IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 28, 2020, accepted March 12, 2020, date of publication March 16, 2020, date of current version March 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980949

Unsupervised Fuzzy Measure Learning for Classifier Ensembles From Coalitions Performance

MIKEL URIZ, DANIEL PATERNAIN, IRIS DOMINGUEZ-CATENA, HUMBERTO BUSTINCE, (Senior Member, IEEE), AND MIKEL GALAR, (Member, IEEE)
Institute of Smart Cities, Public University of Navarra, 31006 Pamplona, Spain
Corresponding author: Mikel Uriz (mikelxabier.uriz@unavarra.es)

This work was supported in part by the Spanish Ministry of Science and Technology (AEI/FEDER, UE) under Project TIN2016-77356-P, and in part by the Public University of Navarra under Project PJUPNA13.

ABSTRACT In Machine Learning an ensemble refers to the combination of several classifiers with the objective of improving the performance of every one of its counterparts. To design an ensemble two main aspects must be considered: how to create a diverse set of classifiers and how to combine their outputs. This work focuses on the latter task. More specifically, we focus on the usage of aggregation functions based on fuzzy measures, such as the Sugeno and Choquet integrals, since they allow to model the coalitions and interactions among the members of the ensemble. In this scenario the challenge is how to construct a fuzzy measure that models the relations among the members of the ensemble. We focus on unsupervised methods for fuzzy measure construction, review existing alternatives and categorize them depending on their features. Furthermore, we intend to address the weaknesses of previous alternatives by proposing a new construction method that obtains the fuzzy measure directly evaluating the performance of each possible subset of classifiers, which can be efficiently computed. To test the usefulness of the proposed fuzzy measure, we focus on the application of ensembles for imbalanced datasets. We consider a set of 66 imbalanced datasets and develop a complete experimental study comparing the reviewed methods and our proposal.

INDEX TERMS Fuzzy measures, Choquet integral, aggregation, ensembles, classification.

I. INTRODUCTION

Classification is one of the most well-known examples of Machine Learning, since many real-world problems can be formulated as classification problems [1], [2]. Specifically, supervised classification consists of learning a classifier from labeled data in such a way that it is able to correctly classify new examples (also called instances) that were not taken into consideration during the learning step [3]. This behavior is called generalization capability, and is the most desirable property of any learned classifier.

In the literature, many learning algorithms have been proposed, e.g., Decision Trees [4], Support Vector Machines [5] or Fuzzy Classifiers [6]. However, it is well-known that none of them is able to outperform the rest in every problem (see [7] for more details). One way of improving the performance of single classifiers is to combine several of them by learning an ensemble of classifiers. In this context, individual classifiers must be diverse to take advantage of the different answers provided by the classifiers.

To construct a classifier ensemble, two main aspects needs to be considered. Firstly, how to generate diversity among classifiers. Bagging [8] and Boosting [9] are considered the most popular techniques for this purpose. Secondly, how to combine the answers provided by the classifiers into a single output label. Several approaches can be found in the literature for classifier combination such as weighted voting, Naive Bayes, Decision Templates or Stacking among

The associate editor coordinating the review of this manuscript and approving it for publication was Corrado Mencar.

others [10], [11]. Another type of combination strategy consists of the usage of fuzzy integrals, such as the Choquet [12] and Sugeno [13] integrals, which are based on an underlying fuzzy measure that models the relations among the sources of information to be combined, i.e., the classifiers of the ensemble. In this paper, we focus on these approaches, where the construction of the fuzzy measure is the key factor.

In this context, a fuzzy measure models the interactions among every possible coalition (subset) of classifiers. The main difficulty when constructing a fuzzy measure is the large number of coefficients to be estimated ($2^N - 2$, being N the number of classifiers). As a consequence, traditionally in the literature there has been two different ways for addressing this issue. 1) Reducing the number of coefficients to be estimated [13]; 2) Estimating the whole set of coefficients, but obtaining them indirectly from information obtained from the individual classifiers or at most pairwise measures between them [14]. In the field of ensemble classification, the majority of fuzzy measure construction methods are unsupervised, in the sense that the desired output of the fuzzy integral is unknown. In this paper, we review the existing methods in the literature and classify them analyzing their characteristics into different categories. Although supervised construction methods exist [15], their analysis is out of the scope of this paper, and we leave the comparison between both types of approaches for future work.

Attending to the drawbacks of existing methods, the objective of this work is to propose a new fuzzy measure construction method based on directly estimating the whole set of coefficients. This method is named as Coalition Performance-based Measure (CPM) and estimates each coefficient by efficiently measuring the performance of the corresponding coalition of classifiers. The novelty of this methodology lies in the avoidance of both indirect measures and simplifications of the fuzzy measure. Our hypothesis is that using the full potential of the fuzzy measure with accurately estimated coefficients can lead to a final improved performance of the ensemble.

To show the usefulness of CPM, we develop an exhaustive empirical study, where we compare the existing alternatives for fuzzy measure construction in classifier ensembles and our proposal. These fuzzy measures are evaluated with both the Choquet and Sugeno integrals. Moreover, we also consider other classical aggregations to complete our study and show the benefits of fuzzy integral-based approaches. Due to our previous experience on the topic and the fact that the importance of aggregation is highlighted in complex problems, we focus on the challenging framework of imbalanced datasets [16], [17] and show that the state-of-the-art performace can be improved using fuzzy measure-based methods.

The complete experimental study is formed of the sixty six datasets from KEEL dataset repository [18]. We consider the UnderBagging ensemble method [19] for generating classifiers and apply the Reduced Error Pruning with Geometric Mean [20] to obtain the final ensemble. This algorithm is chosen for being the best performer in previous studies and we aim to analyze whether fuzzy measure-based aggregations are able to improve its performance. The performance of each method will be measured by the geometric mean (GM) of the performance over each class due to the nature of the class imbalance problem. As suggested in the literature [21], the results will be properly analyzed using non-parametric stastiscal tests.

The remainder of this paper is organized as follows. In Section II we recall the main ideas of aggregation functions. Later, Section III introduces ensembles and the class imbalance problem. In Section IV, existing unsupervised fuzzy measure construction algorithms are reviewed. Then, in Section V we present CPM, our proposal for constructing fuzzy measures. Afterwards, we present the experimental framework in Section VI and the experimental study in Section VII. We end this paper with the conclusions and future research lines in Section VIII.

II. AGGREGATION FUNCTIONS FOR INFORMATION FUSION

Aggregation functions are known to be an important mathematical tool to deal with information fusion.

Definition 1 [22]–[25]: A mapping $f : [0, 1]^n \to [0, 1]$ is called an aggregation function if it satisfies:

- boundary conditions: $f(0, \ldots, 0) = 0, f(1, \ldots, 1) = 1$;
- increasing monotonicity: if $x_i \leq y_i$ for all $i \in \{1, \ldots, n\}$, then $f(x_1, \ldots, x_n) \leq f(y_1, \ldots, y_n)$.

One of the most prominent family of aggregation functions are averaging aggregation functions or means. An aggregation functions is said to have an averaging behavior if $\min(x_1, \ldots, x_n) \leq f(x_1, \ldots, x_n) \leq \max(x_1, \ldots, x_n)$ for every $(x_1, \ldots, x_n) \in [0, 1]^n$. Recall that averaging aggregation functions also satisfy the idempotence property, i.e., $f(x, \ldots, x) = x$ for every $x \in [0, 1]$. Prototypical examples of averaging aggregation functions are the arithmetic mean, the geometric mean, the median, the minimum and the maximum. In fact, the minimum and the maximum are the lowest and greatest averaging aggregation functions, respectively.

When we deal with real-world problems where an aggregation function must be used, it is desirable to incorporate some sources of information in addition to the own inputs to be aggregated. This problem can be solved by applying a certain kind of aggregation functions, called weighted aggregation function. The weighting vector associated to the function allows to model the importance of individual attribute or criterion to be fused.

Definition 2: A vector $\mathbf{w} = (w_1, \ldots, w_n)$ is called a weighting vector if $w_i \in [0, 1]$ and $\sum_{i=1}^{n} w_i = 1$.

Example 1: The weighted arithmetic mean associated to the weighting vector $\mathbf{w}$ is given by $WAM(x_1, \ldots, x_n) = \sum_{i=1}^{n} w_i x_i$. Notice that the weighted arithmetic mean is symmetric only if $w_i = \frac{1}{n}$ for every $i \in \{1, \ldots, n\}$.

A wide family of weighted aggregation functions are the so-called OWA operators.

Definition 3: Let $\mathbf{w}$ be a weighting vector. An OWA operator OWA_w associated with $\mathbf{w}$ is a mapping $OWA_w : [0,1]^n \to [0,1]$ defined by

$$OWA_w(x_1, \ldots, x_n) = \sum_{i=1}^{n} w_i x_{(i)} \tag{1}$$

where $x_{(i)}$ denotes the i-th greatest component of the input $(x_1, \ldots, x_n)$.

In this work we will use weighting vectors induced by increasing quantifiers [26]. An increasing quantifier Q is a mapping $Q : [0,1]^n \to [0,1]$ satisfying $Q(0) = 0$, $Q(1) = 1$ and $Q(x) \geq Q(y)$ whenever $x > y$. Then, given an increasing quantifier Q, the weighting vector $\mathbf{w}$ induced by Q is given, for every $i \in \{1, \ldots, n\}$, by

$$w_i = Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right).$$

Example 2: The following picewise linear functions are examples of increasing quantifiers:

$$Q_{alh}(x) = \begin{cases} 2x & x \leq 0.5 \\ 1 & \text{oth.} \end{cases} \qquad Q_{amap}(x) = \begin{cases} 0 & x \leq 0.5 \\ 2x-1 & \text{oth.} \end{cases}$$

$$Q_{mot}(x) = \begin{cases} 0 & x \leq 0.3 \\ 2x - 0.6 & 0.3 < x \leq 0.8 \\ 1 & \text{oth.} \end{cases}$$

In fact, these quantifiers are well-known in linguistic modeling (see, for example [27]) and the corresponding OWA operators induced by Q are known as *at least half*, *as many as possible* and *most of them*, respectively.

When we deal with complex problems of information fusion, inputs to be fused (criteria, attributes or sources of information) may not be totally independent. This means that we can have either positive or negative interaction among inputs. Under these circumstances, weighted aggregation functions are insufficient. However, a more adequate tool to model these interactions are fuzzy (non-additive) measures [28], [29] and aggregation functions based on these measures.

Definition 4: Let $\mathcal{N} = \{1, \ldots, N\}$. A discrete fuzzy measure is a set function $m : 2^{\mathcal{N}} \to [0,1]$ satisfying boundary conditions $m(\emptyset) = 0$, $m(\mathcal{N}) = 1$ and monotonicity with respect to the inclusion, i.e. $m(A) \leq m(B)$ whenever $A \subset B$ for every $A, B \subseteq \mathcal{N}$.

For defining a fuzzy measure, it is necessary to define its $2^N - 2$ components. This may be a complex task when N is large enough. To simplify this problem, some reformulations have been given in the literature, being the Sugeno λ-measures [13] one of the most well-known examples.

Definition 5: Let $\lambda \in (-1, \infty)$. We say that $m : 2^{\mathcal{N}} \to [0,1]$ is a Sugeno λ-measure if, for every $A, B \subseteq \mathcal{N}$ with $A \cap B = \emptyset$, we have that $m(A \cup B) = m(A) + m(B) + \lambda m(A) m(B)$.

The most widely used fuzzy measure-based aggregation functions are the Choquet and the Sugeno integrals.

Definition 6: Given a fuzzy measure $m : 2^{\mathcal{N}} \to [0,1]$, the discrete Choquet integral is given by

$$C_m(x_1, \ldots, x_N) = \sum_{i=1}^{N} (x_{\sigma(i)} - x_{\sigma(i-1)}) m(\{\sigma(i), \ldots, \sigma(N)\})$$

where $\sigma : \mathcal{N} \to \mathcal{N}$ is a permutation such that $x_{\sigma(1)} \leq \ldots \leq x_{\sigma(N)}$ and $x_{\sigma(0)} = 0$ for convention.

Remark 1: Notice that if m is additive, i.e. for any $A, B \subset \mathcal{N}$, $A \cap B = \emptyset$ then $m(A \cup B) = m(A) + m(B)$, the Choquet integral is the weighted arithmetic mean. If m is symmetric, i.e. for any $A, B \subseteq \mathcal{N}$, $m(A) = m(B)$ whenever $|A| = |B|$, then the Choquet integral is an OWA operator. Finally, if m is symmetric and additive, then the Choquet integral is the arithmetic mean.

Definition 7: Given a fuzzy measure $m : 2^{\mathcal{N}} \to [0,1]$, the discrete Sugeno integral is given by

$$S_m(x_1, \ldots, x_N) = \max_{i=1}^{N} \min\{x_{\sigma(i)}, m(\{\sigma(i), \ldots, \sigma(N)\})\}$$

where $\sigma : \mathcal{N} \to \mathcal{N}$ is a permutation such that $x_{\sigma(1)} \leq \ldots \leq x_{\sigma(N)}$.

III. ENSEMBLES AND THE CLASS IMBALANCE PROBLEM

This section introduces the concept of ensembles, the problem of class imbalance in classification and the ensembles specifically designed for this problem.

A. CLASSIFIER ENSEMBLES

An ensemble in Machine Learning refers to a set of classifiers, which are used together to solve a problem with the assumption that their combination will lead to better performance than using a single classifier. One key aspect in classifier ensembles is diversity, since there is no way to improve the performance other than having complementary classifiers in the ensemble. For this reason, different ways for creating diverse classifiers out of the same original data have been proposed [10], [17]. Once the ensemble has been built, new examples are classified by querying all the classifiers and aggregating their outputs to obtain the final output label. Other terms used to refer to this combination step are classifier fusion or aggregation [10]. This phase is the main focus of this work, although we first recall the most common ways for creating diverse ensembles before going through their combination.

Classifier learning algorithms focus on building classifiers with a good trade-off between accuracy and diversity. Among them, the most well-known algorithms are AdaBoost [9] and Bagging [8]. In both of them, classifiers are learned strategically by altering the dataset used to learn each classifier. These algorithms require the usage of a weak learner to build the set of classifiers (a classifier in which small changes in data produces big changes in the model). In this work, we focus on Bagging because all the classifiers obtained are *a priori* similar and hence, no specific weights are assigned to the classifiers (as in Boosting).

Bagging (*boostrap aggregating*) was proposed by Breiman [8] as a simple but effective way to build ensembles. In this method, diversity is achieved by training each classifier with a different bootstrapped replica of the original dataset. Hence, a new dataset is build for each classifier by randomly drawing (with replacement) instances from the original dataset. This resampling mechanism clearly requires the usage of a weak learner to achieve diversity. Notice that the original size of the dataset is usually considered for the resampling, which results in approximately 63.2% of the instances being present in each bag. The pseudo-code of Bagging is shown in Algorithm 1.

