majority instances when considering two common criteria: the maximization of both accuracy and model generalization. With regard to the former, a good classification performance on majority classes can easily overshadow a very poor recognition of minority instances. For the latter, the regions of the problem space with few minority examples can possibly be discarded in the learning process. Recent studies have also shown that the problem of imbalanced classes usually occurs in combination with various intrinsic characteristics of the data [42] that impose additional learning restrictions. Among them, we stress the overlap between classes [2,27] and the presence of small disjuncts and noisy data [54].

In this paper, we consider the general problem of multi-class imbalance, while many previous works have been limited to the binary imbalanced case. Multi-class imbalanced data is encountered in real-life applications, like microarray research [68], protein classification [71], medical diagnosis [7], activity recognition [24], target detection [52] and video mining [25].

When aiming to solve any classification problem, it is clear that the higher the number of classes, the harder it becomes to correctly determine the output label for a query instance. This is mainly due to the overlap between the different classes in the dataset, which increases as more classes are inter-related. One simple yet effective way to address this task is to apply a divide-and-conquer methodology. Such methods are known as decomposition strategies [44], in which the original problem is divided into several easier-to-solve binary subsets. A different classifier is devoted to distinguish among each pair of classes, and then, in the testing phase, the outputs of all classifiers are aggregated to make the final decision [20]. The difficulty in addressing the multi-class problem is therefore shifted from the classifier itself to the combination stage. Among the proposed decomposition strategies, the one-versus-one (OVO) setting has been shown to outperform the one-versus-all (OVA) setting for imbalanced data (e.g. [16]). One problem related to decomposition schemes is the question of classifier competence [19]. In the OVO setting, this issue refers to the fact that the outputs of all classifiers are equally taken into account when extracting a final prediction, although some of them were not trained to discern the real class of the instance and will usually not provide any relevant information. This can hinder the prediction performance. This phenomenon should be considered when developing a method based on OVO decomposition.

The work of [51] proposed a powerful classifier for two-class imbalanced data based on fuzzy rough set theory [12], a mathematical theory that allows to model vagueness and indiscernibility in data. This method was called IFROWANN and was shown to outperform other state-of-the-art methods. Its limitation is that it was set up as a binary classifier, and it cannot directly deal with more than two classes.

In this work, we propose the extension of IFROWANN to the multi-class setting. To successfully classify multi-class imbalanced datasets, a binarization step is considered. We use IFROWANN within an OVO setting, proposing two new components:

- **IFROWANN-$\mathcal{W}_{\text{IR}}$**: the fuzzy rough component of the IFROWANN method requires the specification of a weighting scheme. The original study in [51] showed that the optimal choice depends on the IR of the two-class problem under consideration. In an OVO decomposition, the IR can greatly differ among the binary sub-problems. We therefore propose an adaptive version of IFROWANN, called IFROWANN-$\mathcal{W}_{\text{IR}}$, that dynamically chooses its weight settings based on the IR of each binary problem at hand. We demonstrate the necessity of an adaptive weight choice in our experiments.
- **WV–FROST**: the second original contribution (and main novelty) of this paper is a new approach to deal with the classifier competence issue in an OVO ensemble. Each classifier in the OVO decomposition provides local information, that is, it only discerns between two possible classes. The reduction to two classes results in a loss of information. This is somewhat counteracted by aggregating over all classifiers to obtain a final prediction, as done in existing OVO aggregation schemes. We propose a further performance enhancement by explicitly including two global measures in the decision procedure. In this way, we aim to optimally use all information contained in the dataset. Both global summary terms are based on fuzzy rough set theory, as the binary classifiers are. When classifying an instance, the summary terms evaluate its global affinity with all candidate classes, complementing the local information provided by the OVO classifiers. Our new aggregation method is called weighted voting with fuzzy rough summary terms (WV–FROST).

The use of fuzzy rough set theory for multi-class imbalanced classification is motivated primarily by the excellent performance of the IFROWANN method from [51] in two-class imbalanced problems. The fuzzy rough paradigm has also been used to preprocess such datasets in order to facilitate their classification [56]. The limitation of IFROWANN, as stated above, is that it cannot be directly applied in multi-class problems. We therefore use it within the OVO decomposition scheme. However, the decomposition step poses some challenges that should be appropriately dealt with. The importance of the classifier competence problem has been demonstrated in the works of [21,22], which proposed dynamic classifier selection schemes in conjunction with OVO decomposition. These methods are based on the presence of classes in the local neighborhood of a query instance. In WV–FROST, we replace this local evaluation by two global fuzzy rough summary terms, as local information is already provided by the decomposition process. As we will show, fuzzy rough set theory is ideally suited to capture the global information contained in the dataset. Furthermore, since both steps are based on fuzzy rough set theory, the synergy between IFROWANN in the OVO decomposition on the one hand and WV–FROST on the other allows their combination to show a superior behavior. Our complete novel methodology, the combination of IFROWANN-$\mathcal{W}_{\text{IR}}$ in the OVO decomposition and the WV–FROST aggregation, is referred to as FROVOCO, which stands for Fuzzy Rough OVO COmbination. FROVOCO is a full and novel method, which can be used directly in the classification of multi-class imbalanced data.

We use 18 datasets from various application domains in our experiments. In a first stage, we demonstrate the advantage of our adaptive weighting scheme for IFROWANN in the OVO setting. Secondly, we show that our affinity-based design WV–FROST outperforms the earlier dynamic approaches from [21,22] as well as partially constructed models using no binarization step or only one of the two summary terms. Finally, our complete method FROVOCO is experimentally shown to outperform the state-of-the-art in multi-class imbalanced classification, in particular, C4.5-OVO combined with preprocessing [16], AdaBoost.NC [61] and C4.5 with Mahalanobis Distance Oversampling (MDO) [1].

The remainder of this paper is organized as follows. In Sect. 2, we recall the proposed solutions for the classification of imbalanced data in both the binary and multi-class settings as well as the OVO decomposition scheme and its existing traditional and dynamic aggregation methods. Section 3 describes the original IFROWANN method from [51] and provides the necessary background on fuzzy rough set theory. The original contributions of this work, IFROWANN-$\mathcal{W}_{\text{IR}}$ and WV–FROST, are presented in Sect. 4. Our proposal is carefully evaluated in Sects. 5–7 and shown to outperform the state-of-the-art. Finally, our conclusions and future work are formulated in Sect. 8.

2 Classification approaches for imbalanced data

Despite showing a fairly common occurrence and a strong impact on applications, the problem of imbalanced classes has not been solved properly by machine learning algorithms. Indeed, those methods that perform well in standard classification problems do not necessarily achieve the best performance for imbalanced datasets [15]. The main issue is that they consider equal distributions among classes or the same cost ranking for all classes.

Traditionally, the focus of class imbalance research has been on binary problems, where one class is considerably larger than the other (Sect. 2.1). However, datasets with more than two classes can be imbalanced as well and the attention of the research community has shifted to this more general setting in recent years. This paper focuses on classification problems with more than two classes, for which we apply the OVO decomposition scheme [30]. This strategy and its aggregation mechanisms are described in Sect. 2.2, including some remarks on the dynamic classifier selection procedure for the OVO scheme. Finally, Sect. 2.3 discusses some relevant solutions to deal with multi-class imbalanced data.

2.1 Binary-class imbalance

When the goal is to boost the global performance on both the minority and majority classes, special mechanisms must be applied together with the classifiers. The procedures to address imbalanced classification in two-class problems can be categorized into three groups [40,42]: data level solutions that rebalance the training set [4], algorithmic level solutions that adapt the learning stage toward the minority classes [3] and cost-sensitive solutions which consider different misclassification costs with respect to the class distribution [11].

Among these methodologies, the advantage of the data level solutions is that they are more versatile, since their use is independent of the selected classifier. Three possible schemes can be applied: undersampling of the majority class, oversampling of the minority class and combinations of these two. The simplest approach, random undersampling, removes instances from the majority class until the class distribution is more balanced. A downside is that important majority class examples may be ignored. The random oversampling alternative makes exact copies of existing minority instances. The drawback here is that this method can increase the likelihood of overfitting [4].

More sophisticated algorithms have been proposed based on the generation of synthetic samples, often inspired by the SMOTE oversampling method [6]. The core idea is to form new minority class examples by interpolating between several minority class examples that lie close together. This allows to expand the clusters of the minority class and to strengthen the borderline areas between classes.

2.2 The one-versus-one scheme (OVO)

The use of decomposition strategies in multi-class classification is of great interest in the research community [20,44]. This scheme simplifies the original problem into binary-class subsets, following a divide-and-conquer paradigm. Evidently, boundaries between two classes are easier to learn than in the general case, where they are more likely to highly overlap. Therefore, the critical step is moved toward the decision process, in which the confidence degrees of all binary classifiers must be aggregated in order to output a single class.

In the OVO strategy, an m-class problem is divided into $m(m-1)/2$ two-class problems, one for each pair of classes. Each binary classification sub-problem is addressed by a different classifier, which is built using training instances only from the two classes under consideration. An easy way of organizing the outputs of the base classifiers for an instance x is by means of a score-matrix $R(x)$, given by

$$R(x) = \begin{pmatrix} - & r_{12} & \cdots & r_{1m} \\ r_{21} & - & \cdots & r_{2m} \\ \vdots & & & \vdots \\ r_{m1} & r_{m2} & \cdots & - \end{pmatrix},$$

where $r_{ij} \in [0, 1]$ represents the confidence that x belongs to class i and not class j, obtained by the binary classifier trained on these two classes. The confidence in favor of class j is set to $r_{ji} = 1 - r_{ij}$, if the classifier does not provide it.[1]

Several combination strategies to derive a class prediction from $R(x)$ have been proposed in the literature. Two of the most intuitive aggregation methods are the simple voting strategy (VOTE) proposed in [17] and the weighted voting strategy (WV) from [35]. In the former setting, each binary classifier casts a vote for its predicted class. The class that receives the most votes is predicted. In the weighted alternative, each classifier assigns a confidence to the two classes that it handles. The class with the largest total confidence is the final prediction. More advanced methods are pairwise coupling (PC, [30]), a decision directed acyclic graph (DDAG, [50]), learning valued preference for classification (LVPC, [33,34]), the non-dominance criterion (ND, [14]), a binary tree of classifiers (BTC, [13]), nesting OVO (NEST, [38,39]) and probability estimates by pairwise coupling (PE, [65]). We refer the reader to the review in [20] for clear descriptions of these methods.

Two more recent aggregation techniques consist of a distance-based adaptation of the score matrix. In general, when constructing an ensemble system, an important step is the selection of its constituent classifiers. One seeks to determine the best subset of models, either statically (by using pruning methods [23,46]) or dynamically for each query instance [5]. For the OVO methodology, this step becomes particularly relevant due to the problem of non-competent classifiers [19], which are those binary classifiers whose learned classes do not match the actual class of the query example. General classifier selection methods are based on competence estimation techniques of standard ensembles, computing the local accuracy of each classifier in order to carry out a dynamic selection [60,64]. This approach does not address the OVO non-competence problem, and the adaptation of dynamic classifier selection to this setting is not straightforward. One solution is to take the neighborhood of the query instance into account and adapt the score matrix in order to diminish the impact of those classifiers of which the two classes are not sufficiently relevant based on the neighborhood information. Two different approaches have been proposed:

[1] If the classifier provides both confidence degrees, one must ensure that they are normalized such that $r_{ij} + r_{ji} = 1$.

