Received: 6 July 2018 | Revised: 30 August 2019 | Accepted: 7 October 2019

DOI: 10.1111/coin.12249

Computational Intelligence WILEY

ORIGINAL ARTICLE

CSBF: A static ensemble fusion method based on the centrality score of complex networks

Ronan Assumpção Silva[1,2] | **Alceu de Souza Britto Jr**[1,3] | **Fabricio Enembreck**[1] | **Robert Sabourin**[4] | **Luiz E. S. de Oliveira**[5]

[1]Postgraduate Program in Informatics (PPGIA), Pontifical Catholic University of Parana (PUCPR), Parana, Brazil

[2]Department of Informatics, Federal Institute of Parana (IFPR), Parana, Brazil

[3]Department of Informatics, State University of Ponta Grossa (UEPG), Parana, Brazil

[4]Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle, École de Technologie Supérieure (ÉTS), Montreal, Canada

[5]Department of Informatics, Federal University of Parana (UFPR), Parana, Brazil

Correspondence

Ronan Assumpção Silva, Pontifical Catholic University of Parana (PUCPR), R. Imaculada Conceição, 1155, Curitiba, PR 80215-901, Brazil.
Email: ronan.ras@gmail.com

Funding information

Brazilian National Council for Scientific and Technological Development (CNPq), 306684/2018-2; Coordination for the Improvement of Higher Education Personnel (CAPES), 88881.131663/2016-01

Abstract

Ensemble of classifiers can improve classification accuracy by combining several models. The fusion method plays an important role in the ensemble performance. Usually, a criterion for weighting the decision of each ensemble member is adopted. Frequently, this can be done using some heuristic based on accuracy or confidence. Then, the used fusion rule must consider the established criterion for providing a most reliable ensemble output through a kind of competition among the ensemble members. This article presents a new ensemble fusion method, named centrality score-based fusion, which uses the centrality concept in the context of social network analysis (SNA) as a criterion for the ensemble decision. Centrality measures have been applied in the SNA to measure the importance of each person inside of a social network, taking into account the relationship of each person with all others. Thus, the idea is to derive the classifier weight considering the overall classifier prominence inside the ensemble network, which reflects the relationships among pairs of classifiers. We hypothesized that the prominent position of a classifier based on its pairwise relationship with the other ensemble members could be its weight in the fusion process. A robust experimental protocol has confirmed that centrality measures represent a promising strategy to weight the classifiers of an ensemble, showing that the proposed fusion method performed well against the literature.

KEYWORDS

centrality measures, complex network, fusion of classifiers

1 | INTRODUCTION

In the last decades, social network analysis (SNA) emerged as a multidisciplinary field belonging to network science (NS) in which scientists try to understand the behavior of different types of objects organized in a network.[1] Different types of research use the NS to discover new patterns that traditional approaches could not. One of the most meaningful contributions of the Moreno study[2] in 1934, a precursor of the SNA development, is to consider the set of people as a whole and analyze how the people's choices affect the group structure. For instance, Jackson describes in Reference 3 that the social network pattern of the medieval elites from Scotland using the PoMS database. His work described how SNA found a new pattern, and he also states that the pattern could not be found using traditional historical methods. Jackson work reveals a new opinion leader, a person who developed an important role that possibly leads to important happenings in the course of Scotland history. Recently, NS gains pattern recognition attention in different applications, such as multiagent systems,[4] concept drift,[5,6] recommendation systems,[7] and ensemble of classifiers.[8] The belief is the same as Jackson; NS can find patterns that common approaches could not.

In this work, the classifier combination problem is presented as an ensemble network, which is analyzed by centrality measures of the NS. Such measures are used to estimate the importance of the members of a network in which the vertices represent ensemble members and the edges represent a pairwise relationship among them. In this study, we aim to answer the following questions:

1. May the centrality of each classifier inside the ensemble network built using pairwise diversity contribute to improving the accuracy of the ensemble fusion?
2. Which pairwise diversity measure should be used to represent the ensemble network in which the centrality information will be computed?
3. Which centrality measure is more appropriate to provide the importance of each classifier in the ensemble network?

To answer these questions, we start by describing the centrality measures, associating them with the ensemble learning theory. This theoretical association is the core of the article, providing the background knowledge that is necessary to understand the proposed method. The expectation is that by applying centrality measures is it possible to estimate how important is a classifier for the ensemble concerning the complementarity of its errors when compared with the other ensemble members. With this in mind, we developed a new method for static ensemble fusion, named centrality score–based fusion (CSBF). With the CSBF method, we performed a set of experiments to compare different combinations of pairwise relationship and centrality measures, which is presented in Section 6.1. Besides, the best method setup is compared against the methods that represent the state of art. Some statistical tests suggest that the proposed method differs from other approaches and also is recommended for several classification problems as we can see in Section 6.2.

The contributions of this work are 3-fold. From the theory perspective, this study presents an analysis of undirected and directed ensemble networks, which were structured by the classifiers and their relationships based on pairwise diversity measures. This work aims to clarify the importance of the classifiers according to their prominent position inside the ensemble network from a centrality point of view. The proposed CSBF method is also a contribution, and its novelty is related to the strategy used to combine classifiers considering their score of importance inside the ensemble. Finally, as a practical contribution, the performance of the proposed method was evaluated over 40 different machine learning problems and against 15 fusion methods available in the literature. The results show that our approach is the best strategy for combining classifiers in 70.33% of these problems.

The work is organized as follows: first, it presents in Section 2 a brief NS theory that will introduce the basic concepts used in this work. Then, in Section 3, the main definitions related to ensemble learning and a literature review of classifier fusion methods and diversity measures are presented. The proposed fusion method is presented in Section 5. A robust set of experiments is described in Section 6 and also discussed. Finally, Section 7 presents our conclusions and perspectives of future work.

2 | NETWORK SCIENCE

It is a multidisciplinary field that concerns the analysis of objects by their relationships. The theory behind the area is continuously improved by researchers in the most different scientific fields such as SNA, complex network, chemistry, traffic engineers, and computer science, to cite a few. This section briefly describes the background theory used to propose and analyze classification problems in the light of ensemble learning. To this end, first, a short history is presented to instigate the reader who is not familiarized with the field to understand the powerful contribution of this growing area. Useful and straightforward definitions here aim to be a guide to understanding the properties used to build the ensemble network and finally compute the classifiers importance by centrality measures.

2.1 | Basic definitions

In this study, we describe only the basic definitions of NS to help the reader to understand the possible contributions to ensemble learning. It represents a particular problem as a network in which vertices represent objects, groups, or individuals and edges represent a relationship among them. The edges are undirected in case of symmetric relations or direct in case of asymmetric. An undirected network $G(V, E)$ consists of two sets namely $V \neq \emptyset$ and E. The set $V = v_1, v_2, \dots, v_T$ representing objects (elements, individuals, groups, nodes) that are called vertices. The set $E = e_1, e_2, \dots, e_U$ is a distinct and unordered group representing the pairs of elements of V, so each element e is a pair of vertices i and j (i,j). Objects i and j are referred to as neighbors. Different from undirected networks, in the direct ones (i,j) and (j,i) represent different edges. As one may see, the network concept presented here is the same than graph.

A walk $walk(x, y)$ from vertex x to vertex y is defined by Latora et al[1] as an alternating sequence of vertices and edges $walk = (x \equiv v_0, e_1, v_1, e_2, \dots, e_l, v_l \equiv y)$. It begins at x and ends at y, such that each edge $e_i = (v_{i-1}, v_i)$ for $i = 1, 2, \ldots, l$). A walk is commonly represented as the sequence of the traversed vertices, between two given vertices x and y, resulting in $walk = (x \equiv v_0, v_1, \ldots, v_l \equiv y)$. The length of a walk $\ell(walk)$ is the number of edges in the sequence. A path is a walk where each vertex is visited only once. The shortest path, also known as a geodesic, is a path of minimum length from vertex x to vertex y. The term “distance” is frequently used to describe the alternating vertex/edge relations between two vertices. In this article, we consider a graph $G(V, E)$, in which $V = c_1, c_2, \ldots, c_T$ represents the pool of classifiers where each c_i is an ensemble member of a team composed of T classifiers, and the set of edges E represents their relationship or diversity. From that network representing an ensemble of classifiers, we compute centrality measures, which are described in the following section.