Algorithm 1 Bagging

Input: S: Training set; N: Number of iterations; n: Bootstrap size; I: Weak learner

Output: Bagged classifier: $Class(x) = \arg\max_{y\in\mathcal{C}} \left(\frac{1}{N}\sum_{i=1}^{N} p_{c_i}(y|x)\right)$ where $p_{c_i}(y|x) \in [0,1]$ is the probability of x belonging to class y given by the classifier c_i

1: **for** $i = 1$ to N **do**
2: $S_i \leftarrow$ RandomSampleReplacement(n, S)
3: $c_i \leftarrow$ I(S_i)
4: **end for**

Bagging belongs to classifier fusion strategies and hence, the outputs of all the induced classifiers are taken into account to classify new instances. The weighted majority voting is commonly used for aggregation, where the confidences given by the classifier are considered. Therefore, the final class is decided from the following formula:

$$Class(x) = \arg\max_{y\in\mathcal{C}} \left(\frac{1}{N}\sum_{i=1}^{N} p_{c_i}(y|x)\right) \tag{2}$$

where $p_{c_i}(y|x) \in [0,1]$ corresponds to the output probability given by classifier c_i for class y and $\mathcal{C}$ is the set of classes (the number of classes $|\mathcal{C}|$ will be denoted by C). Notice that classifiers not giving probabilities as outputs but confidence degrees in favor of each class can also be used (which would substitute $p_{c_i}(y|x)$). In this case, the confidence degrees given by each classifier should be properly calibrated so that none of them dominates the aggregation. Otherwise, if we assume that the output probabilities are discrete (either 0 or 1), then the majority voting strategy would be recovered.

The key point in this paper is that instead of using this simple averaging formula, one could chose to substitute the aggregation by any of the functions presented in the previous section. In fact, one can take the interactions among classifiers into account in the aggregation phase, instead of simply averaging their outputs. In this work, besides from proposing a new way for constructing fuzzy measures for classifier combination, substituting Eq. 2 by the Choquet or Sugeno integrals, we will develop an in depth experimental comparison of the performance of different aggregations and different ways of creating fuzzy measures. The experimental framework will consider the problem of imbalanced datasets. We focus on this scenario because the importance of aggregation is highlighted in complex problems such as this one.

B. THE CLASS IMBALANCE PROBLEM

Imbalanced datasets pose a challenging scenario to classifier learning algorithms [30]. By definition, a dataset is said to be imbalanced whenever the number of examples from the different classes are not nearly the same. Focusing on two-class problems, the issue is that the class of interest is usually under-represented in the dataset [16]. Unfortunately, standard classifier learning algorithms tend to favor the majority class due to their accuracy-oriented design.

To deal with this situation, four main types of approaches are usually considered: algorithm adaptations [31], data preprocessing [32], cost-sensitive methods [33] and ensemble-based methods [17]. This paper focuses on the last ones, which mainly consists of the combination of a traditional ensemble learning method with one of the other types of approaches, especially data preprocessing and cost-sensitives methods. Accordingly, ensemble-based approaches can be further divided into three main classes, depending on which ensemble learning algorithm they are based on (Bagging, Boosting or Hybrid). A thorough empirical analysis of these solutions was carried out in [17], where the combination of Bagging and random undersampling, named as UnderBagging, stood out. Some of these ensemble methods were further developed in [20], where classifier pruning was considered to improve the final performance and yet UnderBagging (coupled with reduced error pruning) achieved the best results. For this reason, this is the method considered for our experimental study as it is the state-of-the-art on ensembles for imbalanced datasets. We want to highlight this fact because we will check whether fuzzy measure-based aggregations are able to make a difference when considering highly optimized ensembles. Notice that the better the ensemble and its base classifiers are, the less the margin for improvement due to the aggregation is. Hence, different from other works, we will study the behavior of aggregations in a challenging scenario.

In the following, we briefly recall UnderBagging with Reduced Error Pruning with Geometric Mean (UnderBagging_RE-GM). UnderBagging is the ensemble learning algorithm in charge of constructing the pool of classifiers. Since a pruning method is used afterwards (RE-GM), more than the necessary number of classifiers are generated. Following [34] and [20], 100 classifier are generated in this work. Notice that we will fully replicate the experimental study in [20] so as to carry out a fair comparison with the state-of-the-art methods. After classifier generation, the pruning method is applied to the reduce the total number of classifiers to 21 in this case (as recommended in [34] and in the same way as in [20]). The two components of UnderBagging_RE-GM works as follows:

- *UnderBagging*: It is a slight modification of Algorithm 1, where the random sample with replacement in line 2 is substituted by a random undersampling of the dataset. That is, each bag is created by randomly removing majority class instances from the dataset until the same number of majority and minority class instances are present.
- *Reduced Error Pruning with Geometric Mean (RE-GM)*: Pruning in ensembles refers to the elimination of classifiers that do not contribute to the ensemble performance (redundant classifiers). In RE, rather than eliminating classifiers, the final ensemble is formed by adding classifiers from the pool one by one. First, the classifier achieving the lowest classification error is added. Then, in each iteration, classifiers are ordered by the performance they achieve after being added to the sub-ensemble. Again, the one achieving the largest improvement is finally added. In the particular case of RE-GM, the GM performance measure is considered to decide which classifier is added in each iteration. This measure is explained afterwards.

In any Machine Learning task, properly evaluating the quality of a solution becomes a key factor. When dealing with imbalanced datasets this is not an exception. Furthermore, specific evaluation criteria are required, since the standard accuracy measure is no longer valid becausse it does not reflect the quality of the prediction over both classes. Commonly, the results of a classifier over a dataset are gathered in a confusion matrix (Table 1). From this matrix, several class-wise measure are obtained such as the True Positive Rate ($\mathrm{TP}_{rate} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$) and the True Negative Rate ($\mathrm{TN}_{rate} = \frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}$), which allow one to measure how well the classifier performs in each class. However, considering them separately easily leads to incorrect conclusions as their maximization becomes trivial. For this reason, one usually prefers to assess the performance of the classifier over both classes simultaneously. To do so, the geometric mean (GM) [31] is widely used (Eq. (3)).

$$\mathrm{GM} = \sqrt{\mathrm{TP}_{rate} \cdot \mathrm{TN}_{rate}}. \tag{3}$$

TABLE 1. Confusion matrix for a two-class problem.

	Positive prediction	Negative prediction
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

IV. RELATED WORK ON ENSEMBLE AGGREGATION BY THE CHOQUET INTEGRAL

In the literature it is not difficult to find a variety of data mining and machine learning algorithms where the Choquet integral plays an important role. For example, in [35] several applications of the Choquet integral in data mining problems are presented. However, this review does not consider many other applications, e.g. classification and patter recognition [36], [37], bioinformatics [38], fuzzy rule-based systems [39]–[41], preference learning [42], remote sensing [43], or ensemble reduction [44] and ensemble construction [45], [46], as it is the case of this paper.

The usage of the Choquet integral as an information fusion tool necessarily requires the construction of a fuzzy measure. We can have two main methods for constructing or learning the associated fuzzy measure: supervised and unsupervised learning. The former consists of estimating the coefficients by some optimization algorithm, where target outputs of the Choquet integral are provided and a set of restrictions must be kept, usually monotonicity constraints. In other words, these methods usually start from a predefined set of input vectors together with their corresponding (ideal) output and the objective consists of estimating a fuzzy measure whose corresponding Choquet integral fits the given outputs provided the input vectors. In fact, this can be seen as a regression problem where the underlying function is the Choquet integral. The optimization procedure to estimate the coefficients of the fuzzy measure may vary among neural networks [47], genetic algorithms [48], linear or quadratic programming [49], [50] or gradient-based algorithms [51], [52].

Unsupervised learning methods are the second major procedures to deal with the estimation of the coefficients of the fuzzy measure. In this sort of methods, there is no ideal outputs to be fit, but generally some prior knowledge about the information sources that conduct the learning strategy.

If we restrict ourselves to the specific problem of this paper, we recall that our aim is to combine the outputs of an ensemble method by using a Choquet integral and therefore, the associated fuzzy measure must be computed to model the interactions among the classifiers in the ensemble. In the literature, the mainstream methodologies for constructing the fuzzy measure for ensemble of classifiers are unsupervised. Then, depending on the prior knowledge that conducts the construction, we distinguish two main groups: 1) entropy-based construction algorithm, where the probability distribution given by the classifiers induce the coefficients; 2) classifiers' knowledge-based construction algorithms, where the fuzzy measure is induced by some performance measures, that vary between accuracy, confidence or diversity measures of the classifiers of the ensemble [53].

Aside from the information that conducts the construction algorithm (entropy or classifiers' knowledge), the way the classifiers' predictions are aggregated also influences the estimation of the fuzzy measure. Here, we can distinguish two different scenarios: dynamic and global (or static). In the dynamic scenario, the instance to be classified affects, up to some extent, the aggregation procedure, while in the global case, the same aggregation method is applied for the whole set of instances. Therefore, in the dynamic aggregation based on the Choquet integral, a fuzzy measure must be constructed for each instance. On the contrary, global aggregation requires a single fuzzy measure for the whole classification problem. In the next section, we will see that entropy-based methods

are dynamic, whereas classifiers' knowledge-based methods can be either dynamic or global.

Remark 2: Even though many papers in the literature deal with dynamic aggregation of classifiers using the Choquet integral, we may argue that formally this is not a pure Choquet integral-based procedure. In this sort of methods, the coefficients that take part in the aggregation are exclusively estimated, since the rest are only necessary for a formal definition of the fuzzy measure. However, we think that this proceeding fits better with the name of dynamic OWA or OWA dependent (see for example [54]–[56]) operators, rather than with the Choquet integral. However, since these works appear in the literature with the name of Choquet integral, we have taken into consideration for this study.

In the next subsections, we recall the most prominent methods we have found in the recent literature about the construction of a fuzzy measure for aggregating classifiers, both from the entropy-based approach and from the classifiers' knowledge-based approach. In the latter, we also distinguish between dynamic and global methods.

A. ENTROPY-BASED CONSTRUCTION ALGORITHMS

We will call entropy-based methods those whose construction method depends on the probability distribution given by each classifier for a specific instance of the problem, assuming that the lower the entropy of the distribution is, the more accurate the classifier is. The most recent proposal is given in [57], where the proposed construction method solves several complexity issues of previous entropy-based approaches, such as [58] and [59] (see also [60]). Essentially, it starts setting the coefficients of the singletons (sets whose cardinality is one) to be inversely proportional to the Shannon's entropy of the probability distribution of the considered classifier for the given instance. Later, the coefficients of those classifiers with very high measure value that strongly disagree with the rest of classifiers (those whose value is 1.5 times greater than the average value of the singletons) are truncated, since these classifier are treated as outliers. The rest of the coefficients associated with any combination of classifiers are computed in an additive way. The pseudocode of this approach is given in Algorithm 2.

B. CLASSIFIERS' KNOWLEDGE-BASED CONSTRUCTION ALGORITHMS

We will denote by classifiers knowledge-based methods those whose construction method depends on some prior knowledge or heuristic about the classifiers of the ensemble, which can be either characterized by a confidence degree, by a measure of diversity among classifiers, or by both simultaneously [14].

While the entropy-based methods are inherently dynamic, in this subcategory we can distinguish between dynamic and global aggregation of classifiers. For this reason, for some of the explained methods we will distinguish between its dynamic and global version.

Algorithm 2 EB

Input: x: instance of the problem; $\mathcal{N}$: Set of classifiers;
Output: fuzzy measure: $m : 2^{\mathcal{N}} \rightarrow [0, 1]$ associated to the instance x.

for $i = 1$ to N **do**
  $m(\{i\}) \leftarrow -\sum_{y \in C} p_{c_i}(y|x) \log(p_{c_i}(y|x))$
end for
$m_{max} \leftarrow \max_{i=1,\ldots,N} m(\{i\})$
for $i = 1$ to N **do**
  $m(\{i\}) \leftarrow m_{max} - m(\{i\})$
end for
$M \leftarrow \frac{1}{N} \sum_{i=1}^{N} m(\{i\})$
$O \leftarrow \{i | m(\{i\}) \leq 1.5M\}$
$M_r = \frac{1}{|O|} \sum_{i \in O} m(\{i\})$
for $i = 1$ to N **do**
  if $m(\{i\}) > 1.5M$ **then**
    $m(\{i\}) \leftarrow M_r$
  end if
end for
for $A \subseteq \mathcal{N}$ **do**
  $m(A) \leftarrow \sum_{i \in A} m(\{i\})$
end for
Normalize m to have $m(\mathcal{N}) = 1$

1) INTERACTION-SENSITIVE FUZZY MEASURE (ISFM) [14]

In this algorithm the fuzzy measure is constructed by means of a confidence vector collecting the confidence degree of each individual classifier, namely $\kappa = (\kappa_1, \ldots, \kappa_N)$ and a pairwise similarity measure, namely $S : \mathcal{N} \times \mathcal{N} \rightarrow [0, 1]$, whose objective is to approximate a measure of how similar (inversely, how dissimilar) two classifiers of the ensemble are. Then main idea under this algorithm is to construct a fuzzy measure where the coefficients of each coalition is given by the sum of the individual confidences weighted by the diversity of the classifiers among the group. Thus, when adding a new classifier to the coalition, if it is very similar to the existing ones, the coefficients remain stables, while if the incorporation increases diversity, the new coefficient of the fuzzy measure also increases.

- Dynamic ISFM (D-ISFM): the dynamic behavior of ISFM comes from the fact that, for each instance, the specific arrangement of the probabilities to be aggregated must be known in advance, thus generating different fuzzy measures depending on the explicit input vector. The construction of the fuzzy measure starts with the coefficient of the classifier predicting the highest probability, whose value is set to its corresponding confidence degree. Later, the second classifier is added to the coalition and the corresponding coefficient is calculated. The process is repeated until the whole set of classifiers is considered. The pseudocode of this approach is given in Algortithm 3.
- Global ISFM (G-ISFM): in this approach, instead of considering the specific ordering induced by the input

Algorithm 3 D-ISFM

Input: x: instance of the problem; $\mathcal{N}$: Set of classifiers; κ: confidence vector; S: similarity measure.

Output: fuzzy measure: $m : 2^{\mathcal{N}} \to [0, 1]$ associated to the instance x.

$\sigma \leftarrow$ permutation such that $p_{c_{\sigma(1)}(y|x)} \leq \cdots \leq p_{c_{\sigma(1)}(y|x)}$
$m(\{\sigma(n)\}) \leftarrow \kappa_{\sigma(n)}$
for $i = n-1, \ldots, i$ **do**
  $m(\{\sigma(i), \ldots, \sigma(n)\}) \leftarrow m(\{\sigma(i+1), \ldots, \sigma(n)\}) + \kappa_{\sigma(i)} \left(1 - \max_{k=i+1}^{r} S(\sigma(i), \sigma(k)\right)$
end for
Normalize m to have $m(\mathcal{N}) = 1$

vector of probabilities, the ordering is induced by the confidence of classifiers that are being taken into account. This means that the classifier with the highest confidence is the "original" classifier, and the rest of classifiers are later added to the coalition, thus generating an additive fuzzy measure. The pseudocode of this approach is given in Algortithm 4.