– Dynamic-OVO (DynOVO) [21]. The WV scheme is used as a basis. Prior to its computation, the score matrix is filtered by considering only those classifiers whose classes are in the neighborhood of the query instance. A neighborhood of size $3 \cdot m$ is used.
– Distance relative competence weighted approach (DRCW) [22]. This methodology consists of carrying out a dynamic adaptation of the score matrix. It alters the confidence degrees by assigning a higher weight to those classifiers whose predicted classes are in the neighborhood of the query instance, setting the final prediction to $\arg\max_{i=1,\ldots,m} \sum_{1 \leq j \neq i \leq m} (r_{ij} \cdot w_{ij})$, where w_{ij} is the relative classifier competence computed as $w_{ij} = \frac{d_j^2}{d_i^2 + d_j^2}$, with d_i the average distance of the k neighbors of the i-th class to the query instance.

2.3 Multi-class imbalance

The scenario of imbalanced classification with more than two classes imposes a strong restriction on the correct recognition of the different concepts present in the data [16]. This is not only due to the larger number of boundaries to consider. All data characteristics that must be considered in the scenario of classification with binary imbalanced data [42] are further accentuated when working in a context with more than two classes as well. The occurrence of multi-minority and multi-majority classes, the dependency among these classes (including overlapping) and relations between same-class examples are possibly the main causes of performance degradation in this case.

The three solution groups listed in the Sect. 2.1 are designed for two-class problems, and their extension to the multi-class scenario is not straightforward. On the one hand, data level solutions (preprocessing) are not directly applicable as the search space is increased. On the other hand, algorithmic level solutions become more complicated since there can be more than one minority class. Several alternatives have been developed to address this task [16]. We emphasize three different schemes: two approaches acting at the data level, namely OVO with preprocessing [16] and MDO [1], and a third one considering the use of ensembles for multi-class imbalanced learning, like the AdaBoost.NC method [61].

OVO combined with preprocessing Binarization techniques are very useful in overcoming the gap between two-class and multi-class imbalanced datasets. They allow the application of any of the standard solutions and particularly those that rebalance the training set. These methods are composed of three simple steps [16]:

1. First, the original multi-class problem is divided into simpler binary sub-problems by means of a decomposition strategy [20], for instance with the OVO scheme. In this way, the skewed class distribution is somehow controlled, as the sizes of two given classes can be similar.
2. Then, for each sub-problem, any technique for instance preprocessing in two-class imbalanced datasets can be applied. In this paper, we use the SMOTE method from [6]. After this step, every binary dataset is processed by a classifier. Recall that in the learning stage only instances from the two classes that the classifier is responsible for are taken into account.
3. Finally, when a new instance is presented, every individual classifier is fired to provide the confidence degrees for the two classes it is responsible for. These values are then aggregated using one of the schemes discussed in Sect. 2.2.

Mahalanobis distance oversampling (MDO) This method was proposed in the recent contribution of [1]. It is a novel oversampling method for multi-class imbalanced problems and is inspired by the Mahalanobis distance [45]. The core idea is to not generate synthetic minority instances at random, but rather to guarantee that an artificial instance has the same Mahalanobis distance to its class mean as the seed element from which it was constructed. We note that there is a small error in the description of this method, which makes the implementation presented in [1] invalid. We have fixed this by slightly modifying Algorithm 3 from [1]. Instead of choosing the value r from the interval $\left[-\sqrt{AlphaV(j)}, \sqrt{AlphaV(j)}\right]$, we divide both boundaries by $(|\mathcal{A}| - 1)$, where $\mathcal{A}$ is the feature set. This fix is required to guarantee that a solution can be found in line 14. Leaving Algorithm 3 as it was presented in [1] results in failures of the method.

AdaBoost.NC ensemble The binary version of this method was proposed in [62] and incorporates negative correlation learning. It is based on the AdaBoost algorithm and extends it by introducing diversity between the constructed classifiers. The instance weights are not solely used to better recognize misclassified elements in later iterations, but also to enhance the diversity. AdaBoost.NC was extended to handle more than two classes in [61]. The authors noted that the application of random oversampling is required to improve the recognition of minority instances. To avoid increasing the training time, we incorporate this instruction in our experiments by a modified initialization of the ensemble weights, in order to give a higher significance to smaller classes. AdaBoost.NC is an important standard to measure the performance of new methods in multi-class imbalanced learning against. We note that the use of ensembles for multi-class imbalanced learning has been evaluated in [28,67] as well, albeit in conjunction with feature selection. In the study of [70], ensembles for binary imbalanced classification were used within the OVO decomposition, showing competitive results with AdaBoost.NC.

3 The IFROWANN algorithm

In this second preliminary section, we recall the original IFROWANN classification method, the classifier using fuzzy rough sets for binary imbalanced classification. We provide the necessary background on fuzzy rough set theory (Sect. 3.1) and the classification model itself (Sect. 3.2).

3.1 Fuzzy rough set theory

Fuzzy rough set theory [12] is a mathematical tool dealing with two distinct types of uncertainty in data, namely vagueness and incompleteness. It was developed by integrating fuzzy set theory [69] into rough set theory [49]. Rough sets approximate a concept C described by an incomplete feature set in two ways. The *lower approximation* contains elements certainly belonging to C, while the *upper approximation* consists of elements possibly belonging to it. When these two sets are equal, there is no uncertainty in the data. In every other case, C cannot be described conclusively based on the observed features and can only be approximated. A limitation of rough set theory is that it requires discretization of any real-valued features in order to obtain useful results. The extension to fuzzy rough set theory addresses this issue. By measuring similarity between instances with fuzzy relations, the discretization requirement is removed. Fuzzy rough set theory has been used in many machine learning applications, see [58] for a recent review.

The fuzzy rough lower approximation of a concept C is a fuzzy set, defined as

$$\underline{C}(x) = \min_{y \in T} [\mathcal{I}(R(x, y), C(y))], \tag{1}$$

where T is the training set and the $R(\cdot, \cdot)$ relation expresses the similarity or indiscernibility between elements. The values $C(\cdot)$ represent the membership degree of the training instances to the concept. In this paper, where C corresponds to a decision class, $C(y)$ can take on only two values: 0 (element y not in class C) and 1 (element y in class C). Finally, $\mathcal{I}$ is a fuzzy logic operator called an implicator. This is a $[0, 1]^2 \rightarrow [0, 1]$ mapping decreasing in its first argument, increasing in its second and satisfying the boundary conditions $\mathcal{I}(0, 0) = \mathcal{I}(0, 1) = \mathcal{I}(1, 1) = 1$ and $\mathcal{I}(1, 0) = 0$. Since $C(y)$ can only take on values 0 and 1, derivations show that, using popular choices for $\mathcal{I}$ like the Kleene-Dienes implicator ($\mathcal{I}(a, b) = \max(1 - a, b)$) or the Łukasiewicz implicator ($\mathcal{I}(a, b) = \min(1 - a + b, 1)$), expression (1) can be simplified to

$$\underline{C}(x) = \min_{y \notin C} (1 - R(x, y)). \tag{2}$$

The membership of x to the lower approximation of C is therefore given by the complement to 1 of its similarity with the most similar instance not in C.

The membership to the fuzzy rough upper approximation of C is given as

$$\overline{C}(x) = \max_{y \in T} [\mathcal{T}(R(x, y), C(y))], \tag{3}$$

where the fuzzy logic operator $\mathcal{T}$ is a triangular norm (t-norm), a commutative and associative $[0, 1]^2 \rightarrow [0, 1]$ mapping increasing in both arguments and satisfying the boundary condition $(\forall a)(\mathcal{T}(a, 1) = a)$. As above, since $C(y)$ only takes on 0–1 values, (3) can be shown to simplify to

$$\overline{C}(x) = \max_{y \in C} R(x, y) \tag{4}$$

and is therefore given as the highest similarity of x with a training instance belonging to C.

Both (2) and (4) show that the membership degree of an instance to the fuzzy rough approximations of decision classes are determined based on their similarity with a single training instance. These procedures are therefore highly susceptible to noise. Several noise-tolerant versions of fuzzy rough sets have been proposed in the literature [9]. We recall the OWA-based fuzzy rough sets from [8]. This model replaces the minimum and maximum operators in (1) and (3) by ordered weighted average (OWA, [66]) aggregations. An OWA aggregation of a set of values $V = \{v_1, \ldots, v_n\}$ uses a weight vector $W = \langle w_1, \ldots w_n \rangle$ of which all elements are drawn from $[0, 1]$ and sum to 1 in total. These weights are assigned to elements in V based on their position in an ordered sequence. Concretely, two steps are performed:

1. Sort the elements in V in descending order. Let $S = \langle s_1, \ldots s_n \rangle$ be this sorted sequence, where s_i is the ith largest value in V.
2. Compute the OWA aggregation of V as $\text{OWA}^W(V) = \sum_{i=1}^{n} w_i s_i$.

The OWA-based fuzzy rough set model proposed in [8] uses appropriate weight vectors W_l and W_u for the lower and upper approximations that soften the minimum and maximum, respectively. It replaces (1) and (3) by

$$\underline{C}(x) = \underset{y \in T}{\text{OWA}^{W_l}} [\mathcal{I}(R(x, y), C(y))] \tag{5}$$

and

$$\overline{C}(x) = \underset{y \in T}{\text{OWA}}^{W_u}[\mathcal{T}(R(x, y), C(y))]. \tag{6}$$

We note that to measure the similarity between two instances x and y in this study, we use the fuzzy relation $R(\cdot, \cdot)$ defined as

$$R(x, y) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} R_a(x, y),$$

where $\mathcal{A}$ is the feature set. The function $R_a(\cdot, \cdot)$ expresses the feature-wise similarity between instances, calculated as

$$R_a(x, y) = 1 - \frac{|a(x) - a(y)|}{\text{range}(a)}$$

when a is a numeric feature and as

$$R_a(x, y) = \begin{cases} 1 & \text{if } \ a(x) = a(y) \\ 0 & \text{otherwise,} \end{cases}$$

when a is nominal. This relation has previously been used in various other works, including [51], on which we base our proposal.

3.2 The classification model

In [51], a classification method for two-class imbalanced data based on fuzzy rough set theory was proposed. It is an extension of the fuzzy rough nearest neighbor classifier (FRNN, [36]), modified to deal with class imbalance. To classify an instance x, FRNN computes its membership degree to the fuzzy rough lower and upper approximations of each class. The score for a class is set to the average membership degree of x to its approximations. The membership degrees are computed by expressions (1) and (3).