2.2 | Centrality measures

Centrality measures are designed to weight the members of a network based on their topological importance. In other words, centrality measures evaluate the importance of the individuals, considering the role they play with others, based on the relationships represented on the network. The following four classic measures are commonly used in the literature: degree,[9] betweenness,[10] closeness,[9] and eigenvector.[11] These measures have focused on different pieces of information, such as (a) the sum of edges (degree), (b) shortest paths (closeness, betweenness), and (c) walks (eigenvector) also known as interactive refinement.[12]

The degree centrality K_{c_i} is related to the number of edges of a network member. It can be also the sum of the weight of the edges in the context of weighted networks. It is defined by Equation (1):

$$K_{c_i} = \sum_{c_j=1}^{T} E_{c_i c_j}, \tag{1}$$

where edge $E_{c_i c_j}$ connects the members c_i and c_j, and T is the total amount of members (classifiers) of the network or the total weight of these members. So, it is worth noting that the weight associated with $E_{c_i c_j}$ can be the original weight of the edge in a weighted network (weighted degree) or simply 1.0 in the case of an unweighted representation (unweighted degree), which only indicates the presence of an edge between c_i and c_j.

The degree centrality is a very simple measure to estimate. It considers only the local aspect of the network to be estimated, for example, a member degree is evaluated as direct (unweighted or weighted) connections with neighbors. As a result, the unweighted version of this measure cannot be used in a complete network because every member has the same degree ($T - 1$). An alternative, in this case, is to apply a simplification based on a pruning method.[13,14] Therefore, using the weighted degree in a weighted network does not demand a simplification process, which can lead to an extra computational effort.

Another classic centrality measure is betweenness.[9,10] This measure considers the number of shortest paths from each member of the network to all others that pass through each particular member, as denoted by Equation (2).

$$B_{c_i} = \sum_{c_j c_k} \frac{g_{c_j c_k}^{c_i}}{g_{c_j c_k}}, \tag{2}$$

where $g_{c_j c_k}$ is the number of geodesics between the vertex c_j and the vertex c_k; $g^{c_i}_{c_j c_k}$ refers to the geodesic between c_j and c_k that pass through c_i. As in degree centrality, the weight of the edges can be considered. In this case, the betweenness centrality computes the shortest paths based on the cost of each possible path, where the cost is the sum of weights of the edges belonging to the path. This centrality measure imposes a high computational cost, but the authors in Reference 15 suggest a different strategy to compute the shortest paths, which allows estimating the betweenness centrality efficiently.

A commonly used measure that also is based on the shortest paths is the closeness centrality.[9,16] This measure estimates the average distance of a member to all others in the network, taking into account the length of the average shortest paths. It considers that members with high centrality are those closest to all others. Like betweenness, the closeness centrality also depends on a connected network. Equation (3) can be used to compute the closeness centrality:

$$C_{c_i} = \frac{1}{l_i} = \frac{|T|}{\sum_{c_j} h_{c_i,c_j}}, \tag{3}$$

where l_i is the average shortest path length of each member to other members. The geodesic length h_{c_i,c_j} of a member c_i to any other member c_j inside the network must be estimated, and consequently summed. Averaging the number of vertices by the geodesics found leads to a higher centrality when the value is low. The network must have only one component to estimate closeness centrality and betweenness centrality, so it is possible to calculate all the distances (paths) between two given network members. For networks with more components, some alternatives are presented in Reference 17.

One last classic measure is the eigenvector centrality,[11,18,19] which can be used to compute the centrality x_i of a given node c_i using a matrix equation, or sum, as shown in Equation (4).

$$\lambda x = Ax, \quad \lambda x_i = \sum_{j=1}^{n} a_{ij} x_j, \qquad i = 1, \dots, n, \tag{4}$$

where $A = (a_{i,j})$ is the adjacency matrix, that is, $a_{i,j} = 1$ if vertex c_i is linked to vertex c_j and $a_{i,j} = 0$ otherwise, λ is the largest eigenvalue of A, and $(a_{i,j} x_j)$ is the each possible neighbor of c_i, whereas n is the number of vertices in the network. According to Reference 20, this centrality is based on the idea that a given vertex can depend not only on the number of its adjacent vertices but also on their value of centrality.

There are variants of the mentioned classic centrality measures for directed networks in which we have asymmetric relationships. For the most simple measure, the degree centrality considers indegree and outdegree of a vertex. The first counts the number of incoming arcs (or compute the sum of the weights of these arcs), whereas the second is the opposite, that is, the number of arcs directed to the node neighbors (or compute the sum of the weights of these edges). In this section, we have presented classic centrality measures appearing in the literature. The choice of a centrality measure depends on what a network represents and which questions the network analysis intends to answer. Each measure focuses on different aspects of the network to evaluate the role of its members, as shown in Figure 1. In Reference 21, the authors present several comparisons between centrality measures and conclude that degree, betweenness, closeness, and eigenvector are distinct, although related to the conceptual point of view. So, despite the most central vertex could be the same for different centralities, the score of centrality is different and ranking them

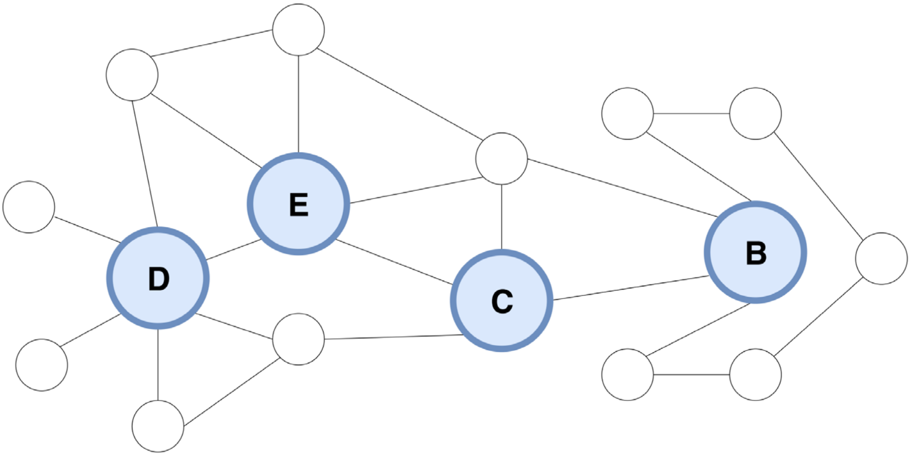

FIGURE 1 Example of a network and its most important vertices according to specific centrality measures. B, C, D, and E stands for highest betweenness, closeness, degree, and eigenvector, respectively [Color figure can be viewed at wileyonlinelibrary.com]

can present different order. It is also known that the network presents some complexity to build, to walk, and to view. The reader can find some information about these concerns in Reference 22.

3 | ENSEMBLE LEARNING

Ensemble learning is concerned to explore the diversity in an ensemble team to improve the accuracy of classification problems. Several models are combined, and the final ensemble prediction is the prediction of at least one ensemble member. For weighted combination methods, the ensemble members differently influence the final prediction. The weight of the classifiers is usually an equation using the number of the correct predictions over a validation set. This section aims to elucidate necessary information concerning a static weighted combination of classifiers.

3.1 | Basic concepts

Ensembles have been used as an attractive alternative to avoid the risk of selecting a single classifier as the solution for a pattern recognition problem. The ensemble divides the responsibility of covering the entire problem space among the members of a team composed of diverse and accurate classifiers. Merging the decisions of classifiers in which errors are different may lead to an improvement in the classification performance. Ensemble learning is used in different applications such as data stream,[23,24] concept drift,[25] class imbalance,[26] and sentiment analysis.[27]

The ensemble is usually described as a system composed of three phases: generation, selection, and fusion.[28-30] In the first phase, generation methods provide a set of accurate and diverse classifiers. The second phase is responsible for finding a subset of classifiers, which can be more accurate than using the whole pool. The selection phase is not obligatory. Finally, in the fusion phase, methods use some rule for combining different classifiers expecting that any ensemble member performs well but not so well as the final combined classifier.