Algorithm 4 G-ISFM

Input: $\mathcal{N}$: Set of classifiers; κ: confidence vector; S: similarity measure.

Output: fuzzy measure: $m : 2^{\mathcal{N}} \to [0, 1]$.

$\sigma \leftarrow$ permutation such that $\kappa_{\sigma(1)} \leq \cdots \leq \kappa_{\sigma(n)}$
for $i = 1, \ldots, N$ **do**
  $m(\{\sigma(i)\}) = \kappa_{\sigma(i)} \left(1 - \max_{j=\sigma(i)+1,\ldots,\sigma(n)} S(\sigma(i), j)\right)$
end for
for each $A \subseteq \mathcal{N}$ **do**
  $m(A) \leftarrow \sum_{i \in A} m(\{i\})$
end for
Normalize m to have $m(\mathcal{N}) = 1$

2) MODIFIED HÜLLERMEIER MEASURE (MHM) [14], [61]

The measure proposed in this approach was originally given in [62] for a modification of the K-NN algorithm using the Choquet integral, and later adapted in [14] for the aggregation of classifiers. This algorithm starts from a given additive fuzzy measure m' constructed from the confidence degree of each classifier and transforms it into a non-additive measure by considering the diversity of classifiers. Here again, the diversity of the coalitions controls the additivity of the fuzzy measure by a parameter α.

- Dynamic MHM (D-MHM): as in D-ISFM, the specific arrangement of the probabilities predicted for an instance is considered. We start with the classifier predicting the highest probability, whose corresponding coefficient is inherited from the additive measure (confidence degree). As we add new classifiers, the coefficients are calculated by weighting the additive coefficient (sum of confidence degrees) with a measure of the "relative diversity" of the set of classifiers.

Algorithm 5 D-MHM

Input: x: instance of the problem; $\mathcal{N}$: Set of classifiers; κ: confidence vector; S: similarity measure; $\alpha \in [0, 1]$: parameter.

Output: fuzzy measure: $m : 2^{\mathcal{N}} \to [0, 1]$ associated to the instance x.

Construct additive fuzzy measure m'
for $A \subseteq \mathcal{N}$ **do**
  $m'(A) = \sum_{i \in A} \kappa_i$
end for
Normalize m' to have $m'(\mathcal{N}) = 1$
$\overline{ds} \leftarrow \max_{i \neq j \in \mathcal{N}} S(i, j) = 1 - \min_{i \neq j \in \mathcal{N}} S(i, j)$
$\sigma \leftarrow$ permutation such that $p_{c_{\sigma(1)}(y|x)} \leq \cdots \leq p_{c_{\sigma(1)}(y|x)}$
$m(\{\sigma(n)\}) \leftarrow m'(\{\sigma(n)\})$
for $i = n-1, \ldots, i$ **do**
  $A \leftarrow \{\sigma(i), \ldots, \sigma(n)\}$
  $div(A) \leftarrow \frac{2}{|A|^2 - |A|} \sum_{i<j \in A} 1 - S(i, j)$
  $rdiv(A) \leftarrow 2 \cdot div(A) \cdot \overline{ds} - 1$
  $m(A) \leftarrow m'(A)(1 + \alpha \cdot rdiv(A))$
  if $m(A) > m(\{\sigma(i+1), \ldots, \sigma(n)\})$ **then**
    $m(A) \leftarrow m(\{\sigma(i+1), \ldots, \sigma(n)\})$
  end if
end for
Normalize m to have $m(\mathcal{N}) = 1$

After this step, an adjustment for satisfying monotonicity is performed. The pseudocode is given in Algorithm 5.

- Global MHM (G-MHM): the static approach does not consider any arrangement, since the whole measure must be constructed. Therefore, an extra step for enforcing monotonicity of the fuzzy measure must be added to the proposal. Observe that the measure constructed by this method is not an additive measure. The pseudocode is given in Algorithm 6.

3) OVERLAP INDEX-BASED FUZZY MEASURE (OIFM) [63]

This algorithm was originally given in [63] for constructing fuzzy measures for fuzzy rule-based classification systems. However, it can be easily adapted for calculating a fuzzy measure associated with the confidence degree of each classifier. The construction method is based on the use of overlap indices (see [64] for more details). These indices allow to measure the degree of overlaping between two fuzzy sets, assuming that the higher the membership degree of an element to both sets, the higher the overlap. For this purpose, the overlap index tries to measure the importance of the coalition in terms of confidence degree. Therefore, the higher the confidences of the coalition, the higher the value of its coefficient. This algorithm is global and we do not consider any dynamic counterpart. Moreover, the resulting measure is non-additive. The pseudocode can be seen in Algorithm 7.

Finally, in Figure 1 we show an schematic summarization of Algorithms 1-6, where the taxonomy allows to visually

Algorithm 6 G-MHM

Input: $\mathcal{N}$: Set of classifiers; $m' : 2^{\mathcal{N}} \to [0, 1]$ original fuzzy measure; S: similarity measure; $\alpha \in [0, 1]$: parameter.
Output: fuzzy measure: $m : 2^{\mathcal{N}} \to [0, 1]$.

$\overline{ds} \leftarrow \max_{i \neq j \in \mathcal{N}} S(i, j) = 1 - \min_{i \neq j \in \mathcal{N}} S(i, j)$
for each $A \subseteq \mathcal{N}$ **do**
  if $|A| \leq 1$ **then**
    $div(A) \leftarrow 0$
    $rdiv(A) \leftarrow 0$
  else
    $div(A) \leftarrow \frac{2}{|A|^2 - |A|} \sum_{i<j \in A} 1 - S(i, j)$
    $rdiv(A) \leftarrow 2 \cdot div(A) \cdot \overline{ds} - 1$
  end if
  $m(A) \leftarrow m'(A)(1 + \alpha \cdot rdiv(A))$
  $m(A) = \max_{B \subseteq A} m(B)$
end for
Normalize m to have $m(\mathcal{N}) = 1$

Algorithm 7 OIFM

Input: $\mathcal{N}$: Set of classifiers; $O : [0, 1]^N \times [0, 1]^N \to [0, 1]$ overlap index; κ: confidence vector;
Output: fuzzy measure: $m : 2^{\mathcal{N}} \to [0, 1]$.

Construct fuzzy set $E = \{(i, \kappa_i) | i = 1, \ldots, N\}$ associated to κ
for each $A \subseteq \mathcal{N}$ **do**
  $E_A = \{(i, 0) | i = 1, \ldots, N\}$
  for each $i \in A$ **do**
    $E_A(i) = \kappa_i$
  end for
  $m(A) \leftarrow O(E, E_A)$
end for
Normalize m so that $m(\mathcal{N}) = 1$

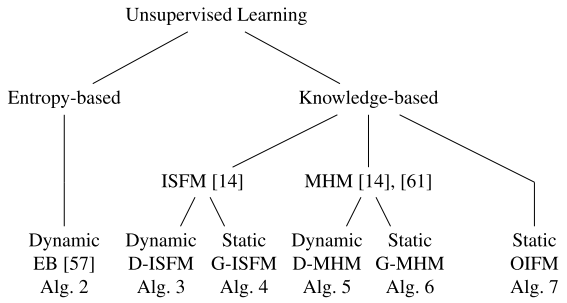


FIGURE 1. Diagram of algorithms for learning fuzzy measures in the context of ensemble aggregation.

identify the group and behavior (dynamic or global) each algorithm belongs to.

V. CPM: CLASSIFER PERFORMANCE-BASED MEASURE

Analyzing the fuzzy measures reviewed in the previous section, one can observe that only G-MHM constructs a global non-additive fuzzy measure. However, notice that in this global measure, the interaction among classifiers is a mere estimation based on a heuristic considering a pairwise similarity measure and a confidence of each individual classifier. The drawback is that interactions among groups of more than two classifiers are being estimated indirectly. Our idea is that we can measure these interactions directly testing how well they interact. Intuitively, directly measuring this degree of synergy should lead to a better measure.

Therefore, our main objective in this work is to propose a methodology to construct a global non-additive fuzzy measure that is based on directly estimating the quality of each coalition. Moreover, we want to show that the full potential behind fuzzy measures can be successfully used, yet efficiently. Our proposal is based on using classifiers' performance over the training set to construct the measure. In this section, we formalize our proposal for the calculation of the fuzzy measure, we show an illustrative example of its behavior and analyze its computational complexity.

A. INTUITION

Our aim with this fuzzy measure is to make each coefficient to truly reflect how well the classifiers are interacting. That is, the combination of certain classifiers could lead to a better solution, whereas adding a new classifier could not always result in an increase in performance as it may be expected from its individual performance. Although previous approaches could perform well in practice, the fuzzy measure constructed is mainly based on the individual confidence (which we measure as their performance) of each classifier and sometimes, either a pairwise measure of similarity or diversity that is considered to reflect how well a pair of classifiers may interact. However, they are measuring the interactions indirectly. The main novelty of the proposed fuzzy measure is that we are able to measure the interactions directly from data, by measuring the performance of each possible subset of classifiers.

The intuition for constructing the fuzzy measure is rather simple: the value of each coefficient should be conditioned by the performance of the specific sub-ensemble. Notice that the performance values cannot be directly used as coefficients as they will not probably result in a proper fuzzy measure.

For this reason, our construction method of the fuzzy measure starts from a uniform fuzzy measure $m_U : 2^{\mathcal{N}} \to [0, 1]$ given for every $A \subseteq \mathcal{N}$ by $m_U(A) = \frac{|A|}{N}$. Then, the construction methods continues level by level (we refer to the i-th level as every subset $A \subseteq \mathcal{N}$ with $|A| = i$). For the first level, we take the performance of each individual classifier and we calculate the average performance (obtained by the arithmetic mean). The value of the fuzzy measure of those classifiers whose performance is greater than the average will be increased (with respect to the value of uniform fuzzy partition). Conversely, those with lower performance will get their value of the fuzzy measure reduced. This is again performed for the second level, but now the performance of each possible pair of classifiers is taken into account, as well as the average performance of pair of each classifiers. This process is repeated until the top level is reached.

B. FORMAL DEFINITION OF THE FUZZY MEASURE

Formally, let $A \subseteq \mathcal{N}$ represent a subset of classifiers with P_A being be the performance of A in a classification problem. For each $i \in \{1, \ldots, n\}$, define μ_i as the average performance of those $B \subseteq \mathcal{N}$ with $|B| = i$, that is,

$$\mu_i = \frac{1}{k} \sum_{B \subseteq \mathcal{N}, |B|=i} P_B,$$

where $k = \binom{n}{i}$ is the number of sets in the ith level of the measure and $\mu_0 = 0$. Let $m_U : \{1, \ldots, n\} \to [0, 1]$ be the uniform fuzzy measure. Then, the fuzzy measure we propose is given by

$$m(A) = m_U(A) + \frac{\tanh(100 \cdot (P_A - \mu_{|A|}))}{2n} \tag{4}$$

where $\tanh : (-\infty, +\infty) \to (-1, 1)$ is the hiperbolic tangent function given by $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Theorem 1: The mapping $m : \{1, \ldots, n\} \to [0, 1]$ given in Eq. 4 is a non-additive fuzzy measure.

Proof: Boundary conditions are clear, since $P_\emptyset = \mu_0$ and $P_\mathcal{N} = \mu_\mathcal{N}$, yielding $m(\emptyset) = m_U(\emptyset) = 0$ and $m(\mathcal{N}) = m_U(\mathcal{N}) = 1$. To proof monotonicity, we have that for every $A \subset B \subseteq \mathcal{N}$, $m_U(B) - m_U(A) \geq 1/n$. However, due to the construction method $|\tanh(100 \cdot (P_A - \mu_{|A|})) - \tanh(100 \cdot (P_B - \mu_{|B|}))| \leq 1$ and, therefore, $m(B) - m(A) \in [0, 1/n]$ and monotonicity holds. ∎

Analyzing Eq. (4), we clearly see that $m = m_U$ only if $P_A = \mu_A$ for every $A \subseteq \mathcal{N}$, which only happens if all the classifiers perform in the same way. It can be easily seen that the value of the fuzzy measure associated with a good combination of classifiers (better than the average performance) will be increased, and those associated with bad coalitions will be decreased with respect to the uniform measure. In fact, we can easily prove the following result

Observe that the construction of the fuzzy measure proposed in this work is not unique, in the sense that we can generalize the method to consider any other performance measure. For instance, one can measure the performance of classifier by means of the accuracy rate (percentage of correctly classified examples). However, as we have mentioned earlier, this may not be a proper measure for the imbalance framework, where the GM may be more suitable.

Finally, in Algorithm 8 we describe the proposed CPM construction method.

C. ILLUSTRATIVE EXAMPLE

We consider a two-class problem ($|\mathcal{C}| = 2$) for which an ensemble with four classifiers ($N = 4$) has been learned. We will use the Choquet integral with the fuzzy measure obtained to show the final classification result with this measure. For illustrative purposes, we also consider the classification with the arithmetic mean as aggregation to show that using an adequate fuzzy measure together with the Choquet integral can lead to better classifications.

Table 2 shows the scores obtained by each one of the classifiers (namely, C_1, C_2, C_3 and C_4) for the 10 data instances (examples) in the problem, whose true labels are given in column y ($\hat{y}_\mu$ and $\hat{y}_C$ will be the outputs obtained using the arithmetic mean and the Choquet integral with the proposed measure, respectively). Each score represents the probability of each instance belonging to class 1 ($p(y = 1|x)$, thus $p(y = 0|x) = 1 - p(y = 1|x)$).

Algorithm 8 CPM

Input: $\mathcal{N}$: Set of classifiers; P_A: accuracy of any coalition $A \subseteq \mathcal{N}$; Performance measure p
Output: fuzzy measure: $m : 2^\mathcal{N} \to [0, 1]$.

1: **for** $A \subseteq \mathcal{N}$ **do**
2: $\quad m_U(A) = \frac{\sum_{i \in A} P_i}{\sum_{i=1}^n P_i}$
3: **end for**
4: **for** $A \subseteq \mathcal{N}$ **do** $P_A \leftarrow$ performance of A according to p
5: **end for**
6: **for** $i = 1, \ldots, N$ **do**
7: $\quad k \leftarrow \binom{n}{i}$
8: $\quad \mu_i \leftarrow \frac{1}{k} \sum_{B \subseteq \mathcal{N}, |B|=i} P_B$,
9: **end for**
10: **for** $A \subseteq \mathcal{N}$ **do**
11: $\quad m_(A) = m_U(A) \frac{\tanh(100(P_A - \mu_{|A|}))}{2n}$
12: **end for**

TABLE 2. Data.