The proposal of [51] was called IFROWANN and is developed to deal with class imbalance in two-class datasets. With only two possible class labels, it was shown that the upper approximation of classes does not carry any additional information on top of that represented by the lower approximation. Consequently, only the membership to the lower approximations of the two classes is computed in the classification step and the class for which this value is largest is used as prediction. The OWA-based alternative (5) is used. The distinction between minority and majority classes is incorporated in a class-dependent weight selection for the OWA aggregation, that is, the definition of the weight vector W_l can be different for the two classes. Our definition of these weights differs slightly from the description given in the original proposal, but is equivalent and somewhat clearer to interpret. In particular, we base this description on the integration of an OWA operator in the simplified form (2) rather than (5). The lower approximations of the positive (minority) and negative (majority) classes P and N, using the weight vectors W_l^+ and W_l^-, are, respectively, given by

$$\underline{P}(x) = \underset{y \notin P}{\text{OWA}}^{W_l^+}(1 - R(x, y)) = \underset{y \in N}{\text{OWA}}^{W_l^+}(1 - R(x, y)) \tag{7}$$

and

$$\underline{N}(x) = \underset{y \notin N}{\text{OWA}}^{W_l^-}(1 - R(x, y)) = \underset{y \in P}{\text{OWA}}^{W_l^-}(1 - R(x, y)). \tag{8}$$

Six different weighting schemes were proposed in [51], of which two were shown to yield the best results in an extensive experimental study. In this paper, we denote them as $\mathcal{W}_e$ and $\mathcal{W}_\gamma$. They, respectively, correspond to schemes $\mathcal{W}_4$ and $\mathcal{W}_6$ of [51]. Version $\mathcal{W}_e$ uses the same weight definition for W_l^+ and W_l^-, namely that of exponential weights, which, for a vector of length n, are defined as

$$W = \left\langle \frac{1}{2^n - 1}, \frac{2}{2^n - 1}, \ldots, \frac{2^{n-2}}{2^n - 1}, \frac{2^{n-1}}{2^n - 1} \right\rangle.$$

Although $\mathcal{W}_e$ uses the same general definition for W_l^+ and W_l^-, the actual vectors can be very different, since their lengths can differ greatly. Indeed, note that the aggregations in (7) and (8) are over sets of sizes $|N|$ and $|P|$, respectively. For imbalanced datasets there can be a large difference in these values. The second scheme $\mathcal{W}_\gamma$ does use different weight definitions for W_l^+ and W_l^-. For W_l^-, the exponential weights remain in place. For W_l^+, a different approach is taken by actively removing elements from the aggregation (7). Instead of using the contribution $1 - R(x, y)$ for all instances $y \in N$, only the smallest values are used. These correspond to the instances $y \in N$ most similar to x. Their number is set to $r = \lceil |P| + \gamma(|N| - |P|) \rceil$, with $\gamma \in [0, 1]$ a user-defined parameter. This step is achieved by setting the first $|N| - r$ positions of the vector to 0, as they correspond to the highest values. The weight vector is therefore given by

$$W_l^+ = \left\langle 0, \ldots, 0, \frac{2}{r(r+1)}, \frac{4}{r(r+1)}, \ldots, \frac{2(r-1)}{r(r+1)}, \frac{2}{r+1} \right\rangle. \tag{9}$$

The remaining r values are assigned linear increasing weights as opposed to exponential weights, in order to ensure that they have a more balanced contribution. More details and motivation can be found in [51]. With respect to the parameter γ, the value 0.1 was proposed and experimentally shown to perform better than alternative values for this parameter. The related study [59] also confirmed that setting γ to 0.1 is a good choice.

4 FROVOCO: novel algorithm for multi-class imbalanced problems

In this section, we describe our proposal for multi-class imbalanced classification using fuzzy rough set theory:

- In Sect. 4.1, we describe our IFROWANN-$\mathcal{W}_{\text{IR}}$ binary classifier. We propose a new adaptive weighting scheme that selects appropriate weights depending on the IR of the pair of classes at hand.
- In Sect. 4.2, we describe our new OVO aggregation scheme: weighted voting with fuzzy rough summary terms (WV–FROST). We introduce two global summary terms, measuring the affinity of a test instance with the possible classes in two ways. We combine these terms with the WV aggregation scheme to enhance the performance of the latter. WV–FROST deals with the classifier competence issue in an OVO decomposition in a global way.
- As a summary, we present a flowchart of our full proposal FROVOCO in Sect. 4.3.

4.1 Binary classifier within OVO: IFROWANN-$\mathcal{W}_{\text{IR}}$

In the first stage of our experimental study (Sect. 6.1), we will evaluate the performance of IFROWANN in the OVO schemes discussed in Sect. 2.2. For each pair of classes, we apply

an IFROWANN classifier to discern between them. The smallest class of the two is used as positive class. The original method solely yields class predictions, while the construction of the OVO score matrix requires the output of class confidence scores. To this end, when classifying x and applying IFROWANN to classes C_1 and C_2, we set the score for C_1 to $\frac{\underline{C_1}(x)}{\underline{C_1}(x)+\underline{C_2}(x)}$ and that for C_2 to $\frac{\underline{C_2}(x)}{\underline{C_1}(x)+\underline{C_2}(x)}$.

An important question with respect to IFROWANN regards the choice of the weighting scheme. As indicated above, the original study put forward two good candidate schemes: $\mathcal{W}_e$ and $\mathcal{W}_\gamma$. It was shown that $\mathcal{W}_e$ performs well for mildly imbalanced data with an IR up to 9, while for higher imbalance $\mathcal{W}_\gamma$ obtained better results. An IR of 9 is traditionally used (e.g. [41,57]) as a threshold above which datasets are considered highly imbalanced. In a binarization scheme, the IR can be different for each pair of classes. It may therefore not be prudent to decide on the weighting scheme of IFROWANN beforehand, but rather choose this based on the imbalance between the two classes. In our experiments, we therefore evaluate three separate settings, two where the weighting scheme is the same in all IFROWANN classifiers (either $\mathcal{W}_e$ or $\mathcal{W}_\gamma$) and a third one where $\mathcal{W}_e$ is used when the IR between a pair of classes is at most 9 and $\mathcal{W}_\gamma$ otherwise. We denote the third scheme as $\mathcal{W}_{\text{IR}}$ and the corresponding classifier as IFROWANN-$\mathcal{W}_{\text{IR}}$.

Our experimental results in Sect. 6.1 will show that the third scheme outperforms the other two. The motivation of the threshold of 9 follows from the conclusions of [51]. However, based on the performance of IFROWANN evaluated on 102 two-class imbalanced datasets in the original experimental study, we also construct a visual motivation in Fig. 1. The interested reader may find the full experimental results on the web page http://www.cwi.ugent.be/sarah.php. In the figures, we plot the difference in obtained AUC values by IFROWANN using the weighting scheme $\mathcal{W}_6$ ($\mathcal{W}_\gamma$) or $\mathcal{W}_4$ ($\mathcal{W}_e$) against the IR of the dataset. Figure 1a contains the results of all 102 datasets. The only clear conclusion that can be drawn from this plot is that for very highly imbalanced data (IR above 65), $\mathcal{W}_\gamma$ has the clear advantage. In Fig. 1b we zoom in on Fig. 1a on the section containing datasets with an IR at most 20. We observe that the benefits of $\mathcal{W}_e$ can only be found for the mildly imbalanced datasets. Finally, in Figs. 1c and 1d, we take the group of 88 datasets with IR at most 65 in order to decide on a threshold above which $\mathcal{W}_\gamma$ can be preferred. In these plots, we present averages over consecutive data points in order to obtain a smoother figure. In Fig. 1c, we averaged over 4 observations, meaning that the leftmost point was taken as the average value of the 4 least imbalanced datasets in the study. In Fig. 1d, averages were taken over 8 observations, providing an even smoother visual. From both figures, a threshold of 9 certainly seems appropriate. When the IR of the dataset is higher than 9, $\mathcal{W}_\gamma$ yields better results for IFROWANN than $\mathcal{W}_e$. Otherwise, the latter is preferred. We are aware that the choice of this threshold may still appear artificial. It follows from a tradition in imbalanced learning and the empirical validation referenced and presented above. This threshold seems appropriate for our purposes and by setting it beforehand, we avoid the cost of choosing between $\mathcal{W}_e$ and $\mathcal{W}_\gamma$ based on a validation of the training data.

We want to note that it would also be possible to extend IFROWANN to a multi-class classifier without applying a binarization step. However, due to the definition of the fuzzy rough approximation operators (2–6), this corresponds largely to an OVA aggregation, which has been shown in previous studies (e.g. [16]) to not perform well compared to an OVO setting. Indeed, preliminary results for this evaluation (Sect. 6.2) showed that a multi-class version of IFROWANN does not perform at the same level as the integration of the classifier in binarization schemes. As this path seemed less promising, we do not pursue it further in this paper.

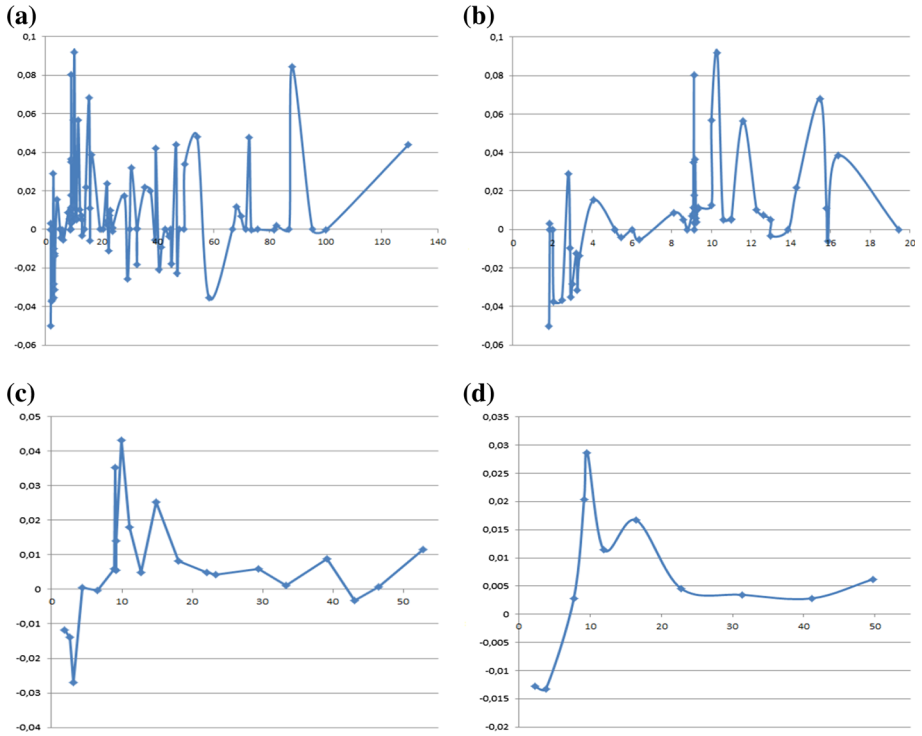

Fig. 1 Motivation for the definition of $\mathcal{W}_{\text{IR}}$. The horizontal axis represents the IR of the dataset and the vertical axis the difference in performance (AUC) between $\mathcal{W}_6$ ($\mathcal{W}_\gamma$) and $\mathcal{W}_4$ ($\mathcal{W}_e$), based on the results on the 102 datasets in [51]. Positive values indicate that $\mathcal{W}_6$ ($\mathcal{W}_\gamma$) performs better than $\mathcal{W}_4$ ($\mathcal{W}_e$). Lines between points were drawn for the sake of visual clarity. **a** 102 datasets (IR: 1.82–129.44), **b** 55 datasets (IR: 1.82–20), **c** 88 datasets (IR: 1.82–65), **d** 88 datasets (IR: 1.82–65)

4.2 New OVO aggregation scheme: WV–FROST

We present a new OVO aggregation scheme, called weighted voting with fuzzy rough summary terms (WV–FROST). It enhances the WV method with a global evaluation represented by two summary terms. This can be regarded as an alternative for traditional dynamic classifier selection models. Whereas those methods only take into account the locality of the input instance, we propose to counteract the information loss resulting from the dataset reduction to class pairs by the inclusion of global measures in the decision procedure. The addition of these terms deals with the problem of classifier competence, as the aggregation not only relies on the binary classification performance, but uses the global information available in the dataset as well.