Formally, an ensemble of classifiers is a set $C = \{c_1, c_2, \ldots, c_T\}$. Each ensemble member represents an independent function $c_t : R^n \rightarrow W$ that assigns a class label $w_i \in W$ to $x \in R^n$, where $W = \{w_1, w_2, \ldots, w_M\}$. A final decision takes into account the decisions of all the classifiers in the ensemble, or part of them, selected statically or dynamically.[29,31] If the final group is $T \geq 2$, so a fusion method must combine their decisions. The literature provides a variety of fusion methods;[32-34] some of them assume that all classifiers perform equally well and combine the classifiers' decisions without considering any mechanism to differentiate them concerning competence.

Considering that classifiers can differ, an alternative has been to weight the vote of the classifiers while assuming that they compete with one another in assigning the correct class label.[35] The competition among the classifiers in the ensemble through the use of weights is very promising. The literature thus bears witness to a wide variety of static and dynamic strategies for weighting the decision of each classifier in an ensemble. A static weighting strategy assesses the classifier confidence during the training phase of a classification system, and the weights obtained remain the same during the test phase, whereas in a dynamic weighting strategy, each classifier receives a different score for each test instance.

Several authors proved that the accuracy of the ensemble depends on individual accuracy, and the classifiers must commit different errors, in the sense of one complement the fails of another.[29,36-39] Selection methods made use of these information,[40-42] although it has been poorly used in fusion phase, as the number of works presented in section 4 suggests.

3.2 | Pool generation

There are a wide variety of pool generation methods,[33,34] also called ensemble generators or ensemble creation methods. In this section, we also present some classic approaches for the ensemble creation, such as bagging, boosting, and random subspaces.

Bagging[43] requires a labeled dataset S_{train}. Then, it creates T bootstrap samples from the training dataset to train classifiers $c_1, \ldots, c_{\mathrm{T}}$, one classifier on each sample. Every new dataset is generated by sampling from the S_{train} dataset, choosing T bags formed randomly and with replacement. Each bag has equal size, and the size can be the same as the original training dataset S_{train} (%bag = 100), or it can be smaller. The ensemble vote is the combination of the votes of the T classifiers, for example, majority vote (MV) or it can be a decision rule based on the joint probability distribution such as sum rule (SR), max rule, and product rule (PR), just to cite a few. Bagging can create the T classifiers in parallel.

Boosting[44] is similar to bagging considering the classifiers being built in different versions of the original training set S_{train}. However, these classifiers are constructed on weighted versions of the training set, which are dependent on the previous classification results. This approach first creates a sampled training set derived from the original and build the first classifier. This classifier is used to classify the training set, increasing the weight of each instance misclassified. From the second sampled training set to the last, the sampled training set is composed with instances from the original set, giving priority to those instances with high weights, that is, instances being misclassified. Each classifier in boosting is created in sequence.

Random subspaces[45] randomly selects a fixed number of subspaces from the original feature space, building one classifier on each subspace. Each feature of the subspace is chosen without replacement. The classifiers can be created in parallel, but there is no consensus on the size of the subspace,[33] so each researcher defines it differently, being fixed as seen in Reference 46 or the researcher can vary the size of the subspace, searching to evaluate the effect of the subspace dimension as seen in Reference 47.

3.3 | Diversity measures

Diversity can be explored in all phases of an ensemble. It can be estimated taking into account pairs of classifiers (also known as pairwise diversity) or the whole ensemble (nonpairwise). For the first approach, several measures are compared in Reference 39 as follows: Q statistics (QS),[48] correlation coefficient (CC),[49] disagreement (Dis) measure,[45,50] double-fault (DF) measure, [51] and kappa statistic (KS).[52] The second one can be estimated using the following measures: Kohavi-Wolpert variance,[53] interrater agreement,[54,55] entropy measure,[56] measure of difficulty,[57] generalized diversity,[58] and coincident failure diversity,[58] to cite a few. The focus of this section is to present pairwise diversity measures. This kind of measure is the most proper to discover the role of each classifier in the ensemble diversity. Pairwise measures are based on the relationship of classifiers computed on their incorrect/correct predictions. Such a relationship between two classifiers, c_i and c_j, is presented in Table 1. For a given instance, if both classifiers are correct, then N_{11} is increased. If both classifiers are wrong, N_{00} is increased. If c_i is correct, but c_j is incorrect, then N_{10} is increased, otherwise N_{01} is increased. This pairwise relationship can be estimated using the training or validation set.

The following pairwise diversity measures (QS, CC, Dis, DF, and KS) are based on the possible relationship presented in Table 1. The relationship between classifiers can be symmetric, for example, N_{00} and N_{11} or asymmetric, for example, N_{01} and N_{10}. Symmetric relationships are also observed in diversity measures; thus, pairwise diversity measures can be used as the edge score of an undirected network.

A well-known pairwise diversity measure is QS, which is denoted by Equation (5), assuming values in the interval $[-1, 1]$. For statistically independent classifiers, QS assumes 0. If the classifiers tend to recognize the same instances correctly, QS will be positive. Otherwise, if they commit errors in different instances, QS will be negative.

$$\text{QS} = \frac{N_{11} \times N_{00} - N_{01} \times N_{10}}{N_{11} \times N_{00} + N_{01} \times N_{10}}. \tag{5}$$

A similar measure is the CC that can be computed as shown in Equation (6). The CC values lie within the interval $[-1, 1]$.

$$\text{CC} = \frac{N_{11} \times N_{00} - N_{01} \times N_{10}}{\sqrt{\Delta}}, \tag{6}$$

where $\Delta = (N_{11} + N_{10}) \times (N_{11} + N_{01}) \times (N_{01} + N_{00}) \times (N_{10} + N_{00})$.

The Dis estimates the number of observations in which one classifier is incorrect, while the other is correct. The diversity is represented by higher scores of Dis in the interval $[0, 1]$. This measure is denoted by Equation (7):

$$\text{Dis} = \frac{N_{01} + N_{10}}{N}. \tag{7}$$

TABLE 1 Pairwise relationship between two classifiers c_i and c_j adapted from Reference 39

	c_i Correct (1)	**c_j Incorrect (0)**
c_i correct (1)	N_{11}	N_{10}
c_i incorrect (0)	N_{01}	N_{00}
Total $N = N_{00} + N_{01} + N_{10} + N_{11}$		

Another interesting pairwise measure is the DF.[51] It is defined as the number of examples that have been misclassified by both classifiers c_i and c_j, as denoted by Equation 8. In contrast to the Dis measure, DF is represented by lower scores in the interval [0, 1].

$$\mathrm{DF} = \frac{N_{00}}{N}. \tag{8}$$

Kappa statistic, proposed by Margineantu and Dietterich,[52] considers three scenarios: classifiers agree in each tested instance $\mathrm{KS} = 1$, $\mathrm{KS} = 0$ if the agreement is uncorrelated, and $\mathrm{KS} < 0$ if they agree less than expected by chance (rarely happens). KS is presented in Equation (9):

$$\mathrm{KS} = \frac{\Theta_1 - \Theta_2}{1 - \Theta_2}, \tag{9}$$

where

$$\Theta_1 = \frac{N_{11} + N_{00}}{N}, \tag{10}$$

and

$$\Theta_2 = \frac{(N_{11} + N_{10}) \times (N_{01} + N_{00}) + (N_{11} + N_{01}) \times (N_{10} + N_{00})}{N^2}. \tag{11}$$

The measurement of interrater agreement (MOIA) has a pairwise version, proposed by Fleiss[55] and used in Reference 39. Equation (12) presents this measure:

$$\mathrm{MOIA} = \frac{2(N_{11}N_{00} - N_{01}N_{10})}{(N_{11} + N_{10})(N_{01} + N_{00}) + (N_{11} + N_{01})(N_{10} + N_{00})}. \tag{12}$$

Product-moment (PM) correlation measure is proposed by Sharkey and Sharkey[59] and presented in Reference 60. The pairwise version is denoted by the Equation (13):

$$\mathrm{PM} = \frac{N_{00}}{\sqrt{N_{*0}N_{0*}}}. \tag{13}$$

This section presented classical pairwise diversity measures used in a static context, that is, using an entire dataset S_{val}. To estimate diversity dynamically, the reader can find it in References 61 to 63. They proposed different measures to deal with the difference between the ensemble members vote for each test instance of the S_{test}. Ending this section, it worth mention that the complexity of the presented pairwise measures differs from the nonpairwise. In Reference 64, the authors present the complexity of these two approaches, for nonpairwise the complexity is $O(T \times I)$, whereas for pairwise diversity, the complexity is $O(T^2 \times I)$. Besides, the ensemble size T and the size of a validation set S_{val} (used to estimate diversity) must be considered.