Example	C_1	C_2	C_3	C_4	y	$\hat{y}_\mu$	$\hat{y}_C$
x_1	0.29	0.33	0.08	0.93	0	0	0
x_2	0.93	0.67	0.53	0.04	1	1	1
x_3	0.68	0.87	0.15	0.04	1	0	1
x_4	0.58	0.10	0.40	0.96	0	1	0
x_5	0.69	0.70	0.41	0.19	1	0	1
x_6	0.03	0.14	0.11	0.03	0	0	0
x_7	0.46	0.61	0.76	0.57	1	1	1
x_8	0.91	0.10	0.99	0.24	1	1	1
x_9	0.71	0.93	0.25	0.33	1	1	1
x_{10}	0.96	0.18	0.52	0.43	1	1	1

To construct the fuzzy measure, we need to iterate over all the possible sets of classifiers and compute their performance. To compute the performance of a set of classifiers, we will use the arithmetic mean to combine the scores of the classifiers in the set. Then, the class for each instance is predicted taking the class corresponding to the largest aggregated value. Finally, the performance is obtained by comparing the outputs of the sub-ensemble with the true targets. Following our example, we compute the performance of the set of classifiers of cardinality 1, 2, 3 and 4. These performances are presented in Table 3.

Given the data in Table 3 and Eq. 4, we are now able to construct the fuzzy measure. To see one of the steps, consider the value of the fuzzy measure associated with C_1. We have that

$$m(\{1\}) = \frac{1}{4} + \frac{\tanh(100(0.7559 - 0.6438))}{8} = 0.3750.$$

The final coefficients of the fuzzy measure obtained following this rule are presented in Figure 2.

TABLE 3. GM per node and level.

Sub-ensemble	GM	μ_{GM}
C_1	0.7559	0.6438
C_2	0.8452	
C_3	0.7559	
C_4	0.2182	
C_1, C_2	1.0000	0.6374
C_1, C_3	0.8452	
C_1, C_4	0.4364	
C_2, C_3	0.9258	
C_2, C_4	0.3086	
C_3, C_4	0.3086	
C_1, C_2, C_3	1.0000	0.6716
C_1, C_2, C_4	0.5345	
C_1, C_3, C_4	0.6172	
C_2, C_3, C_4	0.5345	
C_1, C_2, C_3, C_4	0.6901	0.6901

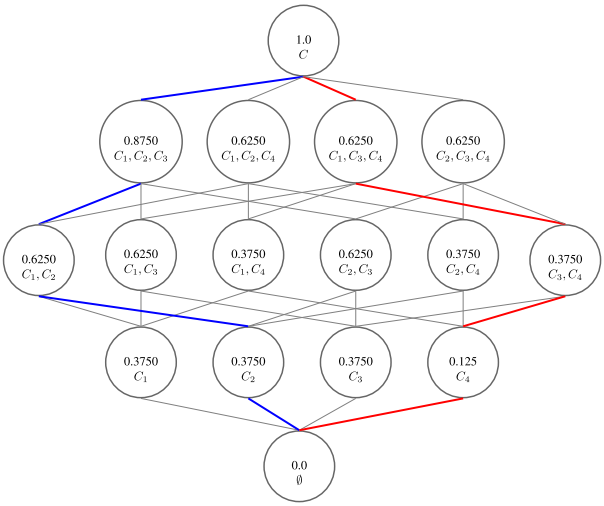


FIGURE 2. Fuzzy measure obtained applying the proposed construction method with the performances in Table 3.

Once the fuzzy measure has been constructed (see Figure 2), we can now predict the class of each of the instances using the Choquet integral insted of the arithmetic mean. For example, let us show the process of the third instance x_3. Recall that the probabilities of x_3 belonging to class 1 provided by each classifier are $p_{C_1}(y = 1|x_3) = 0.68$, $p_{C_2}(y = 1|x_3) = 0.87$, $p_{C_3}(y = 1|x_3) = 0.15$ and $p_{C_4}(y = 1|x_3) = 0.04$. Since

$$p_{C_4}(y = 1|x_3) < p_{C_3}(y = 1|x_3) < p_{C_1}(y = 1|x_3) < p_{C_2}(y = 1|x_3),$$

the coefficients of the fuzzy measure that play role in the classification of x_3 are $m(\{1, 2, 3, 4\})$, $m(\{1, 2, 3\})$, $m(\{1, 2\})$ and $m(\{2\})$ (see blue path in Figure 2). Thus, the aggregation of scores for class 1 is given by

$$\begin{aligned} &C_m(0.68, 0.87, 0.15, 0.04) \\ &= 0.04 \cdot m(\{1, 2, 3, 4\}) \\ &\quad + (0.15 - 0.04) \cdot m(\{1, 2, 3\}) + (0.68 - 0.15) \cdot m(\{1, 2\}) \\ &\quad + (0.87 - 0.68) \cdot m(\{2\}) = 0.53875. \end{aligned}$$

As mentioned earlier, we also need to compute the aggregation score for class 0. Here, the reversed order implies that coefficients that are taken into account are $m(\{1, 2, 3, 4\})$, $m(\{1, 3, 4\})$, $m(\{3, 4\})$ and $m(\{4\})$ ((see red path in Figure 2). Then,

$$\begin{aligned} &C_m(0.32, 0.13, 0.85, 0.96) \\ &= 0.13 \cdot m(\{1, 2, 3, 4\}) \\ &\quad + (0.32 - 0.13) \cdot m(\{1, 3, 4\}) + (0.85 - 0.32) \cdot m(\{3, 4\}) \\ &\quad + (0.96 - 0.85) \cdot m(\{4\}) = 0.46125. \end{aligned}$$

Therefore, since $0.53875 > 0.46125$, we predict class 1 for the third instance. Notice that in Figure 2, the coefficients applied in the aggregation of class 1 and class 0 are given by the blue and red paths, respectively. Otherwise, if we had considered the arithmetic mean, we would had obtained $\frac{0.68+0.87+0.15+0.04}{4} = 0.435$ for class 1, and $\frac{0.32+0.13+0.85+0.96}{4} = 0.565$ for class 0. Since $0.565 > 0.435$, class 0 would had been predicted. Notice that the true class label is 1 and hence, using CPM the correct output is predicted as the interactions among classifiers are taken into account, whereas using the arithmetic mean the example is incorrectly classified.

Of course, we can perform the same process with the rest of the examples and the results obtained are presented in Table 2. We recall that for the aggregation of each instance, we will consider a different set of coefficients of the fuzzy measure, which are determined by the ordering of the probability vector. We should note that using the arithmetic mean 7 out of 10 examples are correctly classified, whereas the Choquet integral with the proposed fuzzy measure is able to correctly classify all the examples. Obviously, this is an illustrative example and the benefit of Choquet-based aggregations should be appropriately tested with real-world problems, which is done in the empirical study in Section VII.

D. COMPUTATION COMPLEXITY

Although computational capacity increases continuously, the computational complexity of the algorithms should be reduced to the minimum to increase their applicability. In our case, CPM has a complexity of $\mathcal{O}(2^N|X|)$, where $|X|$ refers to the number of examples in the dataset. Notice that 2^N coefficients needs to be estimated (as in the rest of fuzzy measure-based methods) and the performance should be computed for each possible combination (complexity of $|X|$). However, notice that this process is only carried out once for global methods, whereas in dynamic ones it is performed for each instance. Anyway, recall that for dynamic methods the whole fuzzy measure need not be estimated. With respect to global methods such as G-ISFM and G-MHM, complexity is the same for larger number of classifiers (N), since $|X|$ can be disregarded. In the case of lower number of classifiers, G-ISFM and G-MHM compute pairwise similarities for all pairs of classifiers and hence, complexity is $\mathcal{O}(2^N + |X|^2)$. Although more complex, our method takes advantage of more

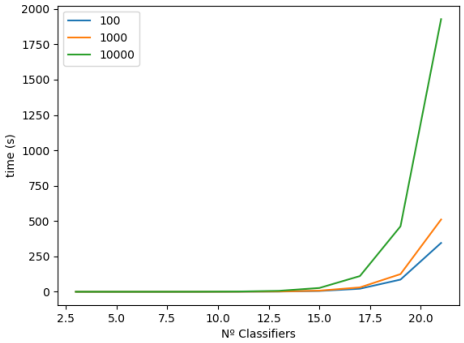


FIGURE 3. Time (s) requirements for constructing a fuzzy measure with CPM for varying number of classifiers and dataset sizes (100, 1000 and 10000 examples).

information as described earlier. Finally, recall that in global methods the fuzzy measure is estimated just once in the training procedure.

In Figure 3, we provide a study of the time required for the computation of the fuzzy measure with CPM for varying number of classifiers (from 3 to 21) and dataset sizes (100, 1000 and 10000 examples). Code is written in Python and only the code for computing the performance is slightly optimized with numba (no GPU or parallel approaches have been considered). As expected, an exponential increase in the computational complexity is found with greater number of classifiers, whereas complexity scales linearly with greater number of examples. Nevertheless, we want to stand out that all the previously proposed fuzzy measure construction methods were tested with much less classifiers in the ensemble than the number of classifiers we use in our experimental study which is 21. Additionally, one should take into account that methods for fuzzy measure compression [49] exist, which can help reducing the computational burden. Moreover, estimating only the necessary coefficients would highly improve the times required for estimating the fuzzy measure, although the usage of these optimizations is out of the scope of this paper.

VI. EXPERIMENTAL FRAMEWORK

In this section, we introduce the experimental framework considered for the empirical study in the next section. We first introduce the general details of our framework: the datasets considered, the base classifier for the ensemble, its parameters, performance measures and statistical tests (Section VI-A). Then, we summarize the aggregation methods selected for the experiments and how their parameters are obtained in Section VI-B. Finally, Section VI-C presents our research questions and details how we aim to answer them with the experiments carried out.

A. GENERAL SETTINGS

We have considered the set of sixty-six imbalanced datasets from KEEL dataset [18], which is a commonly considered benchmark for evaluating classifiers for the class imbalance problem [17], [20]. For each dataset, the total number of examples, number of attributes and IR (ratio between the majority and minority class examples) are presented in Table 4. Two-class imbalanced problems were obtained modifying originally multi-class problems by joining one or more classes as positive and doing the same for the negative one. Notice that although we restrict the experimental study to two-class imbalanced problem, the proposed method is also applicable to either balanced or multi-class problems. However, a challenging scenario like this one highlights the importance of the aggregation in ensembles.

TABLE 4. Summary description of the imbalanced datasets considered in this study.

No.	Data-sets	#Ex.	#Atts.	IR	No.	Data-sets	#Ex.	#Atts.	IR
1	glass1	214	9	1.82	34	Glass04vs5	92	9	9.22
2	ecoli-0_vs_1	220	7	1.86	35	Ecoli0346vs5	205	7	9.25
3	wisconsin	683	9	1.86	36	Ecoli0347vs56	257	7	9.28
4	pima	768	8	1.87	37	Yeast05679vs4	528	8	9.35
5	iris0	150	4	2	38	Ecoli067vs5	220	6	10.00
6	glass0	214	9	2.06	39	Vowel0	988	13	10.10
7	yeast1	1484	8	2.46	40	Glass016vs2	192	9	10.29
11	haberman	306	3	2.78	41	Glass2	214	9	10.39
8	vehicle2	846	18	2.88	42	Ecoli0147vs2356	336	7	10.59
9	vehicle1	846	18	2.9	43	Led7digit02456789vs1	443	7	10.97
10	vehicle3	846	18	2.99	44	Glass06vs5	108	9	11.00
12	glass-0-1-2-3_vs_4-5-6	214	9	3.2	45	Ecoli01vs5	240	6	11.00
13	vehicle0	846	18	3.25	46	Glass0146vs2	205	9	11.06
14	ecoli1	336	7	3.36	47	Ecoli0147vs56	332	6	12.28
16	new-thyroid1	215	5	5.14	48	Cleveland0vs4	177	13	12.62
15	new-thyroid2	215	5	5.14	49	Ecoli0146vs5	280	6	13.00
17	ecoli2	336	7	5.46	50	Ecoli4	336	7	13.84
18	segment0	2308	19	6.02	51	Yeast1vs7	459	8	13.87
19	glass6	214	9	6.38	52	Shuttlec0vsc4	1829	9	13.87
20	yeast3	1484	8	8.1	53	Glass4	214	9	15.47
21	ecoli3	336	7	8.6	54	Pageblocks13vs4	472	10	15.85
22	page-blocks0	5472	10	8.79	55	Abalone9vs18	731	8	16.68
23	ecoli-0-3-4_vs_5	200	7	9	56	Glass016vs5	184	9	19.44
24	yeast-2_vs_4	514	8	9.08	57	Shuttlec2vsc4	129	9	20.5
25	ecoli-0-6-7_vs_3-5	222	7	9.09	58	Yeast1458vs7	693	8	22.10
26	ecoli-0-2-3-4_vs_5	202	7	9.1	59	Glass5	214	9	22.81
27	glass-0-1-5_vs_2	172	9	9.12	60	Yeast2vs8	482	8	23.10
28	yeast-0-3-5-9_vs_7-8	506	8	9.12	61	Yeast4	1484	8	28.41
30	yeast-0-2-5-7-9_vs_3-6-8	1004	8	9.14	62	Yeast1289vs7	947	8	30.56
29	yeast-0-2-5-6_vs_3-7-8-9	1004	8	9.14	63	Yeast5	1484	8	32.78
31	ecoli-0-4-6_vs_5	203	6	9.15	64	Ecoli0137vs26	281	7	39.15
32	ecoli-0-1_vs_2-3-5	244	7	9.17	65	Yeast6	1484	8	39.15
33	ecoli-0-2-6-7_vs_3-5	224	7	9.18	66	Abalone19	4174	8	128.87

A 5-fold cross-validation scheme has been used to obtain the results for each method and dataset. We repeat this scheme 5 times in order to account for the randomness of the partitioning and the model construction. Hence, each result is computed by averaging the results over 25 runs.

The performance of the classifiers should be properly measured as we have already explained in Section III. Consequently, we consider the GM performance measure to evaluate the quality of the methods. We should notice that we have obtained similar conclusions using the Area Under the ROC Curve, AUC).

As recommended in the literature [21], we make use of non-parametric statistical tests to analyze the results obtained. When comparing a pair of methods, the Wilcoxon test is considered, whereas the Friedman aligned-ranks test is used when the comparison involves a group of methods. In the latter case, when significant differences are found, a *post-hoc* test should be performed to check whether the null hypothesis of equivalence between each method and the selected control method (the best one) is rejected. We consider Holm's test for this purpose.

Since we base our experimental framework on previous works [20], we consider C4.5 decision tree [4] as base classifier for our ensemble. With respect to UnderBagging, we recall that we use the variant that performed best

TABLE 5. Parameters for C4.5 and UnderBagging algorithm.

Algorithm	*Parameters*
C4.5	Prune = True, Confidence level = 0.25 Minimum number of item-sets per leaf = 2 Confidence = Laplace Smoothing
UnderBagging	Number of final classifiers = 21 Pruning method = RE_GM Pool of classifiers = 100

in [20], which is considered to obtain state-of-the-art results. As explained in Section III-B, this variant is UnderBagging_RE. The parameters considered for C4.5 and UnderBagging are summarized in Table 5.