When classifying an instance x, the WV aggregation method constructs a vector V_x based on the score matrix obtained after the OVO process. For each class C, the position $V_x(C)$ presents the weighted vote for C. This is the sum of the off-diagonal elements on the row corresponding to C in the score matrix. The class with the highest value is selected. Our proposal WV–FROST adds the values of two additional measures to the V_x vector before selecting the best class:

– Positive affinity term: we compute the membership degree of an instance to each class, based on the full training set, by means of the fuzzy rough approximation operators.
– Negative affinity term: each class C can be represented by a vector containing the expected membership degrees of an instance of class C to all classes. For a test instance, such a signature vector can be constructed as well. We penalize the distance from an instance to a class based on these vectors.

Positive affinity For a class C, the value $mem(x, C)$ is the average membership degree of x to the fuzzy rough lower and upper approximations of C, that is,

$$mem(x, C) = \frac{\underline{C}(x) + \overline{C}(x)}{2},$$

using the OWA-based model. The definition of $mem(x, C)$ is based on the decision procedure of the FRNN classifier, on which IFROWANN was inspired. It is important to note that the *mem* values are determined in the full dataset and not based on versions reduced to two classes. This measure is therefore a global evaluation of the class to which x is most likely to belong. To determine the membership degrees $\underline{C}(x)$ and $\overline{C}(x)$, weight vectors related to scheme $\mathcal{W}_{\text{IR}}$ are used. The sets of values to be aggregated for the lower and upper approximations of C are of sizes $|co(C)|$ and $|C|$, respectively, where $co(\cdot)$ denotes the complement of a class. When the IR between C and its complement is at most 9, exponential weights are used for both the lower and upper approximation. In the other case, the shortest weight vector uses exponential weights and the longest weights (9), respectively, replacing P and N by $\min(|C|, |co(C)|)$ and $\max(|C|, |co(C)|)$ in this definition. When adding the *mem* values to the V_x vector, we replace each value $V_x(C)$ with the average $\frac{V_x(C)+mem(x,C)}{2}$. This means that we replace the single local measure $V_x(C)$ by the average of the local and global measures.

Negative affinity The second summary term evaluates the affinity of an instance with a class at a higher level. For every class C, a signature vector S_C can be constructed that consists of the expected membership degrees of an instance of that class to every possible class. In the literature (e.g. [37]), such a signature is also called a decision template. This vector has a length equal to the number of classes and position $S_C(C')$ corresponds to the average membership value of the training instances of class C to class C', calculated in the same way as the *mem* measure. The vector is therefore determined as

$$\begin{aligned} S_C &= \langle S_C(C_1), S_C(C_2), \ldots, S_C(C_m) \rangle \\ &= \left\langle \frac{1}{|C|} \sum_{y \in C} mem(y, C_1), \ldots, \frac{1}{|C|} \sum_{y \in C} mem(y, C_m)) \right\rangle, \end{aligned}$$

where m is the number of classes in the training set. A similar vector S_x can be computed for a test instance x, gathering its *mem* values for all classes. The mean squared error between S_x and each of the vectors S_C expresses to what extent x is similar to the training instances of class C. It is computed as

$$mse(x, C) = \frac{1}{m} \sum_{i=1}^{m} (mem(x, C_i) - S_C(C_i))^2 . \tag{10}$$

This is a negative affinity term, as a higher value implies less similarity between the instance and the class. To motivate the inclusion of this term, consider the following example. Suppose that in a dataset of three classes, classes C_1 and C_2 have a high overlap in feature space. In this

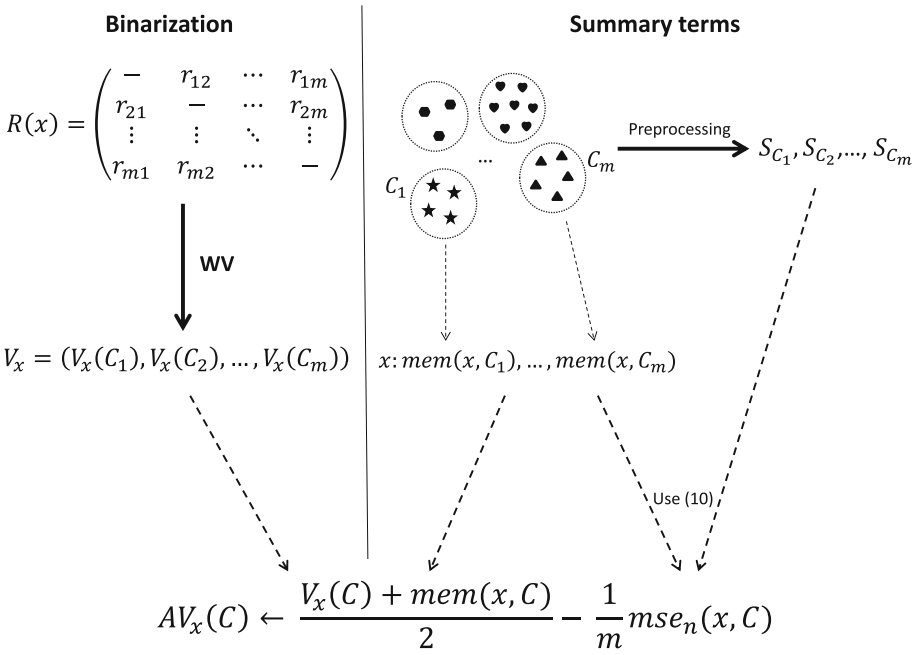

Fig. 2 Overview of the proposed WV–FROST aggregation method

case, the computed values $mem(x, C_1)$ and $mem(x, C_2)$ can be expected to be very similar and therefore not contain sufficient information to make a decision between classes. On the other hand, the signature vectors S_{C_1} and S_{C_2} present the expected membership degrees of instances of these two classes to all classes. In particular, comparing the value $mem(x, C_3)$ to $S_{C_1}(C_3)$ and $S_{C_2}(C_3)$ may provide the final clue in the decision process. We note that, in order to bring the mse values to a similar scale as $V_x(C)$ and $mem(x, C)$, we divide each of them by their sum and define the normalized value as

$$(\forall C)\left(mse_n(x, C) = \frac{mse(x, C)}{\sum_{i=1}^{m} mse(x, C_i)}\right).$$

We use the mem_n values as summary terms. We subtract them from $V_x(C)$ with a weight $\frac{1}{m}$, since we can expect the information in $mse_n(x, C)$ to be less reliable when more classes are present in the dataset. The reason is that when the number of classes increases, the vector S_C becomes longer and the constituent membership degrees more similar, such that the mean squared error is less able to make a distinction between classes.

In conclusion, in our proposed method WV–FROST, for an instance x and a class C, the value $V_x(C)$ obtained from the OVO score matrix by the WV procedure is replaced by the affinity-based value

$$AV_x(C) = \frac{V_x(C) + mem(x, C)}{2} - \frac{1}{m} mse_n(x, C).$$

A visual representation of WV–FROST is provided in Fig. 2. As part of our experimental study, we show that the inclusion of both summary terms outperforms settings where we only include one of the two. The final prediction for an instance x is obtained as

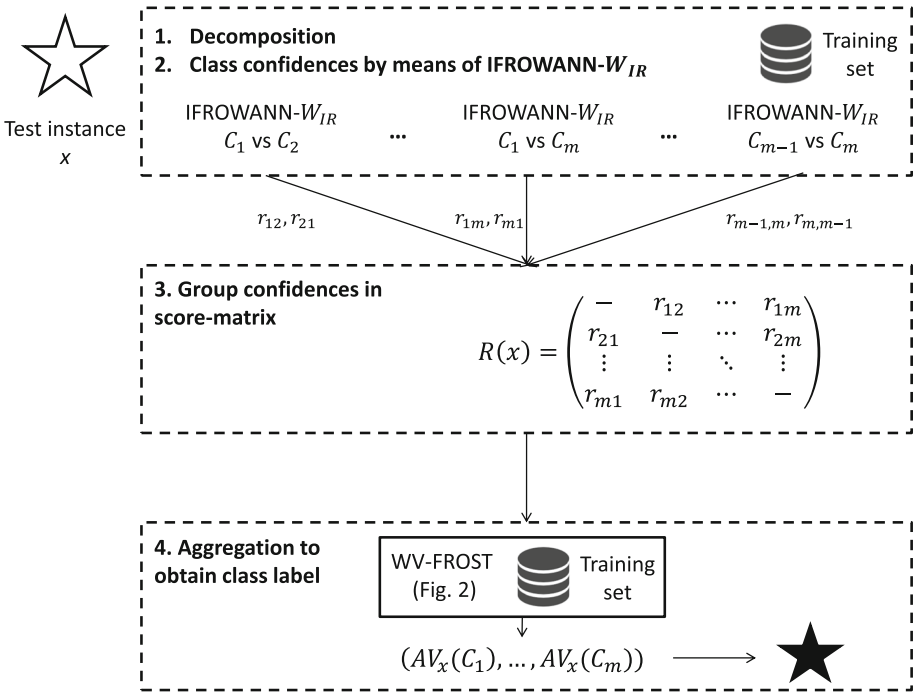

Fig. 3 Flowchart of the complete FROVOCO method

$\arg\max_{i=1,\ldots,m} [AV_x(C_i)]$. By including the global summary terms, a dynamic aggregation in the spirit of the recent methods described in Sect. 2.2 from [21,22] is achieved, although it differs from them in a crucial point. The proposals of [21,22] modify the score matrix of an instance before aggregating it to a prediction value with the WV method. In our approach, we do not modify the matrix, but rather the aggregated values.

4.3 Overview of the FROVOCO proposal

We refer to the combination of our IFROWANN-$\mathcal{W}_{\text{IR}}$ in the OVO setting and our WV–FROST aggregation as FROVOCO (Fuzzy Rough OVO COmbination). Fuzzy rough set theory is used in both the binary classification step and the aggregation phase and the combined benefits of the two is shown in our experimental analysis. For the sake of clarifying the two-stage FROVOCO methodology, Fig. 3 summarizes the actions of this proposal. To classify a test instance x, the following steps are performed:

1. In the decomposition phase, x is send to all IFROWANN-$\mathcal{W}_{\text{IR}}$ methods, each using a pair of classes as training set.
2. Class confidence scores are obtained from the IFROWANN-$\mathcal{W}_{\text{IR}}$ classifiers.
3. These scores are grouped into a score matrix.
4. WV–FROST is applied to aggregate the score matrix to the vector containing the values $AV_x(C)$ for all classes C (Fig. 2). Instance x is assigned to the class corresponding to the largest $AV_x(\cdot)$ value.

5 Experimental setup

In this section, we lay out the details of our study, specifying the datasets, the evaluation measures, the tests used in the statistical analysis and the state-of-the-art methods to which we compare our proposal. The experimental study is conducted in Sects. 6–7.

5.1 Datasets

The 18 datasets used in this study are presented in Table 1. We list the number of instances, features (numeric and nominal) and classes. In the presentation of the experimental results, we use the shorter ID names to refer to the datasets. To give an indication of the imbalance present in a dataset, we list the minimum, average and maximum values of the IR between its class pairs. In all datasets, the minimum IR is low, meaning that there are at least two classes with relatively similar sizes. The maximum IR expresses the highest degree of pairwise imbalance encountered in a dataset. As the table shows, the values for this measure differ greatly across the 18 datasets and are often very high. The average IR is computed as the average of the imbalance ratios in the OVO decomposition. For most datasets, this value is moderate, although a few outliers are present. For the sake of completeness, we also provide the distribution of the instances over the classes (in brackets below the IR information). In our evaluation, we use the tenfold DOB-SCV partitioning scheme from [47] in order to avoid the data-shift problem, the issue of having different distributions in the training and test partitions of the evaluation. In the fold construction of DOB-SCV, regions of same-class elements are divided over different folds to guarantee a proper representation of such a region in all partitions. The use of this partitioning scheme for imbalanced data was advised in [43]. The full datasets and all partitions are available from the KEEL dataset repository (http://www.keel.es) for any interested reader that wishes to repeat our analysis.