4 | RELATED WORKS

Several works[28,65-70] present some interesting reviews of topics related to ensembles of classifiers, and one common observation running through them is that weighted fusion methods are truly promising. Another general observation is that the fusion methods differ from each other. In Reference 68, the authors describe a taxonomy for fusion methods, which consists of the following three types: class label fusion, trainable fuser, and support functions fusion. Each one is based on class labels or support functions, and may demand extra training (trainable) or not (nontrainable). The related works are focused on the trainable class label and also presents the nontrainable support functions. The first group is related to the proposed method, whereas the second group presents some classical approaches, also referred to as simple rules, such as sum, min, max, product, mean, and average. Support functions also present more sophisticated methods, such as decision templates (DTE)[34] and Dempster-Shafer (DS) combination.[71] One may find some recent works on support functions in References 70,72, [75].

The nontrainable class label fusion implements variations of MV.[74] There are three variations of MV[75] as follows: unanimous voting, simple majority, and plurality voting. The first one assigns a final ensemble decision if all of its individuals have a voting agreement. The second is considered the agreement of at least half ensemble members. In the last, a final ensemble decision is simply the class most voted. Usually, the ensemble learning researchers referred to majority voting this last variation. MV variations assume that all classifiers have the same influence on the final decision, regardless of their respective ability to predict the sample correctly. Considering that classifiers are different, trainable class label fusion (weighted voting) are distinct:[74,75]

- the weights Ψ_{c_t} are assigned to each classifier;
- the weights $\Psi_{c_{t,w}}$ are assigned to classifiers and classes;
- the weights $\Psi_{c_t}(x)$ are assigned to classifiers and are dependent on the feature space;
- the weights $\Psi_{c_{t,w}(x)}$ are assigned to classifiers, classes, and are dependent on the feature space.

This section presents only the methods of the first group, that is, approaches concerned to assign the weight Ψ_{c_t} to each classifier, which is independent of classes and features. To weight classifiers, conventional approaches use variations of the weighted majority vote (WMV),[74] also known as the simple weighted vote in Reference 76. In classical WMV, the final ensemble decision is computed from the vote of each classifier in the ensemble, as shown in Equation (14):

$$\mathrm{vote}_{c_t} = \begin{cases} \Psi_{c_t} & \text{if } c_t \text{ picksclass } w_i, \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where Ψ_{c_t} is the weight defined for the classifier c_t, and vote_{c_t} represents its vote for a specific class w_i. The votes are accumulated by class, as denoted in Equation (15), where Θ_c is the total number of votes each class received. Finally, the class with the highest score constitutes the final decision of the ensemble.

$$\Theta_c = \sum_{t=1}^{T} vote_{c_t}. \tag{15}$$

Kuncheva weighted majority vote (KWMV)[34] estimates the weight of each classifier as defined in Equation (16):

$$\Psi_{c_t} = log\left(\frac{\alpha_{c_t}}{1-\alpha_{c_t}}\right), \tag{16}$$

where the weight Ψ of a classifier c_t is logarithmic of the division in which the accuracy α of the classifier c_t is the dividend, and the divisor is the error $(1 - \alpha_{c_t})$ of the same classifier. The accuracy α of each classifier is estimated on a validation set S_{val}.

The performance weighting (PW)[66] approach normalizes the accuracy of each classifier before using it as a weight, considering it as a proportion of the total accuracy of the ensemble, as denoted in Equation (17):

$$\Psi_{c_t} = \frac{1 - \varepsilon_{c_t}}{\sum_{j=1}^{T} \varepsilon_{c_j}}, \tag{17}$$

where ε_{c_t} is the normalized error of the classifier evaluated on a validation set S_{val}.

Bayesian combination (BC)[77,78] uses the posterior probability and the classifier's individual accuracy α. BC is presented in Equation (18):

$$\Psi_{c_t} = P(\alpha_{c_t} | S_{\text{val}}) \times P(w_i | x), \tag{18}$$

where $P(\alpha_{c_t} | S_{\text{val}})$ is the accuracy of the tth classifier c_t in a validation set S_{val}, and $P(w_i|x)$ is the posterior probability of the classifier c_t for a class w_i, given an instance x. After the sum of all classifier outputs is computed, the new pattern will be labeled according to the highest score given to the class w_i.

In Reference 79, the authors proposed a set of empirical scores named "power value" (PV) as a variation of WMV. The approach is shown in Equation (19):

$$\Psi_{c_t} = \frac{\alpha_{c_t}^{\text{pv}}}{\sum_{j=1}^{T} \alpha_{c_j}^{\text{pv}}}, \tag{19}$$

where pv is an empirical value. The authors evaluate 13 different values and conclude that $\text{pv} = 10$ provides the best accuracy. For $\text{pv} = 0$, the Ψ_{c_t} is defined as an simple average.

$$\Psi_{c_t} = \log\left(\frac{a_{c_t}}{1 - a_{c_t}}\right). \tag{20}$$

The following approaches rescaled weighted vote (RSWV), best-worst weighted vote (BWWV), and quadratic best-worst weighted vote (QBWWV) compute first the authority a_{c_T} of classifiers and then estimate the final weight using the Equation (20) inspired by Kuncheva.[34] Some of these can assign zero weights for classifiers with poor individual accuracy (in this case is a pruning of the original ensemble instead of a combination of the ensemble). These methods are presented as originally described in Reference 76. The first approach, RSWV, assigns zero as a score for bad-performing classifiers. Equation (21) assesses the authority of classifiers:

$$a_{c_t} = \max\left\{0, 1 - \frac{M \times \varepsilon_{c_t}}{V \times (M - 1)}\right\}, \tag{21}$$

where M is the size of class set W, V is the size of S_{val}, and ε_{c_t} is the number of errors of classifier c_t. If $a_{c_t} \leq \frac{1}{M}$, the classifier weight is zero, otherwise the weight is proportional to its accuracy performance.

The BWWV assigns zero to the worst classifier, and one to the best classifier after the accuracy of the entire ensemble is evaluated in a validation set S_{val}. As a consequence, the worst classifier is removed. The remaining classifiers are weighted linearly between these edges. Equation (22) presents the BWWV weighting method.

$$a_{c_t} = 1 - \frac{\varepsilon_{c_t} - \varepsilon_{\text{BEST}}}{\varepsilon_{\text{WORST}} - \varepsilon_{\text{BEST}}}, \tag{22}$$

where $\varepsilon_{\text{WORST}} = >\max_T\{\varepsilon_{c_t}\}$ and $\varepsilon_{\text{BEST}} = >\min_T\{\varepsilon_{c_t}\}$.