B. SUMMARY OF AGGREGATION METHODS AND PARAMETER ESTIMATION

Our main focus in this paper is on fuzzy measure learning algorithms applied to classification problems. Accordingly, we consider all the methods described in Section IV for fuzzy measure construction, and its application with both the Choquet and the Sugeno integrals. We do the same with CPM, the proposed fuzzy measure construction method. Moreover, we consider its λ-measure and additive measure counterparts, where only the coefficients for the first level (singletons) are computed from data and the rest of the coefficients are obtained from them.

Otherwise, we believe that it is also interesting to consider classical aggregation functions in the comparison, both weighted and unweighted ones, so that we can study whether the usage of the more complex aggregations can significantly improve the most commonly used ones. In these cases, we need to detail how the parameters for weighted means and OWA operators are obtained.

All the methods in the empirical study are summarized in Table 6, where the family of the method, abbreviation and description are presented. Notice that we assign each method to a family so that we can carry out intra- and inter-family comparisons (explained next). We also highlight whether each method can be categorized as global (G) or dynamic (D) and, if the method is based on learning a global fuzzy measure, we also show if the corresponding measure is additive (+), Sugeno λ-measure (λ) or non-additive (−).

The first method in Table 6, O (Original), makes reference to the aggregation used in [20] for UnderBagging_RE. This aggregation is similar to the AM in spirit, but only the probabilities for the predicted class are summed for each classifier (not averaged).

The setting-up of the parameters of the weighted aggregation functions shown in Table 6 is as follows. For the weighted arithmetic mean, each weight of the corresponding weighting vector is constructed from the normalized performance obtained by each classifier in the the training dataset. For example, if $P_1, \ldots, P_N$ are the performance of each of the N classifiers, then $w_i = \frac{P_i}{\sum_{j=1}^{N} P_j}$ for all $i \in \{1, \ldots, N\}$. Since we focus on imbalanced datasets, we use the geometric mean (GM) as performance measure in order to better model

TABLE 6. Methods considered for the comparison.

Family	Abb.	Description	Behavior
Unweighted	O	Original	G
	AM	Arithmetic Mean	G
	MED	Median	G
	GM	Geometric Mean	G
	HM	Harmonic Mean	G
	MIN	Minimum	G
	MAX	Maximum	G
Weighted	WAM	Weighted Arithmetic Mean	G
	WGM	Weighted Geometric Mean	G
	WHM	Weighted Harmonic Mean	G
	Q_{alh}	OWA using Q_{alh}	D
	Q_{amap}	OWA using Q_{amap}	D
	Q_{mot}	OWA using Q_{mot}	D
Choquet	C_m	Choquet using CPM	G(−)
	C_{m+}	Choquet using additive CPM	G(+)
	$C_{m\lambda}$	Choquet using λ-measure	G(λ)
	C_{OIFM}	Choquet using OIFM measure	G(−)
	$C_{G\text{-}ISFM}$	Choquet using G-ISFM measure	G(+)
	$C_{D\text{-}ISFM}$	Choquet using D-ISFM measure	D
	$C_{D\text{-}MHM}$	Choquet using D-MHM measure	D
	$C_{G\text{-}MHM}$	Choquet using G-MHM measure	G(−)
	C_{EB}	Choquet using EB measure	D
Sugeno	S_m	Sugeno using CPM	G(−)
	S_{m+}	Sugeno using additive CPM	G(+)
	$S_{m\lambda}$	Sugeno using λ-measure	G(λ)
	S_{OIFM}	Sugeno using OIFM measure	G(−)
	$S_{G\text{-}ISFM}$	Sugeno using G-ISFM measure	G(+)
	$S_{D\text{-}ISFM}$	Sugeno using D-ISFM measure	D
	$S_{D\text{-}MHM}$	Sugeno using D-MHM measure	D
	$S_{G\text{-}MHM}$	Sugeno using G-MHM measure	G(−)
	S_{EB}	Sugeno using EB measure	D

the quality of each classifier. Notice that in this case, the individual quality of each classifier is only taken into account, and no information is added on how they interact with each other. With respect to OWA operators, each weighting vectors are induced by the fuzzy quantifiers shown in Example 2.

For fuzzy measure-based methods, classifier confidence and performance are considered to be the same value and are given by the GM value obtained in the training set. We consider the GM as we are dealing with class imbalance problems. This score is obtained from training set performance as it is done for classifier pruning [20]. Notice that further dividing the training set into two sets for training and fuzzy measure coefficients' estimation leads to a loss of performance of the ensemble that cannot be recovered by the aggregation. For CPM, we will also consider its additive and $\lambda-$measure [65], [66] counterparts. In these cases, we only need to establish the first level of the fuzzy measure, since the rest of the fuzzy measure can be derived from them. To have a fair comparison, the first level of the fuzzy measure will be obtained in the same way as in the proposed CPM fuzzy measure.

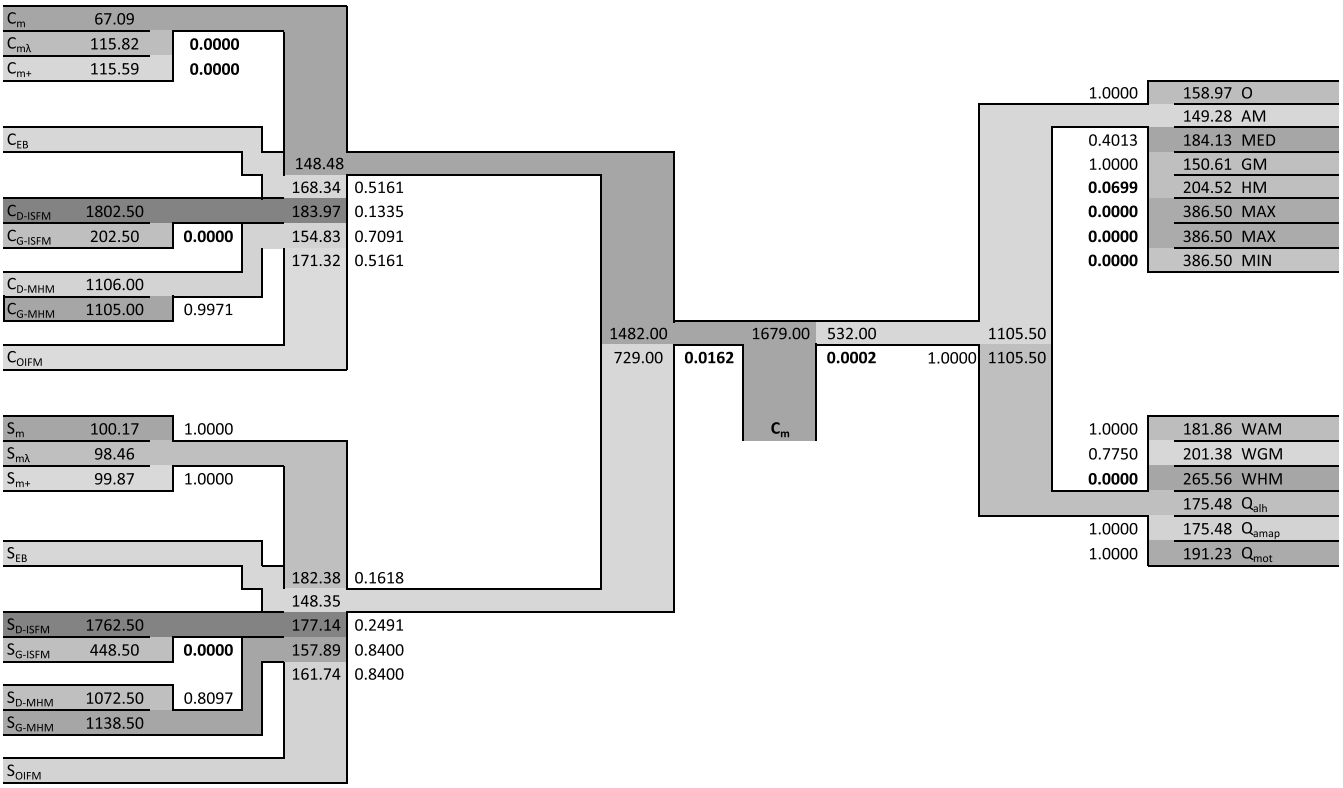


FIGURE 4. Hierarchical statistical study comparing the fusion functions in each family and the best performers of each family for each performance measure using Friedman Aligned ranks test.

C. RESEARCH QUESTIONS AND DESCRIPTION OF THE EXPERIMENTS

With the empirical study, we want to answer six main questions:

- Can global non-additive fuzzy measures outperform its simpler counterparts? (λ-measure and additive measure).
- Do dynamic methods for fuzzy measure estimation perform better than global ones?
- Is CPM capable of improving existing alternatives?
- Which fuzzy integral performs better? (Choquet vs. Sugeno).
- Are fuzzy measure-based methods able to significantly improve the performance of classical aggregation functions?
- Can the usage of fuzzy measures allow one to improve the state-of-the-art results on ensembles for classification of imbalanced data?

We will perform a hierarchical empirical analysis of the results obtained to answer this questions one by one. Notice that all of them are relevant questions. With the first one, we will analyze whether the flexibility of a global non-aditive fuzzy measure really allows for an improvement in the final classification accuracy (Section VII-A). Similarly, we focus on comparing global and dynamic approaches and looking for the best existing alternative (Section VII-B). One can expect a better performance from dynamic approaches due to their greater adaptability to each example. However, as already mentioned, we believe that they are more like "dynamic OWAs" rather than fuzzy measures. Afterwards, we focus on evaluating the quality of our proposal (Section VII-C) and compare the two most widely used fuzzy integrals (Section VII-D). Finally, we will test whether fuzzy measure-based approaches are able to overcome classical aggregations in this task, checking if they are also able to improve state-of-the-art performance. (Section VII-E)

VII. EXPERIMENTAL STUDY

In this section, we carry out the empirical study aiming to answer the questions raised in Section VI-C. Before going through them one by one in the next sections, we first present the testing performance of each method in Table 7. One key-point in these results is that their differences are only due to the aggregation method considered. All of them get the same outputs from the base classifiers of UnderBagging_RE and the only difference among them is how these outputs are combined. This allows us to only focus on the differences of classifier aggregation, leaving apart other issues that could be caused by random effects in the ensemble construction. As mentioned earlier, we should notice that the same conclusions holds using other standard metrics for class imbalance problems such as AUC, but we focus on GM for brevity.

TABLE 7. Average GM score of each method for each dataset.