5.2 Evaluation measures

To evaluate the performance of the included methods, we use two popular evaluation measures used in multi-class imbalanced classification. The first is the average accuracy (AvgAcc), defined as the average of the class-wise accuracies, that is,

$$\text{AvgAcc} = \frac{1}{m}\left(\sum_{i=1}^{m}\frac{\text{corr}_i}{|C_i|}\right),$$

where corr_i is the number of correctly classified instances of class C_i. This measure is also referred to as balanced accuracy. The second measure is the mean area under the curve (MAUC, [29]), which is the mean of the pairwise AUC values of all pairs of classes, defined as

$$\text{MAUC} = \frac{2}{m(m-1)}\sum_{i<j}\text{AUC}(C_i, C_j) = \frac{2}{m(m-1)}\sum_{i<j}\left(\frac{A(C_i, C_j) + A(C_j, C_i)}{2}\right).$$

For two classes C_i and C_j, the value $\text{AUC}(C_i, C_j)$ represents the probability that a randomly selected element from the first class also has a higher probability of being assigned to that class by the classifier compared to a randomly selected element of the other class ($A(C_i, C_j)$) and vice versa ($A(C_j, C_i)$). We include both measures to capture two different aspects of the

Table 1 The 18 multi-class imbalanced datasets used in our experimental study

Dataset	ID	# inst	# feat	m	Min IR	Av. IR	Max IR
Automobile (3/20/48/46/29/13)	aut	150	25 (15/10)	6	1.04	4.90	16.00
Balance (288/49/288)	bal	625	4 (4/0)	3	1.00	4.25	5.88
Cleveland (164/55/36/35/13)	cle	297	13 (13/0)	5	1.03	3.87	12.62
Contraceptive (629/333/511)	con	1473	9 (9/0)	3	1.23	1.55	1.89
Dermatology (111/60/71/48/48/20)	der	358	34 (34/0)	6	1.00	2.17	5.55
Ecoli (143/77/2/2/35/20/5/52)	eco	336	7 (7/0)	8	1.00	15.27	71.50
Glass (70/76/17/13/9/29)	gla	214	9 (9/0)	6	1.09	3.60	8.44
Led7digit (45/37/51/57/52/52/47/57/53/49)	led	500	7 (7/0)	10	1.00	1.16	1.54
Lymphography (2/81/61/4)	lym	148	18 (3/15)	4	1.33	18.30	40.50
Newthyroid (150/35/30)	new	215	5 (5/0)	3	1.17	3.48	5.00
Pageblocks (4913/329/28/87/115)	pag	5472	10 (10/0)	5	1.32	31.65	175.46
Satimage (1533/703/1358/626/707/1508)	sat	6435	36 (36/0)	6	1.01	1.73	2.45
Shuttle (45,586/49/171/8903/3267/10/13)	shu	58,000	9 (9/0)	7	1.30	561.92	558.60
Thyroid (166/368/6666)	thy	7200	21 (21/0)	3	2.22	20.16	40.16
Wine (59/71/48)	win	178	13 (13/0)	3	1.20	1.30	1.48
Winequality-red (10/53/681/638/199/18)	wqr	1599	11 (11/0)	6	1.07	18.83	68.10
Winequality-white (20/163/1457/2198/880/175/5)	wqw	4898	11 (11/0)	7	1.07	61.08	439.60
Yeast (244/429/463/44/51/163/35/30/20/5)	yea	1484	8 (8/0)	10	1.08	11.65	92.60

The number of features is divided between numeric and nominal ones, e.g. the automobile dataset has 15 numeric features and 10 nominal features

classification behavior of the algorithms. AvgAcc considers the actual prediction output of the method and measures how well it recognizes the different classes, while MAUC expresses the ability of the method to separate pairs of classes, as noted in e.g. [61].

5.3 Statistical analysis

We list average results taken over the group of 18 datasets and, for the final comparison in Sect. 7.2, the results on each individual dataset as well. We combine this with an appropriate statistical analysis, applying non-parametric statistical tests as recommended by e.g. [10,26]. For the comparison between two methods, we use the Wilcoxon signed-ranks test [63]. Its null hypothesis is that the two methods have an equivalent performance. In order to find sufficient evidence to reject the null hypothesis, the absolute differences in results of the two methods are ranked. The smallest absolute difference is assigned rank 1 and the largest rank n, with n the number of observations (18 in our study). In a comparison of 'Method 1 versus Method 2', the positive differences are in favor of the first method, the negative differences in favor of the second. The ranks of the positive differences are summed up to R^+ and those of the negative differences to R^-. We report these two values, together with the p value of the test. When it is smaller than the significance level α, the null hypothesis is rejected and the first method is concluded to perform significantly better than the second. We also perform multiple comparisons, that is, we take a group of methods and determine whether any significant performance differences can be found among them. In this case, we use the Friedman test [18] in combination with the Holm post hoc procedure [32]. The null hypothesis of the Friedman test is that all methods under consideration perform equivalently. When it is rejected, the post hoc procedure is applied to detect where the significant differences can be found. The Friedman test is based on a ranking procedure and the method with the lowest rank is concluded to have the overall best performance. It is used as a control method to which the remaining methods are compared in the Holm process. When the p values of these comparisons are lower than α, it is concluded that the control method outperforms the other method with statistical significance. For these multiple comparisons, we list the Friedman ranks of all methods, the p value of the Friedman test (p_{Friedman}) and the adjusted p values of the post hoc procedure (p_{Holm}).

5.4 Structure of experiments and method parameters

The experimental comparison is divided into two parts, discussed in separate sections:

- Section 6: we first conduct an internal comparison of our proposal, in order to clearly show the strength of the separate components IFROWANN-$\mathcal{W}_{\text{IR}}$ and WV–FROST.
- Section 7: we show the benefits of WV–FROST over other dynamic aggregation methods. Finally, we compare our full method FROVOCO to three state-of-the-art methods in multi-class imbalanced classification recalled in Sect. 2.3. In AdaBoost.NC [61], we have set the penalty strength λ to 2, as done in earlier work e.g. [16,61]. The number of classifiers in the ensemble was set to 10, which is a lower value than the one used by these referenced studies. In a preliminary evaluation, we observed that this value provides better average results on our selected datasets. It has been used in ensembles for imbalanced data in earlier studies as well e.g. [42]. For the OVO combination with preprocessing, we use the SMOTE-C4.5 classifier with the same parameter settings as [16]. The use of decision tree learners like C4.5 in ensembles has been highlighted in [53]. Finally, we include the MDO preprocessing method in combination with C4.5 as well. We do not use any cost-sensitive classification method, as an appropriate definition of the cost matrix is usually not readily available [55] and domain experts are required for its specification [72].

6 Experimental evaluation of IFROWANN-$\mathcal{W}_{\text{IR}}$ and WV–FROST

As a first part of our experiments, we empirically justify our proposed components, the IFROWANN-$\mathcal{W}_{\text{IR}}$ classifier and the WV–FROST aggregation. This section is divided as follows:

- Section 6.1: we evaluate the performance of the IFROWANN method in existing OVO schemes, using the three weighting schemes listed in Sect. 4.1. We clearly show the benefit of our novel IFROWANN-$\mathcal{W}_{\text{IR}}$ method.
- Section 6.2: we compare WV–FROST to related partially constructed aggregation models and show the advantages of the full proposal.

6.1 Evaluation of IFROWANN-$\mathcal{W}_{\text{IR}}$

In this section, we evaluate the classification performance of IFROWANN in an OVO setting. We use the nine traditional aggregation procedures listed in Sect. 2.2, postponing the comparison with the dynamic aggregation methods DynOVO and DRCW to Sect. 7.1. With regard to IFROWANN, we evaluate the three weighting schemes listed in Sect. 4.1: the two original schemes $\mathcal{W}_e$ and $\mathcal{W}_\gamma$ as well as our adaptive proposal $\mathcal{W}_{\text{IR}}$. The latter setting corresponds to our IFROWANN-$\mathcal{W}_{\text{IR}}$ method. The results of these experiments can be found in Table 2. We present the average values of the evaluation measures taken over all datasets, accompanied by their standard deviations.

The benefit of scheme $\mathcal{W}_{\text{IR}}$ over $\mathcal{W}_e$ and $\mathcal{W}_\gamma$ is clear. This is particularly reflected in the average accuracy measure, where substantial differences can be observed. For each aggregation method, $\mathcal{W}_{\text{IR}}$ attains the highest average accuracy. The results of $\mathcal{W}_e$ are better than those for $\mathcal{W}_\gamma$. This can be explained based on the description of the datasets in Table 1 and the conclusions drawn from Fig. 1. Indeed, computing the pairwise IR between classes, these values are often found to be less than 9, a situation where $\mathcal{W}_e$ yields better results than $\mathcal{W}_\gamma$. Considering the MAUC, smaller differences in performance for the three alternatives are observed. In most cases, $\mathcal{W}_e$ still outperforms $\mathcal{W}_\gamma$. For five aggregation methods, $\mathcal{W}_{\text{IR}}$ yields the highest MAUC value. In the cases where it does not, the differences with the best performing scheme are small.

We conclude that, when fixing the weighting scheme, $\mathcal{W}_e$ yields better results than $\mathcal{W}_\gamma$, but using our adaptive scheme $\mathcal{W}_{\text{IR}}$ further improves the performance. This largest improvement is made for the average accuracy measure, showing that correct classifications require the use of $\mathcal{W}_{\text{IR}}$. The power of separating between class pairs, evaluated by the MAUC, is more or less comparable for $\mathcal{W}_e$ and $\mathcal{W}_{\text{IR}}$. Deciding on $\mathcal{W}_{\text{IR}}$ as weighting scheme and taking both evaluation measures into account, we can select the WV procedure as favored aggregation method. It attains the highest average accuracy value, among the highest MAUC values and its robustness has been demonstrated in [20,35]. In the next section, we further improve the results of IFROWANN-$\mathcal{W}_{\text{IR}}$-WV by replacing the WV step by our new proposal WV–FROST.