The following QBWWV approach is inspired by BWWV but assigns even more weight to better-performing classifiers. Equation (23) presents the QBWWV approach.

$$a_{c_t} = \left(\frac{\varepsilon_{\text{WORST}} - \varepsilon_{c_t}}{\varepsilon_{\text{WORST}} - \varepsilon_{\text{BEST}}}\right)^2. \tag{23}$$

In the literature, there are some more sophisticated solutions. In Reference 80, the authors proposed a signal strength–based combining approach. As in Reference 32, the algorithm starts by collecting individual votes of the classifiers for a given test instance, creating a decision profile. The decision profile is the basis for estimating some values, such as the signal strength and the signal strength direction and uncertainty degree, which are used to estimate the classifiers' influence and consequently the ensemble's final vote.

A Bayes voting approach is present in Reference 81. The authors proposed to compute the weight in regard to each class rather than to each expert. Those weights are computed by estimating the joint probability distribution of each class concerning all classifiers. The probability distribution is obtained using the naive Bayes probabilistic model.

Another recent solution for weighting classifiers is found in Reference 82. They proposed a weighted multiple classifier framework based on random projections (WMCRP). Initially, some training sets (called down-spaces) derived from the original training set are created, similarly to Bagging approach. Then, the unlabeled observation is projected on the down-spaces, generating a metadata by classifying the feature vectors of the observation in the down-spaces. The prediction is obtained by weighted linear combinations of the predictions made by the base classifiers.

Some combination rules are obtained on the measurement level,[32] such as minimum rule, maximum rule (MAR), PR, SR, average rule, and median rule. The minimum rule estimates the minimum score of each class between the classifiers and assigns the input pattern to the class with the maximum score among the minimum scores. Similar to minimum rule, MAR first finds the maximum score of each class between the classifiers, assigning the input pattern to the class with the highest score among the maximum scores. PR is obtained by multiplying the score provided by each base classifiers and assigns the class label with the highest computed score to the unlabeled input pattern. Similar to PR, the SR assigns the class label with the maximum score to a given input pattern but instead using the product of the outputs it uses the sum. The average rule computes the mean of the scores of each class, concerning every classifier in the ensemble and assigns to the input pattern the class with the highest score among. The median rule is similar average but computes the median instead of the mean of the scores.

A modified PR was proven to be superior to the classical PR.[73] The authors observed the lower performance of PR when increasing the number of classifiers in the ensemble. Therefore, they suggest an adaptation to the classic rule where they divide the joint probability distribution product by the number of classifiers belonging to the given ensemble.

In Reference 83, the authors present a very interesting comparison of different strategies to combine classifiers for the task of handwritten Indic script recognition, observing some improvement against the use of a single classifier. The best result was observed when a logistic regression classifier is used as a secondary model to combine the outputs of primary models. Also dedicated to handwritten Indic script recognition, the authors in Reference 71 present a fusion schema based on the DS theory which is applied to combine the decisions of two MLP classifiers, observing a significant improvement in recognition accuracy.

This section presented several combination methods. MV, the most simple, assigns equal weight to the classifiers expecting they perform similarly; however, it is not observed in all classification problems. Therefore, it is plausible to assume the classifiers are different, and it is an interesting idea to assign more weight to best-performing classifiers.[69] This is the reason behind the weighted majority voting. Concerning the reviews mentioned in this section, the classifier diversity importance to the ensemble is ignored. Measuring the classifier's significance to the ensemble diversity is not trivial, that is, a difference is a property shared by another classifier, a subgroup, or the entire group of classifiers. A method to evaluate the importance of a classifier considering its interactions as a whole is needed. The next section presents an approach for scoring the classifier importance based on the individual contribution to the ensemble using the relationship a classifier has to all other ensemble members. These combination approaches often use only the accuracy of individuals to weight classifiers.[84]

5 | THE PROPOSED FUSION METHOD

Figure 2 presents a general overview of the proposed CSBF method, which is a static weighted combination method organized into three phases. First, a pool of classifiers is generated in the overproduction phase. In the pairwise phase, a measure of diversity is used to represent the relationship between the classifiers in the pool. Finally, the weighted combination phase is built an ensemble network represented as a weighted graph $G(V, E)$, then a centrality measure estimates the importance of each classifier based on the weight of the relationship it has with its peers. The centrality measure suggests a score regarding the classifier's importance to the ensemble, and it is used as the classifier weight in the combination process.

Generally speaking, CSBF requires a dataset S, which is divided into three distinct datasets used for training, validation, and testing. The training set S_{train} provides instances x_{train} to generate a pool C of accurate and diverse classifiers. Then, the pool C is used to estimate pairwise relations ϖ among classifiers using a validation set S_{val}. The pool of classifiers C and the pairwise relationship ϖ are used to build an ensemble network $G(C, \varpi)$ where vertices represent classifiers, and the pairwise relationships are represented by edges (or arcs in the case of direction). A centrality measure χ is used to estimate the importance of classifiers taking into account their relationships. The importance score is the classifier weight Ψ. Finally, a combination rule uses the weight of classifiers and their vote to classify a test instance x_{test}. Each phase is detailed in the following sections.

5.1 | Overproduction

The overproduction phase generates a pool of classifiers $C = \{c_1, c_2, \ldots, c_T\}$, in which each member represents an independent function $c_t : R^n \rightarrow W$ that assigns a class label $w_i \in W$ to $x \in R^n$, where $W = \{w_1, w_2, \ldots, w_M\}$. The most popular pool generators are described in Section 3. Of these, bagging[43,85] is used in this work, even though CSBF as a combination method is not limited to this generator.

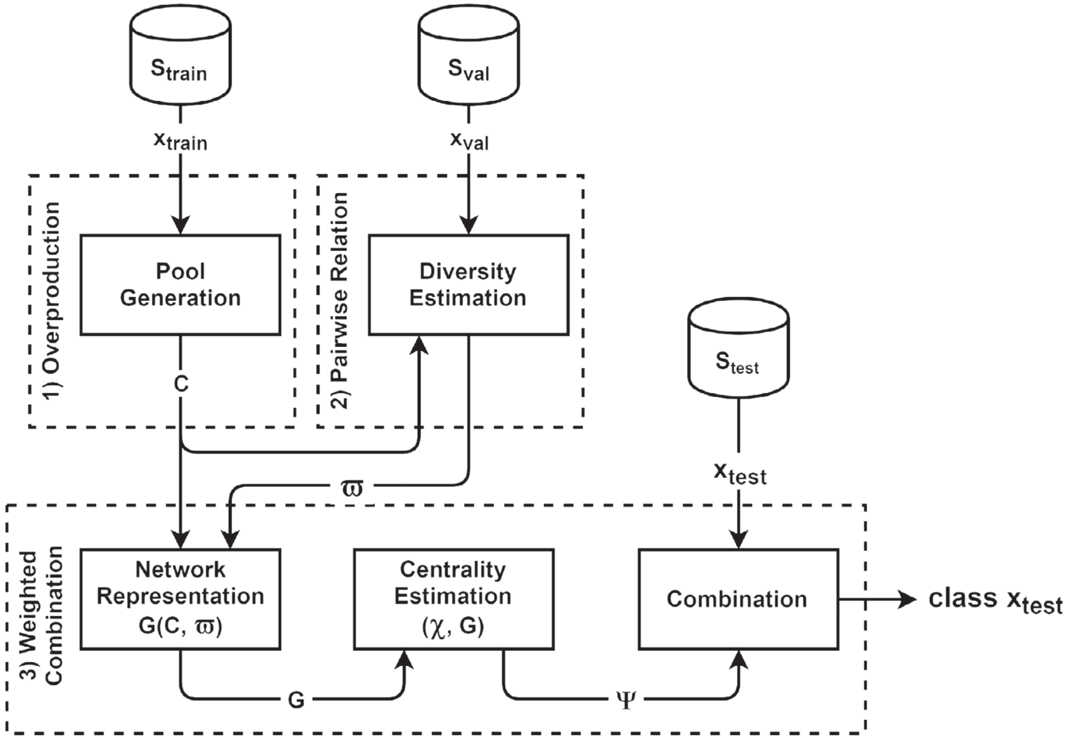

FIGURE 2 Overview of the proposed weighted combination method, centrality scoreŰbased fusion, where S is a dataset, train means training, val means validation, test means test, x is an instance, C is the pool of classifiers, ϖ is the pairwise measure, $G(C,\varpi)$ is the ensemble network, χ is a centrality measure, and Ψ is the computed classifiers's weight

5.2 | Pairwise relationship

The pairwise relationship estimates how each pair of classifiers works together. In pairwise diversity measures, the goal is to know how different one classifier is from another using any diversity measure presented in Section 3.3, such as QS, CC, Dis, DF, or KS. In the case of a pairwise relationship such as N_{10}, the main objective is to know how many times the first classifier is correct, whereas the other is not. In this phase, the pool of classifiers $C = \{C_1, \ldots, C_t\}$ is used to estimate the pairwise relationship $\varpi(Ci, Cj)$ between any classifier to all others. For this, a validation set is used to avoid overfitting.