	Unweighted							Weighted						Choquet										Sugeno									
Dataset	O	AM	MED	GM	HM	MAX	MIN	WAM	WGM	WHM	OWA$_{alh}$	OWA$_{amap}$	OWA$_{mot}$	C$_m$	C$_\lambda$	C$_m$+	C$_{EB}$	C$_{D-ISFM}$	C$_{G-ISFM}$	C$_{D-MHM}$	C$_{G-MHM}$	C$_{OIFM}$	S$_m$	S$_\lambda$	S$_m$+	S$_{EB}$	S$_{D-ISFM}$	S$_{G-ISFM}$	S$_{D-MHM}$	S$_{G-MHM}$	S$_{OIFM}$
glass1	0.7880	0.7834	0.7832	0.7775	0.7791	0.7451	0.7451	0.7812	0.7768	0.7809	0.7834	0.7834	0.7843	0.7812	0.7854	0.7854	0.7731	0.7798	0.7797	0.7809	0.7809	0.7820	0.7832	**0.7887**	**0.7887**	0.7814	0.7810	0.7845	0.7857	0.7857	0.7874
ecoli-0_vs_1	0.9782	0.9817	0.9767	0.9803	0.9796	0.9707	0.9707	0.9817	0.9803	0.9796	0.9817	0.9817	0.9782	0.9817	**0.9824**	**0.9824**	0.9803	0.9776	0.9708	0.9739	0.9739	0.9817	0.9767	0.9795	0.9795	0.9768	0.9783	0.9708	0.9718	0.9718	0.9767
wisconsin	0.9713	0.9705	0.9707	0.9695	0.9668	0.9637	0.9637	0.9705	0.9692	0.9668	0.9705	0.9705	0.9712	0.9705	0.9699	0.9699	0.9703	0.9698	0.9548	0.9701	0.9701	0.9701	0.9707	0.9703	0.9703	0.9706	0.9686	0.9546	0.9712	0.9712	**0.9720**
pima	0.7490	0.7507	0.7475	0.7514	0.7511	0.7468	0.7468	**0.7516**	0.7509	0.7504	0.7507	0.7507	0.7492	**0.7516**	0.7481	0.7481	0.7467	0.7497	0.7386	0.7488	0.7488	0.7511	0.7475	0.7506	0.7503	0.7459	0.7455	0.7413	0.7453	0.7453	0.7494
iris0	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**	**0.9897**
glass0	0.8459	0.8368	0.8435	0.8331	0.8216	0.7824	0.7824	0.8391	0.8339	0.8220	0.8368	0.8368	0.8425	0.8370	0.8415	0.8415	0.8378	0.8345	0.8324	0.8371	0.8371	0.8364	0.8435	**0.8469**	0.8454	0.8377	0.8421	0.8253	0.8403	0.8403	0.8466
yeast1	**0.7341**	0.7282	0.7320	0.7307	0.7251	0.7093	0.7093	0.7286	0.7299	0.7252	0.7282	0.7282	0.7319	0.7292	0.7296	0.7296	0.7304	0.7282	0.7252	0.7294	0.7294	0.7294	0.7320	0.7321	0.7323	0.7294	0.7322	0.7253	0.7325	0.7325	0.7299
vehicle2	0.9747	0.9730	**0.9749**	0.9691	0.9513	0.9251	0.9251	0.9733	0.9691	0.9513	0.9730	0.9730	0.9738	0.9736	0.9725	0.9725	0.9696	0.9718	0.9582	0.9726	0.9726	0.9731	**0.9749**	0.9731	0.9731	0.9713	0.9739	0.9580	0.9738	0.9738	0.9739
vehicle1	0.7924	0.7915	0.7927	0.7899	0.7789	0.7484	0.7484	0.7924	0.7891	0.7791	0.7915	0.7915	0.7901	0.7906	0.7920	0.7920	0.7896	0.7902	0.7857	0.7896	0.7896	0.7907	0.7927	0.7917	0.7917	0.7921	0.7925	0.7829	**0.7936**	**0.7936**	0.7907
vehicle3	0.7917	0.7918	0.7920	0.7893	0.7833	0.7491	0.7491	0.7917	0.7884	0.7833	0.7918	0.7918	0.7937	0.7921	0.7934	0.7934	0.7882	0.7900	0.7878	0.7906	0.7906	0.7907	0.7920	0.7887	0.7887	0.7892	0.7915	0.7831	0.7918	0.7918	**0.7950**
haberman	0.6440	0.6448	0.6478	0.6398	0.6367	0.6287	0.6287	0.6432	0.6371	0.6344	0.6448	0.6448	0.6479	0.6481	0.6414	0.6414	0.6516	0.6405	0.6307	0.6481	0.6481	0.6390	0.6478	0.6456	0.6456	**0.6569**	0.6388	0.6331	0.6429	0.6429	0.6486
glass-0-1-2-3_vs_4-5-6	0.9327	0.9345	0.9327	0.9379	0.9313	0.9140	0.9140	0.9345	0.9379	0.9313	0.9345	0.9345	0.9346	0.9348	0.9298	0.9298	0.9391	**0.9398**	0.9128	0.9372	0.9372	0.9350	0.9327	0.9277	0.9271	0.9364	0.9327	0.9128	0.9284	0.9284	0.9321
vehicle0	0.9543	0.9542	0.9546	0.9519	0.9428	0.9190	0.9190	0.9542	0.9521	0.9428	0.9542	0.9542	0.9552	0.9542	0.9527	0.9527	0.9514	0.9550	0.9387	0.9556	0.9556	0.9542	0.9546	0.9524	0.9524	0.9530	0.9570	0.9379	**0.9583**	**0.9583**	0.9556
ecoli1	0.8994	0.9044	0.8952	0.9075	0.9107	0.9046	0.9046	0.9048	0.9083	**0.9111**	0.9044	0.9044	0.9016	0.9060	0.9070	0.9070	0.8990	0.9056	0.8934	0.9011	0.9011	0.9048	0.8952	0.9040	0.9040	0.8970	0.8984	0.8856	0.8914	0.8914	0.8987
new-thyroid2	0.9498	0.9589	0.9498	0.9620	0.9537	0.9082	0.9082	0.9589	0.9626	0.9543	0.9589	0.9589	0.9589	0.9583	0.9481	0.9481	0.9615	**0.9667**	0.9582	0.9586	0.9586	0.9583	0.9498	0.9492	0.9492	0.9596	0.9650	0.9582	0.9609	0.9609	0.9498
new-thyroid1	0.9476	0.9491	0.9476	0.9444	0.9435	0.9116	0.9116	0.9491	0.9444	0.9441	0.9491	0.9491	0.9461	0.9495	0.9496	0.9496	0.9444	**0.9554**	0.9472	0.9436	0.9436	0.9491	0.9476	0.9451	0.9451	0.9444	0.9519	0.9472	0.9410	0.9410	0.9459
ecoli2	0.9040	0.9030	**0.9054**	0.8954	0.8746	0.8382	0.8382	0.9037	0.8957	0.8746	0.9030	0.9030	0.9040	0.9003	0.8979	0.8979	0.8996	0.8967	0.8865	0.9005	0.9005	0.9013	**0.9054**	0.9028	0.9028	0.8981	0.8997	0.8830	0.9043	0.9043	0.9004
segment0	0.9897	0.9881	0.9896	0.9855	0.9806	0.9779	0.9779	0.9881	0.9855	0.9806	0.9881	0.9881	0.9883	0.9882	0.9887	0.9887	0.9867	0.9866	0.9868	0.9891	0.9891	0.9880	0.9896	**0.9899**	**0.9899**	0.9868	0.9893	0.9868	0.9894	0.9894	0.9895
glass6	0.9244	0.9243	0.9239	0.9233	0.9151	0.8464	0.8464	0.9249	0.9233	0.9162	0.9243	0.9243	0.9243	0.9249	0.9244	0.9244	0.9228	0.9181	0.9189	0.9207	0.9207	**0.9259**	0.9239	0.9244	0.9239	0.9238	0.9204	0.9184	0.9208	0.9208	0.9244
yeast3	0.9329	0.9322	0.9329	0.9322	0.9321	0.9240	0.9240	0.9322	0.9320	0.9322	0.9322	0.9322	0.9325	0.9326	0.9311	0.9311	0.9327	0.9310	0.9200	0.9321	0.9321	0.9316	0.9329	0.9305	0.9305	0.9331	0.9316	0.9193	0.9307	0.9307	**0.9332**
ecoli3	0.8626	0.8666	0.8633	0.8656	0.8671	0.8308	0.8308	0.8670	0.8629	0.8675	0.8666	0.8666	0.8670	0.8667	0.8600	0.8600	0.8650	0.8653	0.8430	0.8649	0.8649	**0.8689**	0.8633	0.8536	0.8536	0.8651	0.8651	0.8357	0.8595	0.8595	0.8662
page-blocks0	**0.9640**	0.9629	0.9638	0.9620	0.9572	0.9492	0.9492	0.9629	0.9620	0.9572	0.9629	0.9629	0.9630	0.9628	0.9627	0.9627	0.9613	0.9621	0.9532	0.9627	0.9627	0.9626	0.9638	0.9639	0.9639	0.9618	0.9627	0.9520	0.9629	0.9629	0.9639
ecoli-0-3-4_vs_5	0.8900	0.8853	0.8900	0.8810	0.8787	0.8152	0.8152	0.8865	0.8808	0.8777	0.8853	0.8853	0.8859	0.8865	0.8772	0.8772	0.8804	0.9127	0.8502	0.9067	0.9067	0.8792	0.8900	0.8733	0.8733	0.8810	**0.9130**	0.8500	0.9064	0.9064	0.8744
yeast-2_vs_4	0.9358	0.9376	0.9358	0.9385	0.9221	0.9120	0.9120	0.9378	0.9387	0.9223	0.9376	0.9376	0.9363	0.9381	0.9370	0.9370	0.9368	0.9398	0.9410	**0.9418**	**0.9418**	0.9378	0.9358	0.9386	0.9386	0.9374	0.9413	0.9410	0.9397	0.9397	0.9361
ecoli-0-6-7_vs_3-5	0.8562	0.8556	0.8562	0.8540	0.8586	0.8385	0.8385	0.8565	0.8538	**0.8596**	0.8556	0.8556	0.8570	0.8556	0.8573	0.8573	0.8539	0.8544	0.8311	0.8583	0.8583	0.8564	0.8562	0.8587	0.8587	0.8552	0.8503	0.8297	0.8550	0.8550	0.8582
ecoli-0-2-3-4_vs_5	0.8976	0.9016	0.8976	0.8998	0.8936	0.8178	0.8178	0.9027	0.9009	0.8952	0.9016	0.9016	0.9022	0.9027	0.8947	0.8947	0.9003	0.9028	0.8872	**0.9044**	**0.9044**	0.8991	0.8976	0.8725	0.8725	0.8959	0.8967	0.8913	0.9001	0.9001	0.8924
glass-0-1-5_vs_2	0.6820	0.7350	0.6722	0.7378	0.7104	0.4665	0.4665	0.7209	0.7358	0.7125	0.7350	0.7350	0.7245	0.7165	0.7121	0.7121	0.7394	0.7269	0.6865	**0.7417**	0.7338	0.6986	0.6722	0.7226	0.7104	0.7370	0.7065	0.6723	0.7122	0.7122	0.7032
yeast-0-3-5-9_vs_7-8	0.7225	**0.7355**	0.6975	0.7321	0.7271	0.7014	0.7014	0.7334	0.7308	0.7279	**0.7355**	**0.7355**	0.7318	0.7333	0.7234	0.7236	0.7300	0.7245	0.7151	0.7235	0.7235	0.7266	0.6975	0.7167	0.7169	0.7290	0.7192	0.7113	0.7144	0.7144	0.7278
yeast-0-2-5-6_vs_3-7-8-9	0.8103	0.8119	0.8094	0.8118	0.8062	0.7818	0.7818	0.8111	0.8124	0.8053	0.8119	0.8119	0.8092	0.8128	0.8099	0.8099	0.8072	**0.8159**	0.8050	0.8127	0.8127	0.8126	0.8094	0.8087	0.8098	0.8055	0.8133	0.8040	0.8115	0.8115	0.8120
yeast-0-2-5-7-9_vs_3-6-8	0.9137	0.9137	0.9142	0.9124	0.9028	0.8767	0.8767	0.9137	0.9123	0.9031	0.9137	0.9137	0.9146	0.9140	0.9135	0.9135	0.9132	0.9049	0.8950	0.9086	0.9086	0.9128	0.9142	0.9128	0.9128	**0.9149**	0.9058	0.8944	0.9076	0.9076	0.9135
ecoli-0-4-6_vs_5	0.9010	0.9000	0.9010	0.9087	0.9196	0.8320	0.8320	0.9000	0.9042	0.9233	0.9000	0.9000	0.9011	0.9075	0.8930	0.8930	0.9157	**0.9269**	0.8586	0.9247	0.9247	0.9012	0.9010	0.8799	0.8799	0.9054	0.9252	0.8569	0.9202	0.9202	0.8958
ecoli-0-1_vs_2-3-5	0.8972	0.8923	0.9003	0.8929	0.8949	0.8295	0.8295	0.8927	0.8920	0.8945	0.8923	0.8923	0.8943	0.8927	0.8945	0.8950	0.8900	0.8866	0.8823	0.8882	0.8882	0.8874	0.9003	**0.9026**	**0.9026**	0.8943	0.8904	0.8819	0.8925	0.8925	0.8908
ecoli-0-2-6-7_vs_3-5	0.8567	0.8553	0.8568	0.8525	0.8555	0.8290	0.8290	0.8558	0.8528	0.8555	0.8553	0.8553	0.8572	0.8557	0.8551	0.8551	0.8513	0.8558	0.8132	0.8563	0.8563	0.8568	0.8568	0.8557	0.8557	0.8532	0.8502	0.8233	**0.8634**	**0.8634**	0.8581
glass-0-4_vs_5	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	0.9713	0.9713	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	0.9822	0.9539	0.9822	0.9822	**0.9939**	**0.9939**	**0.9939**	**0.9939**	**0.9939**	0.9822	0.9539	0.9822	0.9822	**0.9939**
ecoli-0-3-4-6_vs_5	0.8822	0.8860	0.8818	0.9000	0.8919	0.8054	0.8054	0.8865	0.8956	0.8936	0.8860	0.8860	0.8818	0.8867	0.8833	0.8833	0.9010	**0.9074**	0.8478	0.8975	0.8975	0.8818	0.8818	0.8843	0.8843	0.8960	0.9051	0.8522	0.9053	0.9053	0.8818
ecoli-0-3-4-7_vs_5-6	0.8901	0.8896	0.8905	0.8832	0.8544	0.8053	0.8053	0.8900	0.8837	0.8515	0.8896	0.8896	0.8908	0.8908	0.8890	0.8890	0.8875	0.8723	0.8672	0.8780	0.8780	0.8887	0.8905	**0.8933**	**0.8933**	0.8883	0.8717	0.8672	0.8684	0.8684	0.8893
yeast-0-5-6-7-9_vs_4	0.8133	0.8193	0.8152	0.8230	0.8253	0.8163	0.8163	0.8164	0.8254	0.8257	0.8193	0.8193	0.8179	0.8165	0.8220	0.8220	0.8182	0.8256	0.8127	0.8261	0.8261	0.8181	0.8152	0.8145	0.8145	0.8218	0.8278	0.8114	**0.8301**	**0.8301**	0.8159
ecoli-0-6-7_vs_5	0.8723	0.8777	0.8723	0.8743	0.8716	0.8285	0.8285	0.8781	0.8742	0.8726	0.8777	0.8777	0.8792	0.8781	0.8729	0.8729	0.8750	**0.8893**	0.8567	0.8835	0.8835	0.8755	0.8723	0.8759	0.8759	0.8777	0.8889	0.8519	0.8840	0.8840	0.8711
vowel0	0.9691	0.9676	0.9692	0.9650	0.9570	0.9430	0.9430	0.9678	0.9648	0.9570	0.9676	0.9676	0.9682	0.9682	0.9686	0.9686	0.9662	0.9667	0.9630	0.9672	0.9672	0.9679	0.9692	**0.9696**	**0.9696**	0.9666	0.9668	0.9630	0.9684	0.9684	0.9692
glass-0-1-6_vs_2	0.7061	0.7178	0.6766	**0.7386**	0.7234	0.5030	0.5030	0.7092	0.7333	0.7228	0.7178	0.7178	0.6984	0.7085	0.6919	0.6919	0.7147	0.7147	0.6186	0.7291	0.7291	0.7173	0.6766	0.6700	0.6676	0.7165	0.6838	0.6515	0.6950	0.6950	0.6982
glass2	0.7223	0.7092	0.7264	0.7104	0.7385	0.4861	0.4861	0.7109	0.7060	0.7408	0.7092	0.7092	0.7089	0.7168	0.7041	0.7041	0.7044	0.6997	**0.7462**	0.6996	0.6996	0.7129	0.7264	0.7247	0.7250	0.7053	0.7037	0.7456	0.7016	0.7016	0.7099
ecoli-0-1-4-7_vs_2-3-5-6	0.8698	0.8614	0.8729	0.8549	0.8262	0.7646	0.7646	0.8617	0.8562	0.8259	0.8614	0.8614	0.8662	0.8623	0.8574	0.8574	0.8538	0.8522	0.8787	0.8501	0.8501	0.8617	0.8729	0.8675	0.8675	0.8565	0.8542	**0.8788**	0.8552	0.8552	0.8664
led7digit-0-2-4-5-6-7-8-9_vs_1	0.8228	0.8092	0.8257	0.8120	0.8163	**0.8389**	**0.8389**	0.8134	0.8119	0.8168	0.8092	0.8092	0.8193	0.8104	0.8217	0.8217	0.8267	0.8186	0.8270	0.8195	0.8195	0.8190	0.8257	0.8271	0.8271	0.8276	0.8155	0.8288	0.8173	0.8173	0.8230
glass-0-6_vs_5	0.9919	0.9919	0.9919	0.9842	0.9319	0.8731	0.8731	0.9929	0.9885	0.9372	0.9919	0.9919	0.9919	0.9929	0.9939	0.9939	0.9852	0.9646	0.9529	0.9772	0.9772	0.9939	0.9919	0.9939	0.9939	0.9852	0.9243	0.9529	0.9835	0.9835	**0.9949**
ecoli-0-1_vs_5	0.9024	0.9075	0.9076	0.9071	0.9024	0.8271	0.8271	0.9075	0.9013	0.9034	0.9075	0.9075	0.9075	0.9075	0.8941	0.8941	0.9018	0.9401	0.8954	0.9336	0.9336	0.8985	0.9076	0.8992	0.8992	0.9018	**0.9406**	0.8993	0.9381	0.9381	0.9042
glass-0-1-4-6_vs_2	0.7394	0.7391	0.7321	0.7313	0.7507	0.5048	0.5048	0.7148	0.7158	0.7386	0.7391	0.7391	0.7313	0.7343	0.7026	0.7026	0.7275	0.7163	0.7293	0.7437	0.7437	0.7127	0.7321	0.7141	0.7136	0.7212	0.7139	0.7282	**0.7514**	**0.7514**	0.7099
ecoli-0-1-4-7_vs_5-6	0.8929	0.8908	0.8900	0.8767	0.8634	0.7995	0.7995	0.8908	0.8743	0.8637	0.8908	0.8908	0.8929	0.8914	0.8839	0.8839	0.8808	0.8777	0.8596	0.8811	0.8811	0.8918	0.8900	0.8773	0.8773	0.8908	0.8819	0.8560	0.8882	0.8882	**0.8969**
cleveland-0_vs_4	0.8448	0.8538	0.8459	0.8710	0.8619	0.7037	0.7037	0.8544	0.8638	0.8559	0.8538	0.8538	0.8540	0.8543	0.8170	0.8170	0.8839	**0.9019**	0.7509	0.8912	0.8912	0.8279	0.8459	0.8145	0.8145	0.8675	0.8833	0.8132	0.8831	0.8831	0.8276
ecoli-0-1-4-6_vs_5	0.8915	0.9008	0.8915	0.9097	0.9031	0.7969	0.7969	0.8959	0.9104	0.9090	0.9008	0.9008	0.8905	0.9019	0.8960	0.8960	0.9105	0.9190	0.9019	**0.9227**	**0.9227**	0.9032	0.8915	0.8927	0.8927	0.9056	0.9202	0.8942	0.9159	0.9159	0.8972
ecoli4	0.9230	0.9222	0.9232	0.9302	0.9269	0.8231	0.8231	0.9228	0.9315	0.9273	0.9222	0.9222	0.9221	0.9218	0.9194	0.9194	0.9319	0.9294	0.8952	0.9263	0.9263	0.9237	0.9232	0.9186	0.9186	**0.9324**	0.9240	0.8959	0.9260	0.9260	0.9248
yeast-1_vs_7	0.7688	0.7844	0.7410	0.7708	0.7578	0.7116	0.7116	0.7829	0.7742	0.7599	0.7844	0.7844	0.7724	**0.7888**	0.7715	0.7715	0.7826	0.7597	0.7729	0.7683	0.7683	0.7760	0.7410	0.7501	0.7501	0.7749	0.7745	0.7695	0.7795	0.7795	0.7528
shuttle-c0-vs-c4	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
glass4	0.8834	0.8789	0.8839	0.8840	0.8956	0.7110	0.7110	0.8804	0.8731	0.8972	0.8789	0.8789	0.8799	0.8820	0.8808	0.8808	0.8767	0.8541	0.8189	**0.9060**	**0.9060**	0.8810	0.8839	0.8839	0.8839	0.8771	0.8778	0.8189	0.8915	0.8915	0.8820
page-blocks-1-3_vs_4	0.9941	0.9930	0.9941	0.9898	0.9691	0.8684	0.8684	0.9932	0.9903	0.9693	0.9930	0.9930	0.9930	0.9932	0.9952	0.9952	0.9912	0.9751	**0.9971**	0.9778	0.9778	0.9941	0.9941	0.9952	0.9952	0.9921	0.9738	**0.9971**	0.9790	0.9790	0.9948
abalone9-18	0.7358	0.7353	0.7120	0.7381	0.7266	0.7210	0.7210	0.7369	0.7375	0.7283	0.7353	0.7353	0.7374	0.7408	0.7457	0.7457	0.7449	**0.7496**	0.7336	0.7464	0.7464	0.7442	0.7120	0.7287	0.7286	0.7379	0.7279	0.7268	0.7267	0.7267	0.7364
glass-0-1-6_vs_5	0.9684	0.9683	0.9582	0.9618	0.9511	0.7164	0.7164	0.9701	0.9636	0.9511	0.9683	0.9683	0.9683	**0.9731**	**0.9731**	**0.9731**	0.9690	0.9496	0.9541	0.9582	0.9582	0.9707	0.9582	0.9689	0.9683	0.9690	0.9453	0.9541	0.9545	0.9545	0.9707
shuttle-c2-vs-c4	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.8241	0.8241	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.9669	0.9883	0.9669	0.9669	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.9669	0.9883	0.9669	0.9669	**1.0000**
yeast-1-4-5-8_vs_7	0.6222	0.6345	0.5387	0.6341	0.5995	0.5433	0.5433	0.6353	0.6300	0.6011	0.6345	0.6345	0.6326	**0.6422**	0.6347	0.6347	0.6233	0.6186	0.6203	0.6325	0.6394	0.6184	0.5387	0.6103	0.5984	0.6330	0.6273	0.6170	0.6187	0.6208	0.6113
glass5	0.9781	0.9781	0.9766	0.9766	0.9508	0.8586	0.8586	**0.9817**	0.9766	0.9508	0.9781	0.9781	0.9796	**0.9817**	0.9796	0.9796	0.9796	0.9216	0.9644	0.9608	0.9608	0.9791	0.9766	0.9796	0.9791	0.9796	0.9102	0.9644	0.9224	0.9224	0.9771
yeast-2_vs_8	0.7357	0.7713	0.7256	0.7858	0.7829	0.7690	0.7690	0.7696	0.7842	0.7823	0.7713	0.7713	0.7348	0.7710	0.7761	0.7761	0.7626	0.7738	0.7698	0.7778	0.7778	0.7744	0.7256	0.7559	0.7559	0.7402	**0.7899**	0.7678	0.7814	0.7814	0.7465
yeast4	**0.8580**	0.8560	0.8556	**0.8580**	0.8495	0.8415	0.8415	0.8555	0.8577	0.8494	0.8560	0.8560	0.8525	0.8569	0.8569	0.8569	0.8543	0.8565	0.8287	0.8576	0.8576	0.8570	0.8556	0.8526	0.8526	0.8562	0.8506	0.8302	0.8569	0.8569	0.8508
yeast-1-2-8-9_vs_7	0.7060	0.7298	0.6787	0.7288	0.7059	0.6493	0.6493	0.7283	0.7257	0.7062	0.7298	0.7298	0.7032	0.7301	0.7330	0.7330	**0.7383**	0.7208	0.7130	0.7296	0.7296	0.7247	0.6787	0.7004	0.7003	0.7157	0.7173	0.6996	0.7239	0.7239	0.7195
yeast5	0.9569	0.9571	0.9554	0.9643	0.9630	0.9363	0.9363	0.9573	**0.9644**	0.9632	0.9571	0.9571	0.9580	0.9572	0.9550	0.9550	0.9609	0.9582	0.9588	0.9593	0.9593	0.9572	0.9554	0.9550	0.9550	0.9601	0.9580	0.9588	0.9587	0.9587	0.9566
ecoli-0-1-3-7_vs_2-6	0.7774	0.7860	0.7592	0.8894	**0.9076**	0.6093	0.6093	0.7791	0.8832	0.9064	0.7860	0.7860	0.7797	0.7879	0.7628	0.7628	0.8574	0.8926	0.7172	0.8998	0.8998	0.7785	0.7592	0.7585	0.7585	0.8316	0.8831	0.7060	0.8864	0.8864	0.7695
yeast6	**0.8652**	0.8644	0.8560	0.8561	0.8505	0.8383	0.8383	0.8650	0.8568	0.8507	0.8644	0.8644	0.8643	0.8646	0.8593	0.8593	0.8604	0.8545	0.8333	0.8586	0.8586	0.8622	0.8560	0.8554	0.8554	0.8619	0.8526	0.8324	0.8475	0.8475	0.8598
abalone19	0.7108	0.7118	0.6987	0.7283	0.6925	0.6735	0.6735	0.7163	**0.7306**	0.6948	0.7118	0.7118	0.7147	0.7197	0.7188	0.7188	0.7061	0.7166	0.7172	0.7206	0.7206	0.7096	0.6987	0.7035	0.7033	0.6977	0.6946	0.7149	0.6982	0.6982	0.7078
Average	0.8662	0.8686	0.8618	0.8703	0.8634	0.8022	0.8022	0.8681	0.8693	0.8636	0.8686	0.8686	0.8671	0.8692	0.8655	0.8655	0.8694	0.8682	0.8518	**0.8711**	**0.8711**	0.8666	0.8618	0.8632	0.8628	0.8680	0.8660	0.8520	0.8681	0.8681	0.8650