6.2 Evaluation of WV–FROST

We continue the experimental study in this section with an evaluation of WV–FROST, our new OVO aggregation method. We compare our novel proposal to partially constructed versions. As described in Sect. 4.2, WV–FROST includes two summary terms in the score vector constructed by WV. Two important questions need answering, namely:

1. Does WV–FROST improve the performance of WV?

Table 2 Results of the integration of IFROWANN in the OVO setting with the traditional aggregation methods

Method	AvgAcc		
	$\mathcal{W}_e$	$\mathcal{W}_\gamma$	$\mathcal{W}_{\text{IR}}$
VOTE	69.4460 ± 19.6554	61.6819 ± 20.7889	**70.4035 ± 20.5736**
WV	69.4460 ± 19.6554	63.1538 ± 21.6848	**71.4921 ± 19.5685**
PC	69.4959 ± 19.6182	62.6440 ± 21.0826	**71.3500 ± 19.1160**
DDAG	69.9674 ± 19.6337	59.4896 ± 21.0593	**71.1942 ± 19.9944**
LVPC	58.8251 ± 20.3807	61.2524 ± 20.2090	**63.2167 ± 19.5304**
ND	69.5269 ± 19.5921	58.5540 ± 23.3192	**70.2541 ± 21.2630**
BTC	69.4497 ± 19.7924	59.0936 ± 21.4492	**70.5591 ± 20.5641**
NEST	69.5686 ± 19.6920	56.2790 ± 23.1138	**70.0260 ± 21.2530**
PE	69.4785 ± 19.7172	62.1607 ± 21.3814	**71.1413 ± 19.6598**
Method	MAUC		
	$\mathcal{W}_e$	$\mathcal{W}_\gamma$	$\mathcal{W}_{\text{IR}}$
VOTE	0.8566 ± 0.1227	0.8208 ± 0.1379	**0.8613 ± 0.1204**
WV	**0.8921 ± 0.1120**	0.8910 ± 0.1025	0.8895 ± 0.1120
PC	**0.8958 ± 0.1062**	0.8935 ± 0.1058	0.8939 ± 0.1067
DDAG	0.8070 ± 0.1243	0.7366 ± 0.1384	**0.8143 ± 0.1274**
LVPC	**0.8932 ± 0.1110**	0.8919 ± 0.1043	0.8910 ± 0.1107
ND	0.8779 ± 0.1065	0.8457 ± 0.1229	**0.8794 ± 0.1092**
BTC	0.8035 ± 0.1259	0.7341 ± 0.1413	**0.8105 ± 0.1304**
NEST	0.8572 ± 0.1228	0.8208 ± 0.1380	**0.8613 ± 0.1204**
PE	**0.8952 ± 0.1065**	0.8875 ± 0.1089	0.8927 ± 0.1077

For each method and each evaluation measure, we print the result of the best performing weighting scheme in bold

2. Do partially constructed models yield similar or better results?

We address these questions in Table 3. We include the results of the following models:

- IFROWANN-$\mathcal{W}_{\text{IR}}$-WV: the best performing version in Sect. 6.1, the integration of IFROWANN-$\mathcal{W}_{\text{IR}}$ in a traditional OVO setup with WV aggregation.
- *mem*: no binarization, the score of a class C is set to the value $mem(x, C)$ and the class with the highest score is predicted.
- *mem-mse*$_n$: no binarization, the score of class C is $mem(x, C) - \frac{1}{m} mse_n(x, C)$.
- IFROWANN-$\mathcal{W}_{\text{IR}}$-WV-*mem*: the WV-*mem* step is similar to WV–FROST, but only includes one global summary term. The value $V_x(C)$ is replaced by $\frac{V_x(C)+mem(x,C)}{2}$.
- IFROWANN-$\mathcal{W}_{\text{IR}}$-WV-*mse*$_n$: includes only one global summary term. It replaces $V_x(C)$ by $V_x(C) - \frac{1}{m} mse_n(x, C)$.
- IFROWANN-$\mathcal{W}_{\text{IR}}$-WV–FROST: our complete proposal.

The second and third models have been included to verify whether the summary terms on their own are strong enough in the prediction process and whether the binarization is truly required. Both methods correspond to direct generalizations of the binary IFROWANN classifier to the multi-class setting.

Table 3 Results of IFROWANN-$\mathcal{W}_{\text{IR}}$-WV–FROST and partially constructed versions

Method	AvgAcc	MAUC
IFROWANN-$\mathcal{W}_{\text{IR}}$-WV	71.4921 ± 19.5685	0.8895 ± 0.1120
mem	67.6477 ± 18.8233	0.8810 ± 0.1069
mem-mse$_n$	69.2093 ± 18.3121	0.8958 ± 0.1022
IFROWANN-$\mathcal{W}_{\text{IR}}$-WV-*mem*	71.5351 ± 19.3465	0.8946 ± 0.1065
IFROWANN-$\mathcal{W}_{\text{IR}}$-WV-*mse*$_n$	71.8868 ± 19.2429	0.8984 ± 0.0987
IFROWANN-$\mathcal{W}_{\text{IR}}$-WV–FROST	$\mathbf{72.6327 \pm 19.3379}$	$\mathbf{0.9018 \pm 0.0982}$

Table 3 shows the dominance of IFROWANN-$\mathcal{W}_{\text{IR}}$-WV–FROST over all partially constructed models for both evaluation measures. Considering these results in more detail, IFROWANN-$\mathcal{W}_{\text{IR}}$-WV–FROST outperforms the traditional setting IFROWANN-$\mathcal{W}_{\text{IR}}$-WV on 12 out of the 18 datasets for the average accuracy and on 11 out of 18 for MAUC. Placing the results of models *mem* and IFROWANN-$\mathcal{W}_{\text{IR}}$-WV-*mem* next to each other, the benefit of the binarization step is made clear, in particular in the evaluation by the average accuracy. The relatively high MAUC value of *mem* indicates that the fuzzy rough membership degrees form a good tool to separate between pairs of classes. However, this does not necessarily imply correct classification results, as only pairwise comparisons between classes are used in the MAUC evaluation. This is reflected in the clearly inferior average accuracy value of *mem*. This method corresponds to the most straightforward extension of IFROWANN to a multi-class setting without applying any binarization. By also incorporating the comparison between pairs of classes in IFROWANN-$\mathcal{W}_{\text{IR}}$-WV-*mem*, more accurate predictions can be made. This was already noted in Sect. 4.1 and formed part of our motivation to not further pursue a direct extension of the IFROWANN method without binarization. Secondly, comparing *mem* to *mem-mse*$_n$ and IFROWANN-$\mathcal{W}_{\text{IR}}$-WV-*mem* to IFROWANN-$\mathcal{W}_{\text{IR}}$-WV–FROST, the improvement after including the *mse*$_n$ term is shown. The difference in performance between IFROWANN-$\mathcal{W}_{\text{IR}}$-WV–FROST and IFROWANN-$\mathcal{W}_{\text{IR}}$-WV-*mse*$_n$ shows that it is not sufficient to solely include the *mse*$_n$ measure and that both fuzzy rough summary terms carry complementary information needed to improve the baseline performance of IFROWANN-$\mathcal{W}_{\text{IR}}$-WV. In a statistical comparison of IFROWANN-$\mathcal{W}_{\text{IR}}$-WV–FROST to IFROWANN-$\mathcal{W}_{\text{IR}}$-WV using the Wilcoxon test, the p values for the average accuracy and MAUC results were 0.16736 ($R^+ = 118.0$, $R^- = 53.0$) and 0.08866 ($R^+ = 113.0$, $R^- = 40.0$), respectively.

7 Experimental evaluation of FROVOCO

The second part of our experimental study compares our FROVOCO method to the state-of-the-art in multi-class imbalanced classification. We consider two separate aspects:

– We compare the WV–FROST aggregation within FROVOCO to existing dynamic aggregation approaches. We show that the WV–FROST outperforms the alternatives, which justifies its inclusion in FROVOCO (Sect. 7.1).

Table 4 Results of FROVOCO and the combination of IFROWANN-$\mathcal{W}_{\text{IR}}$ with the two other dynamic aggregation methods

Method	AvgAcc	MAUC
IFROWANN-$\mathcal{W}_{\text{IR}}$+DynOVO	71.7930 ± 19.8270	0.7894 ± 0.1151
IFROWANN-$\mathcal{W}_{\text{IR}}$+DRCW	70.8782 ± 20.1452	0.8916 ± 0.1097
FROVOCO	$\mathbf{72.6327 \pm 19.3379}$	$\mathbf{0.9018 \pm 0.0982}$

Table 5 Pairwise statistical comparisons by means of the Wilcoxon test, accompanying the results of Table 4

Comparison	R^+	R^-	p
A: WV–FROST versus DynOVO	100.0	71.0	0.52773
A: WV–FROST versus DRCW	129.0	42.0	0.05994
M: WV–FROST versus DynOVO	171.0	0.0	7.63E−6
M: WV–FROST versus DRCW	100.0	53.0	0.26595

Lines preceded by 'A' correspond to the evaluation by AvgAcc, while those starting with 'M' are related to the evaluation by MAUC

– As a final step, we compare FROVOCO to three state-of-the-art classifiers for multi-class imbalanced data (Sect. 7.2).

7.1 WV–FROST versus other dynamic approaches

In this section, we compare WV–FROST to the existing dynamic aggregation approaches DynOVO [21] and DRCW [22]. For each of these three methods, we use IFROWANN-$\mathcal{W}_{\text{IR}}$ within the OVO method. The combination of IFROWANN-$\mathcal{W}_{\text{IR}}$ with WV–FROST corresponds to our full FROVOCO method.

The average accuracy and MAUC results are given in Table 4. For both measures, our proposal attains the best results. The statistical comparison is presented in Table 5. We observe that the ranks are always in favor of our proposal. WV–FROST significantly outperforms DynOVO for the MAUC evaluation at the 5% significance level. This is due to the strict exclusion of some candidate classes by DynOVO. The corresponding class probabilities are set to zero, which results in lower MAUC values. The preference of WV–FROST over DRCW is clearest for the average accuracy measure. Clearly, IFROWANN-$\mathcal{W}_{\text{IR}}$ within the OVO scheme interacts better with WV–FROST than with the existing dynamic aggregation approaches from [21,22]. Since our classification method and aggregation scheme are both based on fuzzy rough set theory, their superior synergy is an expected outcome. In fact, it motivated the use of this mathematical tool in the WV–FROST aggregation.

7.2 Comparison with the state-of-the-art

We now compare our full proposal to the state-of-the-art in multi-class imbalanced classification. Table 6 lists the full results of AdaBoost.NC (Ada), SMOTE-C4.5-WV (SMT), MDO-C4.5 (MDO) and FROVOCO (FR). For each dataset, we highlight the best performing method. For the average accuracy, this is our proposal for 11 out of the 18 datasets. For the MAUC, our method comes out on top for 15 of the 18 datasets. Its closest competitor

Table 6 Full average accuracy and MAUC results for the three state-of-the-art classifiers and our FROVOCO proposal

Data	AvgAcc				MAUC			
	Ada	SMT	MDO	FR	Ada	SMT	MDO	FR
aut	79.9444	**80.6444**	76.4778	77.1556	0.9370	0.9299	0.8928	**0.9633**
bal	65.8900	55.2701	56.4819	**78.8514**	0.8609	0.5901	0.6768	**0.8854**
cle	26.8750	26.1917	29.0417	**33.7833**	0.5834	0.5772	0.5610	**0.6981**
con	47.9522	**51.7246**	48.7521	47.4449	**0.6669**	0.6560	0.6529	0.6485
der	94.6845	96.2096	95.3709	**97.1553**	0.9857	0.9861	0.9727	**0.9966**
eco	76.2654	71.4609	71.1481	**77.2723**	0.9162	0.8990	0.8556	**0.9304**
gla	71.5516	**75.1885**	63.0437	67.0694	0.9246	0.9204	0.8534	**0.9325**
led	54.3621	63.5466	64.1728	**64.7918**	0.7640	0.9134	0.8780	**0.9189**
lym	72.4355	72.2222	73.2093	**86.2401**	0.7847	0.7717	0.8231	**0.9108**
new	**94.7222**	91.3889	90.4444	91.1111	0.9972	0.9563	0.9276	**0.9981**
pag	**91.9105**	89.2398	83.7520	90.0399	**0.9876**	0.9739	0.9446	0.9736
sat	87.5570	85.2928	84.7142	**89.4955**	0.9817	0.9619	0.9214	**0.9817**
shu	**98.4803**	96.8439	91.3154	91.8527	0.9911	0.9979	0.9600	**0.9987**
thy	**99.4186**	99.2688	97.9360	66.4500	**0.9998**	0.9965	0.9894	0.8494
win	94.7579	95.2698	92.9881	**98.2143**	0.9818	0.9788	0.9482	**1.0000**
wqr	39.6884	34.1986	31.8371	**43.7544**	0.7581	0.7495	0.6432	**0.8342**
wqw	47.6684	39.3455	39.3391	**47.8895**	0.7856	0.7772	0.6811	**0.8309**
yea	49.0789	51.8083	53.3381	**58.8178**	0.8279	0.8472	0.7693	**0.8810**
Mean	71.8468	70.8397	69.0757	**72.6327**	0.8741	0.8602	0.8306	**0.9018**

For each dataset, for each measure, the highest value is printed in bold

is AdaBoost.NC, that attains the highest average accuracy in four datasets and the highest MAUC in three. These results clearly indicate that our proposal, integrating IFROWANN in the OVO scheme with weighting scheme $\mathcal{W}_{\text{IR}}$ accompanied by the WV–FROST aggregation, has a better performance than the state-of-the-art with respect to both evaluation measures.