5.3 | Weighted combination

The CSBF approach differs from the literature in one main aspect: the relationship among the classifiers determines how important is each classifier for the ensemble. This section describes the core of the proposal divided into three steps as follows:

- Network representation: the ensemble network is defined as the pool C and the pairwise relationship ϖ among them. A vertex represents a classifier in C, whereas the pairwise measure

ϖ can be represented by edges in case of symmetric relations or arcs in case of asymmetric. Symmetric relationships are those in which $\varpi(C_i, C_j) = \varpi(C_j, C_i)$. Asymmetric relationships consider $\varpi(C_i, C_j) \neq \varpi(C_j, C_i)$. Any diversity measure in Section 3.3 is symmetric, whereas only the relationship N_{01} and N_{10} are asymmetric. The output of this step is a direct or indirect network $G(V, E)$, where $V = C$ and $E = \varpi$.

- Centrality estimation: a network (represented as a graph) is an input to compute a centrality measure. Classical centrality measures such as betweenness, closeness, degree, and eigenvector are appropriate to deal with symmetric relationships, whereas indegree and outdegree are appropriate to deal with asymmetric. The parameter χ is referred to any of the mentioned centrality measures. The centrality measures present in Section 2.2 score the importance of the classifiers. It can give more importance to classifiers with strong or weak relationships. The output of this step is the weight Ψ for each classifier in C.
- Combination: in this stage, CSBF already weighted the influence of classifiers, Ψ, stated as the importance of the classifier for the whole ensemble. The importance is estimated by a centrality measure which analyzes the classifiers' relationships. Finally, the classifiers' vote is weighted by its influence to compose the ensemble's final decision.

5.4 | The CSBF algorithm

Figure 2 shows a simplified overview of the proposed method. Some aspects were retained to facilitate understanding. Thus, a detailed description of the CSBF method is given through the Algorithm 1 to make clear some important aspects concerning the ensemble network.

Algorithm 1 CSBF($C, \varpi, \chi, \tau, \varphi$)

```
Input: Pool of classifiers $C = \{c_1, c_2, ..., c_T\}$
Input: Pairwise measure ($\varpi$), where $\varpi(c_i, c_j)$
Input: Centrality measure ($\chi$)
Input: Normalization ($\tau$)
Input: Inverse Function ($\varphi$)
Output: Centrality values $\Psi$
$V \leftarrow C$
$E \leftarrow \varpi$
if $\varphi = TRUE$ then
    for $E_i \in E$ do
        $E_i = F^{-1}(E_i)$
    end for
end if
if $\tau = TRUE$ then
    $E = normalize(E)$
end if
Build network $G(V, E)$
$\Psi \Leftarrow$ Compute centrality ($\chi, G$)
return $\Psi$
```

As we can see, first, a pool of classifiers C is required. The parameter ϖ is responsible for scoring the classifiers affinity by a pairwise arrangement. For this purpose, the pairwise diversity measures presented in Section 3.3 are useful. In the case of symmetric pairwise relationships, the number of pairs of classifiers to be evaluated is defined as a combination $\text{Comb}_{T,2}$ where T is the size of pool C and 2 is the pair arrangement. For instance, a pool of $T = 100$ classifiers has $\text{Comb}_{100,2} = 4950$ pairwise symmetric relationships. On the other hand, asymmetric relationships is a permutation $P_{T,2}$, so for the same pool size, there are 9900 pairs to be processed. The centrality measure χ for symmetric evaluation can be betweenness, closeness, degree, and eigenvector, whereas indegree centrality can handle asymmetric evaluation. The normalization τ and the inverse function φ are used to adapt some centrality measures when necessary. After fulfilling the requirements, the pool of classifier C is used to form the group of vertices V in line 1. A computed pairwise measure ϖ for the given pool C is used to define the edges E (line 2). In lines 3 to 7, the inverse function is applied to some centrality measures depending on the meaning of the edge's weight. It is explained later in this section. The normalization process (lines 8-10) is optional, and the goal is to avoid negative weights in E. The graph $G(V, E)$ must be built (line 11), which is referred to in this document as an ensemble network. To obtain the weight of the classifier Ψ, that is, the score of the importance of each classifier in the ensemble, a chosen centrality measure χ and the graph G is needed (line 12). The score of importance Ψ is related to the importance of each classifier, providing a classifier's final weight for the combination. Alternatively, the classifier's accuracy α_{c_t} can be multiplied to Ψ_t to provide the final weight. Therefore, the method differs from the literature by considering the relationships among classifiers instead using the classifier's accuracy only.

Table 2 presents the interpretation of each centrality measure regarding the weight of the edge. From the centrality's point of view, degree assigns high scores to higher weight relationships, whereas betweenness centrality does the opposite. For instance, the pairwise diversity DF assigns lower scores for good relationships, that is, the pair of classifiers usually is correct. In an ensemble network using DF relations, betweenness, and closeness can be estimated directly. On the other hand, degree and eigenvector cannot appropriately estimate the centrality without adapting the set of edges. In these cases, the parameter φ is TRUE. This parameter means that the weight of the pairwise relationships must be reversed, so that the centrality measure can estimate the

TABLE 2 Interpretation of the centrality measures concerning the edge's weight

Centrality measure	Betweenness	Closeness	Degree	Eigenvector
Edges' weight relation with high centrality	↓	↓	↑	↑

TABLE 3 Diversity expressed as a high (↑) or low value (↓) depending on the centrality measure used to estimate the classifiers importance

Centrality	DF	CC	Dis	QS	KS
Degree	↑	↑	↓	↑	↑
Betweenness	↓	↓	↑	↓	↓
Closeness	↓	↓	↑	↓	↓
Eigenvector	↑	↑	↓	↑	↑

Abbreviations: CC, correlation coefficient; DF, double-fault; Dis, disagreement; QS, Q statistics; KS, kappa statistic.

importance of the classifiers correctly. Therefore, the centrality measure and the pairwise measure must be considered together in the CSBF algorithm, adjusting the weight in the relationships in such a way that centrality can use it properly for its estimation, as seen in Table 3.

Table 4 presents the parameters of the proposed method, ϖ stands for the pairwise diversity measure, χ represents the centrality measure, φ represents the inverse function, being true (T) or false (F), and τ is the normalization step, which also can be V or F. Therefore, as seen in Table 4,

TABLE 4 Parameters suggested to use the proposed approach as the combination method

ϖ	χ	φ	τ
DF	Degree	T	T
	Betweenness	F	F
	Closeness	F	F
	Eigenvector	T	T
CC	Degree	T	T
	Betweenness	F	T
	Closeness	F	T
	Eigenvector	T	T
Dis	Degree	F	F
	Betweenness	T	T
	Closeness	T	T
	Eigenvector	F	F
QS	Degree	T	T
	Betweenness	F	T
	Closeness	F	T
	Eigenvector	T	T
KS	Degree	T	T
	Betweenness	F	T
	Closeness	F	T
	Eigenvector	T	T
PM	Degree	T	T
	Betweenness	F	F
	Closeness	F	F
	Eigenvector	T	T
MOIA	Degree	T	T
	Betweenness	F	F
	Closeness	F	F
	Eigenvector	T	T

Abbreviations: CC, correlation coefficient; DF, double-fault; Dis, disagreement; F, false; KS, kappa statistic; MOIA, measurement of interrater agreement; QS, *Q* statistics; PM, product-moment; T, true.

every time φ is TRUE, the normalization process is used. Another suggestion is the use of the normalization process for every diversity measure that can assume values equal or below zero ($\varpi_{ci,cj} \leq 0$). It is worth mention the importance of the parameters as some centrality measures consider more important edges with high weights, whereas others are in the opposite direction.