From the results on this table we have carried out the corresponding statistical analysis in a hierarchical manner, answering the different questions raised. Figure 4 depicts this study. When the comparison involves a pair of methods, the output of the Wilcoxon's test is shown. In this case, the ranks are presented near each method (the greater, the better). In other case, Friedman aligned-ranks test is used for the comparison (the lower the ranks, the better). In both cases, we present the p-values (for Friedman aligned-ranks, the ones given by Holm's post-hoc test in case of significant differences being found). In the next sections, we analyze both the table of results and the outputs of the statistical tests.

A. CPM VS. ADDITIVE AND $\lambda-$MEASURES

We focus on the comparison of CPM and its simpler counterparts, in which the only coefficients that are learnt are the singletons. Looking at the average GM results, CPM-Choquet achieves the best performance. If we focus on Sugeno integral only, the CPM achieves worse overall performance. However, we should drawn the conclusions from the proper statistical analysis, which is performed for Choquet and Sugeno separately.

We performed two aligned-Friedman ranks test to compare CPM vs. additive and $\lambda-$measure (one for each integral). In the case of Choquet, one can observe in Figure 4 that CPM statistically outperforms the other two methods by a large margin (obtaining very low p-values). Otherwise, for Sugeno integral results become much more similar. No significant differences are found among different methods and the $\lambda-$measure is the one that gets the lowest number of ranks by a small margin. It seems that Sugeno does not exploit the full potential behind fuzzy measures for this application, although we will analyze this fact later.

B. GLOBAL VS. DYNAMIC APPROACHES

This section is focused on the question of whether dynamic approaches can overcome global ones. We have both ISFM and MHM measures to answer this question, as both were proposed in their dynamic and global versions. For this reason, we compare these alternatives by pairs for both Choquet and Sugeno integrals.

In the case of ISFM, in both Choquet and Sugeno the dynamic variant (D-ISFM) achieves the highest number of ranks, with significant differences. Regarding MHM, the dynamic model performs slightly better with the Choquet integral, whereas with Sugeno, the global model achieves a higher number of ranks. Anyway, high p-values are obtained in both cases, showing no significant differences between both models.

Interestingly, although one could expect a greater benefit from dynamic approaches due to their greater adaptability to each example, this is only true for ISFM, where significant differences are found, whereas in MHM there are no major differences.

C. CPM VS. EXISTING ALTERNATIVE

For the next comparison, we consider the winners (in term of ranks) from the previous ones and analyze which fuzzy measure works best in the current framework including our proposed CPM.

Looking at Figure 4, the proposed CPM achieves the highest number of ranks in the Friedman aligned-ranks test when the Choquet integral is considered as underlying aggregation. However, no statistically significant differences are found. Anyway, this is an interesting results as it shows that a global estimation of the fuzzy measure (without approximations and avoiding additivity) can give and advantage. Obviously, as it can be concluded from Table 7 and Figure 4, a small margin for improvement exists when only the aggregation is considered. Anyway, there are many applications where a small improvement in terms of numbers can lead to high gains.

Otherwise, when the Sugeno integral is taken as aggregation, EB gest the lowest number of ranks, showing an advantage with respect to the other contenders. There is a synergy between EB measure and Sugeno, which has not been present with Choquet, although no statistical differences are found when applying the Holm's post-hoc test. Nevertheless, whether EB with Sugeno or CPM with Choquet works better remains to be studied and is the focus of the next comparison.

D. FUZZY INTEGRALS: WHICH ONE WORKS BEST?

Among fuzzy measures, CPM has been the best performer coupled with Choquet, whereas EB has been the best with Sugeno. As it can be extracted from the Wilcoxon test in Figure 4, CPM with Choquet gets the highest number of ranks by a large margin, statistically outperforming EB with Sugeno. Hence, the greater advantage of using Choquet can be stressed. At the same time, the synergy between the proposed CPM and Choquet should be highlighted as it can be considered to be the best alternative among fuzzy measure-based aggregations.

E. FUZZY MEASURE BASED APPROACHES VS. CLASSICAL AGGREGATIONS

Finally, we aim to study whether the usage of fuzzy measure-based aggregations pays off in terms of performance when compared to simpler, widely used alternatives such as classical aggregations.

First, following the same idea as before, we have performed a statistical test to compare both unweighted and weighted aggregations, respectively. Among unweighted aggregations, we have considered O (original), which is the way the aggregation is performed in UnderBagging_RE [20]. This will allow us to study whether more advanced aggregations can help in further improving state-of-the-art performance on ensembles for the class imbalance problem.

Observing the results of the Aligned-Friedman tests in Figure 4, the arithmetic mean (AM) is the best performer

among unweighted aggregations, whereas the OWA operator at least half (Q_{alh}) achieved the lowest ranks among weighted ones. In the latter case, the only statistical difference is obtained with respect to WHM (weighted harmonic mean). In the former case, the maximum and minimum aggregations are also significantly outperformed by AM and the same occurs for HM. Among the rest, differences are not significant. Notice, however, that in terms of ranks AM performs better than the usually considered alternative O, which does not take the confidence or probability of the non-predicted class into account. This may tell us that the more the information considered for the aggregation is, the better the decision made can be.

When AM is compared with Q_{alh}, the Wilcoxon test returns no winner, as both perform equally in terms of ranks. Hence, one or the other could be considered for the final comparison and main point of this section.

The last Wilcoxon test compares CPM with Choquet against AM, that is, the best performer fuzzy measure-based aggregation vs. the best classical one. Attending to the results given by this test presented in Figure 4, CPM with Choquet not only achieves the largest number of ranks, but also significantly outperforms the AM. Hence, we can conclude that one can take advantage of more sophisticated aggregations based on fuzzy measures to obtain a significant advantage with respect to the most widely used arithmetic mean.

VIII. CONCLUSION

In this paper, we have studied unsupervised methods for fuzzy measure learning applied to the combination of classifier ensembles. We have categorized these methods depending on what drives the learning strategy, either the probability distributions provided by the classifiers or some a priori knowledge about the classifiers of the ensemble. We further divided these methods into dynamic or global behavior depending on whether a measure is built for each specific instance or a single global measure is learned and used for all the instances. We have put our focus on global methods as dynamic ones result in only using one single path of each fuzzy measure, resembling dynamic OWA operators. To overcome the limitations of current measures, we have proposed CPM, a global non-additive fuzzy measure whose coefficients are directly estimated from the performance of each possible coalition. In thorough empirical study in the challenging framework of imbalanced datasets, CPM has shown to be competitive. It has been the best performer in the hierarchical statistical study carried out, which included all the unsupervised methods for fuzzy measure learning reviewed and other classical aggregations.

For future work, we aim to tackle one of the disadvantages of learning a global fuzzy measure, that is, the complexity of estimating 2^N parameters, especially when N becomes large. For this purpose, fuzzy measure compression could be explored [49]. We want also to test CPM in other classification frameworks such as multi-class, multi-label or multi-instance learning. Likewise, we would like to further analyze the behavior of the different aggregations in similar frameworks to better understand when and why non-classical aggregations can be improved. For example, we could only focus on those examples whose classification can be affected by the aggregation to better analyze the differences among methods. Similarly, we will study other ensemble methods whose resulting classifiers could differ more (e.g., Boosting), increasing the importance of the aggregation phase.

REFERENCES

[1] B. Krawczyk, M. Galar, L. Jeleń, and F. Herrera, "Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy," *Appl. Soft Comput.*, vol. 38, pp. 714–726, Jan. 2016.

[2] J. Sanz, D. Paternain, M. Galar, J. Fernandez, D. Reyero, and T. Belzunegui, "A new survival status prediction system for severe trauma patients based on a multiple classifier system," *Comput. Methods Programs Biomed.*, vol. 142, pp. 1–8, Apr. 2017.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Hoboken, NJ, USA: Wiley, 2001.