Statistical analysis The results of the Friedman test and the Holm post hoc procedure can be found in Table 7. For both evaluation measures, our method is assigned the lowest rank. It is shown to significantly outperform MDO-C4.5 for both metrics and SMOTE-C4.5-WV for the MAUC. Table 8 presents the results of the statistical comparison using the Wilcoxon test, more powerful to study the differences between pairs of methods. We compare our proposal to AdaBoost.NC, SMOTE-C4.5-WV and MDO-C4.5. For each pairwise comparison, the table also lists the number of wins (W) and losses (L) of FROVOCO. In all cases, the assigned ranks are in favor of FROVOCO. It significantly outperforms all state-of-the-art methods for the MAUC. For the average accuracy, it also performs significantly better than MDO-C4.5. Table 8 shows that our method dominates all others in the pairwise comparisons in terms of the number of wins on the 18 datasets as well.

It cannot be shown that the average accuracy of our method is significantly better than that of AdaBoost.NC or SMOTE-C4.5-WV, although, as stated above, the computed ranks indicate that our method is best. The explanation can be found in Table 6 with the thy-dataset. On this single dataset, our proposal performs very poorly compared to the others (as do all other OVO versions using IFROWANN). This difference is assigned the highest rank in favor

Table 7 Results of the Friedman test and Holm post hoc procedure

Method	AvgAcc		MAUC	
	Rank	p_{Holm}	Rank	p_{Holm}
FROVOCO	1.8333 (1)	–	1.4444 (1)	–
AdaBoost.NC	2.3333 (2)	0.245278	2.1667 (2)	0.09329
SMOTE-C4.5-WV	2.5556 (3)	0.18658	2.7222 (3)	**0.00597**
MDO-C4.5	3.2778 (4)	**0.002367**	3.6667 (4)	**0.000001**
p_{Friedman}	**0.008617**		**0.000003**	

Statistically significant differences at the 5% significance level are printed in bold

Table 8 Pairwise statistical comparisons by means of the Wilcoxon test

Comparison	W/L	R^+	R^-	p
A: FROVOCO versus AdaBoost.NC	11/7	108.0	63.0	0.32714
A: FROVOCO versus SMOTE-C4.5-WV	12/6	116.0	55.0	0.19638
A: FROVOCO versus MDO-C4.5	16/2	148.0	23.0	**0.004746**
M: FROVOCO versus AdaBoost.NC	15/3	139.0	32.0	**0.018234**
M: FROVOCO versus SMOTE-C4.5-WV	15/3	149.0	22.0	**0.004006**
M: FROVOCO versus MDO-C4.5	16/2	155.0	16.0	**0.0012894**

Lines preceded by 'A' correspond to the evaluation by AvgAcc, while those starting with 'M' are related to the evaluation by MAUC. Statistically significant differences at the 5% significance level are printed in bold

of the competing methods, which is why no statistical significance in the average accuracy can be detected.

As noted in Sect. 5, the two evaluation measures capture complementary performance information. The average accuracy solely focuses on the number of hits and misses, while the MAUC takes the confidence of the predictions of a classifier into account. Based on the analysis presented in Table 8, we can stress that the prediction confidences of our proposal are significantly more reliable than those of its competitors.

Effect of the IR When we combine the results in Table 6 with the IR information from Table 1, it can be observed that the datasets on which FROVOCO performs sub-optimally are mainly those with a high average IR. In particular, these are the pageblocks, shuttle and thyroid datasets, for which the high average IR is due to the presence of one very large majority class. When we investigated these results in more detail, we observed that the accuracy of our proposal on the single majority class is notably lower than that obtained by the other methods, which resulted in its low average accuracy value. Both the fact that there is only one majority class as well as its absolute volume explain why FROVOCO fails in this situation. Firstly, the method aims to boost the performance on the minority classes to such a degree that the classification of the single majority class is negatively affected. The combined effect of all minority classes against the single majority class leads to a decrease in average accuracy. Secondly (and probably most importantly), when the majority class is very large, the OWA-based fuzzy rough approximation operators are hindered in their recognition ability. In particular, weight vector (9), which is used when the IR between a pair of classes is high, loses some of its strength when the absolute size of one of these classes is very

high. The weight vector becomes very long and the weights on the non-zero positions almost flatten out to an average, due to their linear increasing character and the condition that all weights should sum to one. As a result, its desirable characteristics are lost.

On the winequality-white dataset, which has an average IR of 61.08, our method obtains the best average accuracy. This does not contradict our previous statement, since there are two majority classes present in this dataset, of which the sizes do not differ greatly. Moreover, the size of the largest class is also not that great compared to that of the pageblocks, shuttle and thyroid datasets.

We would like to note that our fuzzy rough approach highly depends on the similarity relation. When the overlap between classes is high (based on this measure), we can expect it to be more difficult for our method to adequately discern them. In such a situation, some confusion between classes is likely to occur, with classification errors as the result.

In general, as demonstrated by the results in Table 6 and analysis in Tables 7 and 8, our method outperforms its competitors, because it combines local (OVO decomposition) and global (WV–FROST) views of the data, thereby improving the recognition of difficult classes. Only in the particular case when a single massive majority class is present, the user may prefer to use AdaBoost.NC in the classification process. For reasons discussed in this paragraph, our method is less suited to handle this one specific type of problem. This situation can easily be checked for by a user before the application of a multi-class imbalanced classifier.

8 Conclusion

The IFROWANN method, which is based on fuzzy rough set theory, is a powerful classifier for binary imbalanced data. In this work, we studied its extension to the multi-class imbalanced setting, by combining it with the OVO decomposition process. In a first stage, we have shown that its success in existing OVO aggregation schemes is boosted by incorporating a newly proposed weighting scheme, represented in our method IFROWANN-$\mathcal{W}_{\text{IR}}$. Secondly, we have proposed a new aggregation scheme WV–FROST that further improves the results of IFROWANN-$\mathcal{W}_{\text{IR}}$ within the OVO setting. WV–FROST enhances the information extracted from the binary classifiers by including two global summary terms. Both are based on fuzzy rough set theory, yielding a nice synergy with the fuzzy rough classifiers. Their global character deals with the non-competent classifier issue encountered in OVO decompositions. Our experiments allow us to conclude that our complete proposal called FROVOCO, which is the combination of IFROWANN-$\mathcal{W}_{\text{IR}}$ in the OVO setting and WV–FROST, outperforms the state-of-the-art in multi-class imbalanced classification.

As future work, we propose to investigate the wider applicability of the WV–FROST step. In this paper, both the classifier within the OVO step and the summary terms in WV–FROST are based on fuzzy rough set theory. It will be interesting to verify whether the strength of WV–FROST transfers to settings where a different internal classifier is used in the OVO procedure. Furthermore, we could easily replace the WV part by any of the other existing OVO aggregation methods and validate whether our proposed summary terms yield similar improvements in those cases. The performance of WV–FROST on balanced datasets is left to be evaluated as well.

Acknowledgements The research of Sarah Vluymans is funded by the Special Research Fund (BOF) of Ghent University. This work was partially supported by the Spanish Ministry of Science and Technology under the Projects TIN2014-57251-P and TIN2015-68454-R; the Andalusian Research Plans P11-TIC-7765 and P12-TIC-2958. Yvan Saeys is an ISAC Marylou Ingram Scholar.

References

1. Abdi L, Hashemi S (2016) To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Trans Knowl Data Eng 28(1):238–251
2. Alshomrani S, Bawakid A, Shim S, Fernández A, Herrera F (2015) A proposal for evolutionary fuzzy systems using feature weighting: dealing with overlapping in imbalanced datasets. Knowl Based Syst 73:1–17
3. Barandela R, Sánchez J, García V, Rangel E (2003) Strategies for learning in class imbalance problems. Pattern Recog 36(3):849–851
4. Batista G, Prati R, Monard MC (2004) A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explor 6(1):20–29
5. Britto AS Jr, Sabourin R, de Oliveira LES (2014) Dynamic selection of classifiers—a comprehensive review. Pattern Recog 47(1):3665–3680
6. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
7. Chen Y (2016) An empirical study of a hybrid imbalanced-class DT–RST classification procedure to elucidate therapeutic effects in uremia patients. Med Biol Eng Comput 54(6):983–1001
8. Cornelis C, Verbiest N, Jensen R (2010) Ordered weighted average based fuzzy rough sets. In: Yu J, Greco S, Lingras P, Wang G, Skowron A (eds) Rough set and knowledge technology. Springer, Berlin, pp 78–85
9. D'eer L, Verbiest N, Cornelis C, Godo L (2015) A comprehensive study of implicator–conjunctor-based and noise-tolerant fuzzy rough sets: definitions, properties and robustness analysis. Fuzzy Sets Syst 275:1–38
10. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
11. Domingos P (1999) MetaCost: a general method for making classifiers cost—sensitive. In: Fayyad U, Chaudhuri S, Madigan D (eds) Proceedings of the 5th international conference on knowledge discovery and data mining (KDD'99). ACM, New York, pp 155–164
12. Dubois D, Prade H (1990) Rough fuzzy sets and fuzzy rough sets. Int J Gen Syst 17(2–3):191–209
13. Fei B, Liu J (2006) Binary tree of SVM: a new fast multiclass training and classification algorithm. IEEE Trans Neural Netw 17(3):696–704
14. Fernández A, Calderon M, Barrenechea E, Bustince H, Herrera F (2010a) Solving multi-class problems with linguistic fuzzy rule based classification systems based on pairwise learning and preference relations. Fuzzy Sets Syst 161(23):3064–3080
15. Fernández A, García S, Luengo J, Bernado-Mansilla E, Herrera F (2010b) Genetics-based machine learning for rule induction: state of the art, taxonomy and comparative study. IEEE Trans Evol Comput 14(6):913–941
16. Fernández A, López V, Galar M, Del Jesus MJ, Herrera F (2013) Analysing the classification of imbalanced data-sets with multiple classes: binarization techniques and ad-hoc approaches. Knowl Based Syst 42:97–110
17. Friedman JH (1996) Another approach to polychotomous classification. Tech rep, Department of Statistics, Stanford University. http://www-stat.stanford.edu/~jhf/ftp/poly.ps.Z
18. Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32(200):675–701
19. Fürnkranz J, Hüllermeier E, Vanderlooy S (2009) Binary Decomposition Methods for Multipartite Ranking. In: Buntine W, Grobelnik M, Mladenić D, Shawe-Taylor J (eds.) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2009. Lecture Notes in Computer Science, vol 5781. Springer, Berlin, Heidelberg
20. Galar M, Fernández A, Barrenechea E, Bustince H, Herrera F (2011) An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. Pattern Recog 44(8):1761–1776
21. Galar M, Fernández A, Barrenechea E, Bustince H, Herrera F (2013) Dynamic classifier selection for one-vs-one strategy: avoiding non-competent classifiers. Pattern Recog 46(12):3412–3424
22. Galar M, Fernández A, Barrenechea E, Herrera F (2015) DRCW-OVO: distance-based relative competence weighting combination for one-vs-one strategy in multi-class problems. Pattern Recog 48(1):28–42
23. Galar M, Fernández A, Barrenechea E, Bustince H, Herrera F (2016) Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced datasets. Inf Sci 354:178–196
24. Gao X, Chen Z, Tang S, Zhang Y, Li J (2016) Adaptive weighted imbalance learning with application to abnormal activity recognition. Neurocomputing 173:1927–1935