In asymmetric networks, the relationship between the classifiers C_i and C_j may be different from the relation between C_j and C_i. A directed network must be used to represent asymmetric relationships, where its arrows indicate the direction of the relationship pointing to the correct classifier. The centrality measure χ for directed networks can be indegree or outdegree. Indegree centrality is more appropriate to deal with the relationship N_{10}, so it can measure the weight of the incoming arrows, that is, the arrows pointing the correct classifier.

6 | EXPERIMENTS

We carried out experiments on 40 classification problems extracted from the following different machine learning repositories: UCI machine learning repository,[86] knowledge extraction based on evolutionary learning (KEEL) repository,[87] Ludmila Kuncheva Collection (LKC) of real medical data,[88] STATLOG project,[89] and artificial datasets generated with the Matlab PRTools toolbox.[90] These datasets present only numeric features with no missing values, a varied number of instances, attributes, classes, and imbalance ratio (proportion of the number of instances in the majority class to the number of instances in the minority class) as presented in Table 5. There are 21 classification problems with 2 classes, 7 problems with 3 classes, 2 problems with 4 classes, 1 problem with 5 classes, 3 problems with 6 classes, 3 problems with 7 classes, 1 problem with 8 classes, and 2 problems with 10 classes. Ecoli, PageBlocks, WineQRed, WineQWhite, and Yeast are the most imbalanced datasets. The motivation for using such different machine learning problems is to obtain a general performance of CSBF compared with the literature.

The experimental protocol was based on 6-fold cross-validation (three for training S_{train} (=50% of the original database), two for validation S_{val} (≅32.3%) and one (≅16.7%) for testing S_{test}. The S_{val} is used to estimate accuracy and diversity measures avoiding overfitting. A stratified sampling was applied for class distribution balance in all subsets (S_{train}, S_{val}, and S_{test}). Bagging was used to create a pool of $T = 100$ weak and diverse classifiers, and each bag used to train a classifier has 66% of the S_{train} size. This bag size should provide more diversity than size 100% as observed in Reference 91 and also is used in Reference 92. The use of perceptron as the base classifier is recommended in Reference 93 to detect small differences in bags. With this in mind, we used a more robust variation of the base classifier named perceptron with minimum square error.[94] This base classifier is found in the Weka Data Mining Tool[95] (version 3.9.1). All other parameters were maintained default which, according to Amancio et al[96] perform well, avoiding the task of finding optimal setup. All individual classifiers trained have more than 50% of accuracy, then the classifiers can perform better than random guessing. CSBF also needs a tool for creating the network, so the graphstream[97](version 1.3) is used to assess all the centrality measures presented in Section 2.2.

Two sets of experiments are performed. First, in Section 6.1, several pairwise diversity measures and centrality measures are evaluated to observe how the importance of the classifier in the ensemble network is related to ensemble's accuracy. The analysis of this section supports the choice of the centrality measure for the proposed method, and also the type of relationship between classifiers is the most promising. Then, in Section 6.2, we compared the proposed method

TABLE 5 The main characteristics of each classification problem

Base	# I	# F	# C	IR	Repository
Australian	690	14	2	1.25	UCI
Banana	2000	2	2	1.00	PRTools
Blood	748	4	2	3.20	UCI
CMC	1473	9	3	1.89	UCI
CTG	2126	21	3	9.40	UCI
Dermatology	358	34	6	5.55	KEEL
Diabetes	766	8	2	1.86	UCI
Ecoli	336	7	8	71.50	UCI
Faults	1941	27	7	12.24	UCI
German	1000	24	2	2.33	UCI
Glass	214	9	6	8.44	UCI
Haberman	306	3	2	2.78	UCI
Heart	270	13	2	1.25	UCI
ILPD	579	10	2	2.51	UCI
Ionosphere	351	34	2	1.79	UCI
Laryngeal1	213	16	2	1.63	LKC
Laryngeal3	353	16	3	4.11	LKC
Lithuanian	2000	2	2	1.00	PRTools
Liver	341	6	2	1.40	UCI
Magic	19 020	10	2	1.84	UCI
Mammo	830	5	2	1.06	UCI
Monk	432	6	2	1.12	KEEL
Optdigits	5620	64	10	1.03	KEEL
PageBlocks	5473	10	5	175.46	UCI
Phoneme	5404	5	2	2.41	ELENA
Ring	7400	20	2	1.02	KEEL
Segmentation	2310	19	7	1.00	UCI
Sonar	208	60	2	1.14	UCI
ThyroidNew	215	6	3	5.00	UCI
Vehicle	846	18	4	1.10	UCI
Vertebral2C	310	6	2	2.10	UCI
Vertebral3C	310	6	3	2.50	UCI
WDBC	569	30	2	1.68	UCI
WDVG	5000	21	3	1.03	UCI
Weaning	302	17	2	1.00	LKC

(Continues)

TABLE 5 (Continued)

Base	# I	# F	# C	IR	Repository
Wifi	2000	7	4	1.00	UCI
Wine	178	13	3	1.48	UCI
WineQRed	1599	11	6	68.10	UCI
WineQWhite	4898	11	7	439.60	UCI
Yeast	1484	8	10	92.60	UCI

Abbreviations: # C, number of classes; # F, number of features; # I, number of instances; IR, imbalance ratio.

against 15 related works. Finally, we discuss about the most important factors that affect the performance of the method in Section 6.3.

6.1 | Evaluation of pairwise diversity and centrality measures

The goal of these experiments is to explain how we compared the pairwise relationships, the classifier prominence, and the ensemble accuracy to define the best set of parameters for the CSBF method. Then, those parameters are explored to explain the relationship between diversity and accuracy.

We assessed a combination of seven pairwise diversity measures (DF, CC, Dis, QS, KS, PM, and MOIA) and also four different weighted centrality measures (betweenness, closeness, degree, and eigenvector), resulting in 28 experiments. Besides these symmetric measures, an asymmetric relationship N_{10} is evaluated with indegree centrality. It is worth mention that in these experiments, sometimes the values of the QS and the CC diversity measures were set to 1.0 to avoid a division by zero, as suggested in Reference 98. It occurs when at least one of the possible relationships (N_{00}, N_{01}, N_{10}, or N_{11}) between two classifiers described in Table 1 is not present, so measuring the pairwise diversity can be an issue. All the centrality measures use the edge weight to compute a score, reflecting the importance of the classifier for the given relationship represented in the ensemble network. For each pairwise diversity, an adaptation of the edge weight is needed. Thus, measures that evaluate prominence using more weight like the degree is capable of finding the best score for classifiers successfully. On the other hand, centrality measures such as betweenness uses the geodesics to score classifiers meant that less is more. Therefore, the network analysis depends on the meaning of the pairwise relationship for the ensemble performance (measure score means high/low diversity) and the centrality measures (high edge weight means more/less important). As a reminder, this important step is presented in Tables 2 to 4 of Section 5.4.

After performing the experiments considering the selected 40 classification problems, we computed Friedman and Nemenyi statistical tests. Friedman assesses the ranks according to the performance of the compared approaches, so the best receives 1, the second 2, and so on. Two or more approaches with the same performance have their nearest ranks (upper and lower) averaged. The average rank considers all datasets. Therefore, as low as the average rank, better the approach.