[4] J. R. Quinlan, *C4.5: Programs for Machine Learning*, 1st ed. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[5] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[6] J. Alcala-Fdez, R. Alcala, and F. Herrera, "A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 5, pp. 857–872, Oct. 2011.

[7] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Comput.*, vol. 8, no. 7, pp. 1341–1390, Oct. 1996.

[8] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[10] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, 2nd ed. Hoboken, NJ, USA: Wiley, 2014.

[11] D. Sannen, E. Lughofer, and H. Van Brussel, "Towards incremental classifier fusion," *Intell. Data Anal.*, vol. 14, no. 1, pp. 3–30, Jan. 2010.

[12] G. Choquet, "Theory of capacities," *Ann. Inst. Fourier*, vol. 5, pp. 131–295, 1953.

[13] M. Sugeno, "Theory of fuzzy integrals and its applications," Ph.D. dissertation, Tokyo Inst. Technol., Tokyo, Japan, 1974.

[14] D. Štefka and M. Holeňa, "Dynamic classifier aggregation using interacion-sensitive fuzzy measures," *Fuzzy Sets Syst.*, vol. 270, pp. 25–52, Jul. 2015.

[15] G. J. Scott, K. C. Hagan, R. A. Marcum, J. A. Hurt, D. T. Anderson, and C. H. Davis, "Enhanced fusion of deep neural networks for classification of benchmark high-resolution image data sets," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 9, pp. 1451–1455, Sep. 2018.

[16] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[17] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

[18] J. Alcalâ-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Multiple-Valued Log. Soft Comput.*, vol. 17, nos. 2–3, pp. 255–287, 2011.

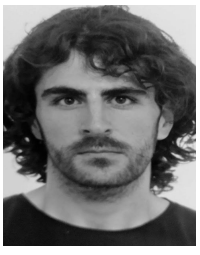
[19] R. Barandela, R. M. Valdovinos, and J. S. Sánchez, "New applications of ensembles of classifiers," *Pattern Anal. Appl.*, vol. 6, no. 3, pp. 245–256, Dec. 2003.

[20] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced datasets," *Inf. Sci.*, vol. 354, pp. 178–196, Aug. 2016.

[21] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced non-parametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, May 2010.

[22] G. Beliakov, A. Pradera, and T. Calvo, *Aggregation Functions: A Guide for Practitioners*. Berlin, Germany: Springer-Verlag, 2007.

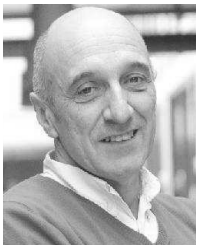
[23] G. Beliakov, H. Bustince, and A. Pradera, *A Practical Guide to Averaging Functions*, 2nd ed. Cham, Switzerland: Springer, 2015.
[24] T. Calvo, G. Mayor, and R. Mesiar, *Aggregation Operators: New Trends and Applications*. Berlin, Germany: Physica-Verlag, 2002.
[25] M. Grabisch, J.-L. Marichal, R. Mesiar, and E. Pap, *Aggregation Functions*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[26] R. R. Yager, "Quantifier guided aggregation using OWA operators," *Int. J. Intell. Syst.*, vol. 11, no. 1, pp. 49–73, 1998.
[27] F. Herrera, E. Herrera-Viedma, and J. L. Verdegay, "Direct approach processes in group decision making using linguistic OWA operators," *Fuzzy Sets Syst.*, vol. 79, no. 2, pp. 175–190, Apr. 1996.
[28] Y. Narukawa and V. Torra, "Fuzzy measure and probability distributions: Distorted probabilities," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 5, pp. 617–629, Oct. 2005.
[29] L. Sun, H. Dong, and A. X. Liu, "Aggregation functions considering criteria interrelationships in fuzzy multi-criteria decision making: State-of-the-art," *IEEE Access*, vol. 6, pp. 68104–68136, 2018.
[30] Q. Yang and X. Wu, "10 Challenging problems in data mining research," *Int. J. Inf. Technol. Decis. Making*, vol. 05, no. 04, pp. 597–604, Dec. 2006.
[31] R. Barandela, J. S. Sánchez, V. García, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognit.*, vol. 36, no. 3, pp. 849–851, Mar. 2003.
[32] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.
[33] N. V. Chawla, D. A. Cieslak, L. O. Hall, and A. Joshi, "Automatically countering imbalance and its empirical relationship to cost," *Data Mining Knowl. Discovery*, vol. 17, no. 2, pp. 225–252, Oct. 2008.
[34] G. Martínez-Muñoz, D. Hernández-Lobato, and A. Suárez, "An analysis of ensemble pruning techniques based on ordered aggregation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 245–259, Feb. 2009.
[35] Z. Wang, K.-S. Leung, and G. J. Klir, "Applying fuzzy measures and nonlinear integrals in data mining," *Fuzzy Sets Syst.*, vol. 156, no. 3, pp. 371–380, Dec. 2005.
[36] K. Xu, Z. Wang, P.-A. Heng, and K.-S. Leung, "Classification by nonlinear integral projections," *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 2, pp. 187–201, Apr. 2003.
[37] H. Fang, M. L. Rizzo, H. Wang, K. A. Espy, and Z. Wang, "A new nonlinear classifier with a penalized signed fuzzy measure using effective genetic algorithm," *Pattern Recognit.*, vol. 43, no. 4, pp. 1393–1401, Apr. 2010.
[38] M. Popescu, J. M. Keller, and J. A. Mitchell, "Fuzzy measures on the gene ontology for gene product similarity," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 3, no. 3, pp. 263–274, Jul. 2006.
[39] E. Barrenechea, H. Bustince, J. Fernandez, D. Paternain, and J. Sanz, "Using the Choquet integral in the fuzzy reasoning method of fuzzy rule-based classification systems," *Axioms*, vol. 2, no. 2, pp. 208–223, 2013.
[40] G. Lucca, J. A. Sanz, G. P. Dimuro, B. Bedregal, R. Mesiar, A. Kolesarova, and H. Bustince, "Preaggregation functions: Construction and an application," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 2, pp. 260–272, Apr. 2016.
[41] G. P. Dimuro, J. Fernández, B. Bedregal, R. Mesiar, J. A. Sanz, G. Lucca, and H. Bustince, "The state-of-art of the generalizations of the Choquet integral: From aggregation and pre-aggregation to ordered directionally monotone functions," *Inf. Fusion*, vol. 57, pp. 27–43, May 2020.
[42] A. F. Tehrani, W. Cheng, and E. Hüllermeier, "Preference learning using the Choquet integral: The case of multipartite ranking," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 6, pp. 1102–1113, Dec. 2012.
[43] X. Du and A. Zare, "Multiple instance Choquet integral classifier fusion and regression for remote sensing applications," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2741–2753, May 2019.
[44] S. Guo, R. Chen, M. Wei, H. Li, and Y. Liu, "Ensemble data reduction techniques and multi-RSMOTE via fuzzy integral for bug report classification," *IEEE Access*, vol. 6, pp. 45934–45950, 2018.
[45] Y. Chen and J. Z. Wang, "Support vector learning for fuzzy rule-based classification systems," *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 6, pp. 716–728, Dec. 2003.
[46] D. Leite and I. Škrjanc, "Ensemble of evolving optimal granular experts, OWA aggregation, and time series prediction," *Inf. Sci.*, vol. 504, pp. 95–112, Dec. 2019.
[47] W. Jia and W. Zhenyuan, "Using neural networks to determine Sugeno measures by statistics," *Neural Netw.*, vol. 10, no. 1, pp. 183–195, Jan. 1997.
[48] W. Wang, Z. Y. Wang, and G. J. Klir, "Genetic algorithms for determining fuzzy measures from data," *J. Intell. Fuzzy Syst.*, vol. 6, no. 2, pp. 171–183, 1998.
[49] M. A. Islam, D. T. Anderson, A. J. Pinar, and T. C. Havens, "Data-driven compression and efficient learning of the Choquet integral," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 4, pp. 1908–1922, Aug. 2018.
[50] A. J. Pinar, J. Rice, L. Hu, D. T. Anderson, and T. C. Havens, "Efficient multiple kernel classification using feature and decision level fusion," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1403–1416, Dec. 2017.
[51] M. Grabisch, "A new algorithm for identifying fuzzy measures and its application to pattern recognition," in *Proc. IEEE Int. Conf. Fuzzy Systems. Int. Joint Conf. 4th IEEE Int. Conf. Fuzzy Syst. 2nd Int. Fuzzy Eng. Symp.*, Mar. 1995, pp. 145–150.
[52] J. Murillo, S. Guillaume, E. Tapia, and P. Bulacio, "Revised HLMS: A useful algorithm for fuzzy measure identification," *Inf. Fusion*, vol. 14, no. 4, pp. 532–540, Oct. 2013.
[53] M. Z. Jan and B. Verma, "A novel diversity measure and classifier selection approach for generating ensemble classifiers," *IEEE Access*, vol. 7, pp. 156360–156373, 2019.
[54] Z. Xu, "Dependent OWA operators," in *Modeling Decisions for Artificial Intelligence*, V. Torra, Y. Narukawa, A. Valls, and J. Domingo-Ferrer, Eds. Berlin, Germany: Springer, 2006, pp. 172–178.
[55] R. Mesiar and J. Špirková, "Weighted means and weighting functions," *Kybernetika*, vol. 42, no. 2, pp. 151–160, 2006.
[56] V. S. Costa, A. D. S. Farias, B. Bedregal, R. H. N. Santiago, and A. M. D. P. Canuto, "Combining multiple algorithms in classifier ensembles using generalized mixture functions," *Neurocomputing*, vol. 313, pp. 402–414, Nov. 2018.
[57] A. G. C. Pacheco and R. A. Krohling, "Aggregation of neural classifiers using choquet integral with respect to a fuzzy measure," *Neurocomputing*, vol. 292, pp. 151–164, May 2018.
[58] I. Kojadinovic, "Estimation of the weights of interacting criteria from the set of profiles by means of information-theoretic functionals," *Eur. J. Oper. Res.*, vol. 155, no. 3, pp. 741–751, Jun. 2004.
[59] H. V. Rowley, A. Geschke, and M. Lenzen, "A practical approach for estimating weights of interacting criteria from profile sets," *Fuzzy Sets Syst.*, vol. 272, pp. 70–88, Aug. 2015.
[60] Y. Cao, "Aggregating multiple classification results using Choquet integral for financial distress early warning," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 1830–1836, Feb. 2012.
[61] E. Hüllermeier and S. Vanderlooy, "Combining predictions in pairwise classification: An optimal adaptive voting strategy and its relation to weighted voting," *Pattern Recognit.*, vol. 43, no. 1, pp. 128–142, Jan. 2010.
[62] E. Hüllermeier, "Cho-k-NN: A method for combining interacting pieces of evidence in case-based learning," in *Proc. 19th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2005, pp. 3–8.
[63] D. Paternain, H. Bustince, M. Pagola, P. Sussner, A. Kolesárová, and R. Mesiar, "Capacities and overlap indexes with an application in fuzzy rule-based classification systems," *Fuzzy Sets Syst.*, vol. 305, pp. 70–94, Dec. 2016.
[64] H. Bustince, J. Fernandez, R. Mesiar, J. Montero, and R. Orduna, "Overlap functions," *Nonlinear Anal.*, vol. 72, pp. 1488–1499, 2010.
[65] S.-B. Cho and J. H. Kim, "Multiple network fusion using fuzzy logic," *IEEE Trans. Neural Netw.*, vol. 6, no. 2, pp. 497–501, Mar. 1995.
[66] L. I. Kuncheva, "'Fuzzy' versus 'nonfuzzy' in combining classifiers designed by Boosting," *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 6, pp. 729–741, Dec. 2003.



MIKEL URIZ received the B.Sc. and M.Sc. degrees in computer science from the Public University of Navarra, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree. His research interests focus on machine learning and deep learning applications, and the application of fuzzy techniques in these topics. He holds a Pre-Doctoral Scholarship from Public University of Navarra.

DANIEL PATERNAIN received the M.Sc. and Ph.D. degrees in computer science from the Public University of Navarra, Pamplona, Spain, in 2008 and 2013, respectively. He is currently a Lecturer with the Department of Statistics, Computer Science and Mathematics, Public University of Navarra. He is also the author or coauthor of almost 25 articles in JCR and more than 50 international conference communications. His research interests include both theoretical and applied aspects of aggregation functions, image processing, and machine learning. His main contributions are based on the use of aggregation techniques and new information fusion procedures, such as aggregation functions, penalty functions or fuzzy integrals in image processing, and supervised classification algorithms.

IRIS DOMINGUEZ-CATENA received the B.Sc. and M.Sc. degrees in computer science from the Public University of Navarra, in 2015 and 2020, respectively. She worked for several private software companies, from 2015 to 2018. She is currently a Researcher with the Public University of Navarra, working for different projects using machine learning. Her research interests focus on machine learning and deep learning, and the application of fuzzy techniques in these topics.



HUMBERTO BUSTINCE (Senior Member, IEEE) received the B.Sc. degree in physics from the University of Salamanca, in 1983, and the Ph.D. degree in mathematics from the Public University of Navarra, Pamplona, Spain, in 1994. He is currently a Full Professor of computer science and artificial intelligence with the Public University of Navarra, where he is also the main Researcher of the Artificial Intelligence and Approximate Reasoning Group, whose main research lines are both theoretical (aggregation functions, information and comparison measures, fuzzy sets, and extensions), and applied (image processing, classification, machine learning, data mining, and big data). He has led more than 10 I+D public-funded research projects, at a national and at a regional level. He has authored more than 210 works, according to Web of Science. He is the coauthor of a monography about averaging functions and coeditor of several books. He is also an Associated Editor of the IEEE TRANSACTIONS ON FUZZY SYSTEMS JOURNAL and a member of the editorial board of the Journals *Fuzzy Sets and Systems*, *Information Fusion*, the *International Journal of Computational Intelligence Systems*, and the *Journal of Intelligent and Fuzzy Systems*.

MIKEL GALAR (Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science from the Public University of Navarra, Pamplona, Spain, in 2009 and 2012, respectively. He is currently an Associate Professor with the Department of Statistics, Computer Science, and Mathematics, Public University of Navarra. He is the author of 35 published original articles in international journals and more than 50 contributions to conferences. He is also a Reviewer of more than 35 international journals. His research interests are machine learning, data mining, classification, ensemble learning, evolutionary algorithms, fuzzy systems, and big data. He is a member of the European Society for Fuzzy Logic and Technology (EUSFLAT) and the Spanish Association of Artificial Intelligence (AEPIA). He has received the Extraordinary Prize for his Ph.D. thesis from the Public University of Navarra and the 2013 IEEE TRANSACTIONS ON FUZZY SYSTEM Outstanding Paper Award for the paper A New Approach to Interval-Valued Choquet Integrals and the Problem of Ordering in Interval-Valued Fuzzy Set Applications (bestowed in 2016).