25. Gao Z, Zhang L, Chen M, Hauptmann A, Zhang H, Cai A (2014) Enhanced and hierarchical structure algorithm for data imbalance problem in semantic extraction under massive video dataset. Multimed Tools Appl 68(3):641–657
26. García S, Fernández A, Luengo J, Herrera F (2010) Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. Inf Sci 180(10):2044–2064
27. García V, Mollineda RA, Sánchez JS (2008) On the k-nn performance in a challenging scenario of imbalance and overlapping. Pattern Anal Appl 11(3–4):269–280
28. Haixiang G, Yijing L, Yanan L, Xiao L, Jinling L (2016) BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification. Eng Appl Artifl Intell 49:176–193
29. Hand DJ, Till RJ (2001) A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn 45(2):171–186
30. Hastie T, Tibshirani R (1998) Classification by pairwise coupling. Ann Stat 26(2):451–471
31. He H, Garcia E (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284
32. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6(2): 65–70
33. Huhn J, Hüllermeier E (2009) FR3: a fuzzy rule learner for inducing reliable classifiers. IEEE Trans Fuzzy Syst 17(1):138–149
34. Hüllermeier E, Brinker K (2008) Learning valued preference structures for solving classification problems. Fuzzy Sets Syst 159(18):2337–2352
35. Hüllermeier E, Vanderlooy S (2010) Combining predictions in pairwise classification: an optimal adaptive voting strategy and its relation to weighted voting. Pattern Recog 43(1):128–142
36. Jensen R, Cornelis C (2011) Fuzzy-rough nearest neighbour classification and prediction. Theor Comput Sci 412(42):5871–5884
37. Kuncheva L, Bezdek J, Duin R (2001) Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recog 34(2):299–314
38. Liu B, Hao Z, Yang X (2007) Nesting algorithm for multi-classification problems. Soft Comput 11(4):383–389
39. Liu B, Hao Z, Tsang ECC (2008) Nesting one-against-one algorithm based on SVMs for pattern classification. IEEE Trans Neural Netw 19(12):2044–2052
40. López V, Fernández A, Moreno-Torres JG, Herrera F (2012) Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. Expert Syst Appl 39(7):6585–6608
41. López V, Fernández A, Del Jesus M, Herrera F (2013a) A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline data-sets. Knowl Based Syst 38:85–104
42. López V, Fernández A, García S, Palade V, Herrera F (2013b) An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci 250:113–141
43. López V, Fernández A, Herrera F (2014) On the importance of the validation technique for classification with imbalanced datasets: addressing covariate shift when data is skewed. Inf Sci 257:1–13
44. Lorena AC, Carvalho AC, Gama JM (2008) A review on the combination of binary classifiers in multiclass problems. Artif Intell Rev 30(1–4):19–37
45. Mahalanobis P (1936) On the generalized distance in statistics. Proc Natl Inst Sci (Calcutta) 2:49–55
46. Martínez-Munoz G, Hernández-Lobato D, Suárez A (2009) An analysis of ensemble pruning techniques based on ordered aggregation. IEEE Trans Pattern Anal Mach Intellig 31(2):245–259
47. Moreno-Torres JG, Sáez JA, Herrera F (2012) Study on the impact of partition-induced dataset shift on-fold cross-validation. IEEE Trans Neural Netw Learn Syst 23(8):1304–1312
48. Orriols-Puig A, Bernado-Mansilla E (2009) Evolutionary rule-based systems for imbalanced datasets. Soft Comput 13(3):213–225
49. Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11(5):341–356
50. Platt JC, Cristianini N, Shawe-Taylor J (2000) Large margin DAGs for multiclass classification. In: Solla S, Leen T, Müller K (eds) Advances in neural information processing systems. MIT Press, Cambridge, pp 547–553
51. Ramentol E, Vluymans S, Verbiest N, Caballero Y, Bello R, Cornelis C, Herrera F (2015) IFROWANN: imbalanced fuzzy-rough ordered weighted average nearest neighbor classification. IEEE Trans Fuzzy Syst 23(5):1622–1637
52. Razakarivony S, Jurie F (2016) Vehicle detection in aerial imagery: a small target detection benchmark. J Vis Commun Image Represent 34:187–203
53. Rokach L (2016) Decision forest: twenty years of research. Inf Fusion 27:111–125

54. Sáez JA, Luengo J, Stefanowski J, Herrera F (2015) SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf Sci 291:184–203
55. Sun Y, Wong AKC, Kamel MS (2009) Classification of imbalanced data: a review. Int J Pattern Recog Artif Intell 23(4):687–719
56. Verbiest N, Ramentol E, Cornelis C, Herrera F (2014) Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. Appl Soft Comput 22:511–517
57. Villar P, Fernández A, Carrasco R, Herrera F (2012) Feature selection and granularity learning in genetic fuzzy rule-based classification systems for highly imbalanced data-sets. Int J Uncertain Fuzz 20(03):369–397
58. Vluymans S, D'eer L, Saeys Y, Cornelis C (2015) Applications of fuzzy rough set theory in machine learning: a survey. Fundam Inform 142(1–4):53–86
59. Vluymans S, Sánchez Tarragó D, Saeys Y, Cornelis C, Herrera F (2016) Fuzzy rough classifiers for class imbalanced multi-instance data. Pattern Recog 53:36–45
60. Vriesmann LM, Britto AS Jr, Oliveira LES, Koerich AL, Sabourin R (2015) Combining overall and local class accuracies in an oracle-based method for dynamic ensemble selection. In: Proceedings of the 2015 international joint conference on neural networks (IJCNN). IEEE, pp 1–7
61. Wang S, Yao X (2012) Multiclass imbalance problems: analysis and potential solutions. IEEE Trans Syst Man Cybern Part B 42(4):1119–1130
62. Wang S, Chen H, Yao X (2010) Negative correlation learning for classification ensembles. In: Proceedings of the 2010 international joint conference on neural networks (IJCNN). IEEE, pp 1–8
63. Wilcoxon F (1945) Individual comparisons by ranking methods. Biom Bull 1(6):80–83
64. Woods K (1997) Combination of multiple classifiers using local accuracy estimates. IEEE Trans Pattern Anal Mach Intell 19:405–410
65. Wu TF, Lin CJ, Weng RC (2004) Probability estimates for multi-class classification by pairwise coupling. J Mach Learn Res 5:975–1005
66. Yager R (1988) On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Trans Syst Man Cybern 18(1):183–190
67. Yijing L, Haixiang G, Xiao L, Yanan L, Jinling L (2016) Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. Knowl Based Syst 94:88–104
68. Yu H, Hong S, Yang X, Ni J, Dan Y, Qin B (2013) Recognition of multiple imbalanced cancer types based on DNA microarray data using ensemble classifiers. BioMed Res Int 2013:1–13
69. Zadeh LA (1965) Fuzzy sets. Inform Control 8(3):338–353
70. Zhang Z, Krawczyk B, Garcìa S, Rosales-Pérez A, Herrera F (2016) Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data. Knowl Based Syst 106:251–263
71. Zhao X, Li X, Chen L, Aihara K (2008) Protein classification with imbalanced data. Proteins: Struct Funct Bioinform 70(4):1125–1132
72. Zhou Z, Liu X (2010) On multi-class cost-sensitive learning. Comput Intell 26(3):232–257

Sarah Vluymans obtained her M.Sc. degree in Mathematical Informatics at Ghent University (Belgium) in 2014. Currently, she is working toward a joint Ph.D. in Computer Science at Ghent University and the University of Granada. She is also affiliated with the Inflammation Research Center, part of the Flemish Institute of Biotechnology, as a member of the DAMBI research group (Data Mining and Modeling for Biomedicine). Her research is funded by the Special Research Fund (BOF) by Ghent University. Her work focuses on the integration of concepts from fuzzy rough theory in a wide variety machine learning techniques.

Alberto Fernández received the M.Sc. and Ph.D. degrees in Computer Science from the University of Granada, Granada, Spain, in 2005 and 2010, respectively. He is currently an Assistant Professor with the Department of Computer Science and Artificial Intelligence, University of Granada. He has published more than 100 papers in highly rated JCR journals and international conferences, collecting up to 5500 citations (according to Google Scholar, September 2017). He has 6 research works included in the "highly cited papers" list from Web of Science (September 2017). In 2013, 2014 and 2017 Dr. Fernández received the University of Granada Prize for Scientific Excellence Works in the field of Engineering. His research interests include classification in imbalanced domains, fuzzy rule learning, evolutionary algorithms and evolutionary fuzzy systems, multi-classification problems with ensembles and decomposition techniques, and data science in Big Data applications.

Yvan Saeys obtained his M.Sc. and Ph.D. in Computer Science at Ghent University. He is currently leading the DAMBI group, where his research focuses on the development and application of data mining and machine learning techniques for biological and medical applications. He has published over 60 papers in international journals and high-profile conferences and has been involved in the organization of many international workshops and conferences on machine learning and bioinformatics. His current research topics include instance and feature selection, large-scale machine learning, and machine learning for structured input and output spaces.

Chris Cornelis received the M.Sc. and Ph.D. degrees in Computer Science from Ghent University, Ghent, Belgium. He is currently a doctor-assistant with Ghent University Belgium. He has co-supervised eight Ph.D. theses and has authored more than 160 papers in international journals, edited volumes, and conference proceedings. His current research interests include fuzzy sets, rough sets, and machine learning. Dr. Cornelis serves as an Executive Board Member of the International Rough Set Society.

Francisco Herrera received his M.Sc. and Ph.D. in Mathematics in 1988 and 1991, both from the University of Granada, Spain. He is currently a Professor in the Department of Computer Science and Artificial Intelligence at the University of Granada. He has been the supervisor of 41 Ph.D. students. He has published more than 350 journal papers, receiving more than 55,000 citations (Scholar Google, H-index 116). He is coauthor of the books "Data Preprocessing in Data Mining" (Springer, 2015), "Multilabel Classification. Problem analysis, metrics and techniques" (Springer, 2016), "Multiple Instance Learning. Foundations and Algorithms" (Springer, 2016), among others. He acts as Editor in Chief of the journals "Information Fusion" (Elsevier) and "Progress in Artificial Intelligence" (Springer), and editorial member of a dozen of journals. He has been selected as a Highly Cited Researcher http://highlycited.com/ (in the fields of Computer Science and Engineering, respectively, 2014 to present, Clarivate Analytics). His research interests includes soft computing (including fuzzy modeling and evolutionary algorithms), information fusion and decision making, data preprocessing, data science and big data.