Figure 3 shows the rank created that suggests the best setup: indegree centrality of classifiers computed in a network using the asymmetric relation N_{10} (CSBF: indegree N_{10}). Indegree centrality computed how much each classifier complement the others. So, it suggests that weight

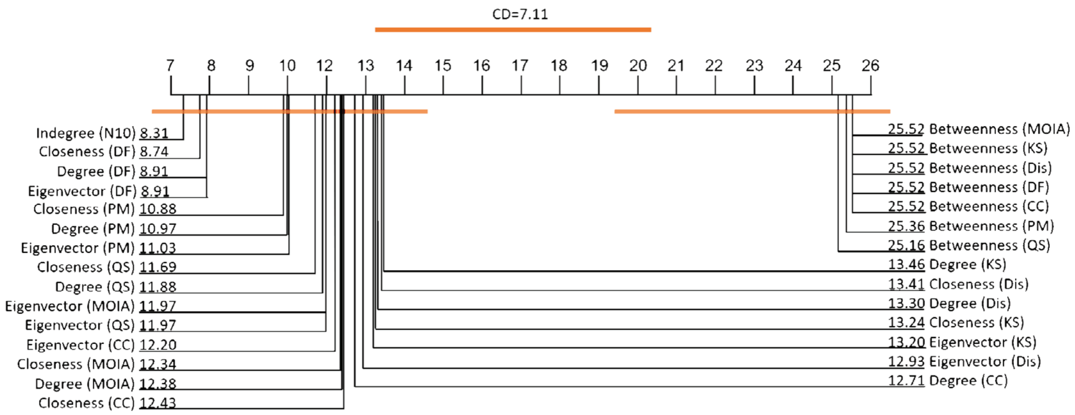

FIGURE 3 Friedman test and Nemenyi post-hoc test to assess pairwise diversity and centrality. The centrality measures are estimated using different weighted graphs: Q statistics (QS), correlation coefficient (CC), disagreement (Dis), double fault (DF), kappa statistics (KS), the pairwise MOIA, product moment (PM), and the asymmetric relation N_{10} [Color figure can be viewed at wileyonlinelibrary.com]

classifiers by their capacity to complement other errors are the most promising strategy. Very close to the indegree N_{10} rank is seen closeness DF, degree DF, and eigenvector DF. Both were estimated using DF as the pairwise relation represented in the ensemble network. Thus, classifiers avoiding mutual errors (DF) with other ensembles are promising. The worse results were observed for betweenness centrality, for any pairwise measure forming the edges of the given network. It suggests that classifiers frequently present in small groups that highly disagree with each other are poorly related to the ensemble's accuracy. Those small groups are the geodesics in which the betweenness centrality scores the classifiers by their presence in these geodesics.

Figure 4 presents the Friedman and post-hoc Nemenyi tests for the best seven approaches. It is observed that the diversity measures that most contribute to the ensemble accuracy are PM, DF, and N_{10}. The approaches performed similarly according to the critical distance (CD). However, Indegree N_{10} is still the best performing approach in regard to the rank score.

Considering the pairwise diversity measures, DF is a tested pairwise relation that detects classifiers that commit simultaneous errors. It is particularly useful for combining classifiers because classifiers equally wrong cannot improve the performance of an ensemble of classifiers. On the contrary, avoiding these classifiers should improve classification performance. Therefore, DF is the symmetric measure more related to the ensemble's accuracy, according to the average rank. The best centrality measures for symmetric relations were closeness, eigenvector, and degree, as seen in DF pairwise relations (the most promising ranks observed). This experiment also shows the lack of relationship between the analysis of diversity relations with ensemble accuracy, due

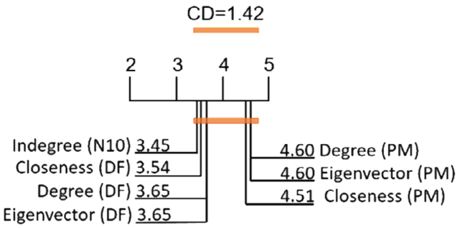

FIGURE 4 Friedman test and Nemenyi post-hoc test to assess pairwise diversity and centrality for the best seven approaches. The centrality measures are double fault (DF), product moment (PM), and the asymmetric relation N_{10} [Color figure can be viewed at wileyonlinelibrary.com]

to classic measures such as KS, Dis, and CC did not perform so well as DF, which only detects mutual errors and is not concerned to how different the pair of classifiers are.

Although the CSBF:indegree N_{10} is statistically different and perform better than the other measures evaluated, why this weighting procedure works must be clarified. Table 6 presents the CSBF approach in a directed ensemble network, in which the centrality measure indegree estimates the classifier's weight considering N_{10} relationship, that is CSBF:indegree N_{10}. The weights of the classifiers is compared with individual accuracy and average diversity. The average diversity is the average of the incoming arrows. Classifier id "52" presents the higher weight and also is one of the most accurate (estimated at the S_{val}). At the end of Table 6, Pearson correlation tests are reported. The correlation between classifier's weight and diversity relation (N_{10}) is strong (as expected) and also a strong correlation between the classifier's weight with accuracy (0.68). The average diversity of a classifier is the average of its pairwise relationships; in this case, high values are preferred. The average diversity of the ensemble is the average of instances in which only one classifier is correct, so a classifier prominent is the one whose average is higher than the ensemble average. The classifier "3" presented the lower weight and also one of the worst individual accuracy. Its average diversity is lower than the ensemble. Table 6 showed that the most important classifier is accurate, but, most important, frequently it performs correct classification while its pairs fail.

In this section, we provided the information concerned to the best set of pairwise relationship and centrality measure to use in the proposed approach. The amount of 40 classification problems and the k-fold cross-validation ($k = 6$) result in the creation of 240 different ensembles. Then, the 29 combinations of the pairwise relationship with the centrality measure used to score the classifiers lead to the analysis of 6960 ensembles scored differently. Therefore, we choose CSBF:indegree N_{10} as the CSBFs best set based on the statistical analysis of the results, which suggests that computing the importance of the classifiers by the CSBF approach is promising when the ensemble network is built using N_{10} as the pairwise relationship and analyzed by the indegree centrality. The following section compares the CSBFs best set with the literature.

6.2 | Comparison with state-of-art methods

We presented the performance of CSBF:Indegree N_{10}—which is the CSBF approach computed in the symmetric network (N_{10} relation) using indegree centrality to calculate the most influent classifier—against 15 literature methods: (i) the MV, (ii) the WMV by accuracy, (iii) the PW, (iv) the KWMV, (v) the BC, (vi) PV, (vii) RSWV, (viii) QBWWV, (ix) BWWV, (x) DTE, (xi) DS combination, (xii) SR, (xiii) PR, (xiv) MAR, and (xv) medium rule (MER). All of these approaches were described or at least referenced in Section 4.

Table 7 shows the new approach compared to the combination rules. The approach reaches the best result in 26 out of 40 classification problems and is also statistically different from DS, DTE, MAR, and PR combination rules. Table 8 presents the comparison of the approach with weighted majority combination schemes and the classic MV.

Table 8 shows the average accuracy and corresponding SD of the proposed method against nine fusion methods of the literature. As can be seen, the CSBF method reaches the best possible result in 17 of 40 classification problems, whereas the best competitors (PW and PV) show the best possible result in ninth and eighth classification problems, respectively. CSBF is also statistically different from 9 out of the 15 related works (Tables 8 and 7).

TABLE 6 A case of success of indegree centrality estimated on N_{10} relations (directed graph)

Id	Weight	Accuracy	Avrg div	Id	Weight	Accuracy	Avrg div
0	5858	67.23	59.17	35	5293	65.07	53.46
1	6331	66.46	63.95	36	5235	65.38	52.88
2	6044	67.23	61.05	37	5047	65.69	50.98
3	4361	61.67	44.05	38	5696	66.00	57.54
4	6343	67.54	64.07	39	5985	66.00	60.45
5	5460	64.45	55.15	40	5658	64.91	57.15
6	6532	68.16	65.98	41	5988	67.39	60.48
7	5205	62.29	52.58	42	5735	67.23	57.93
8	5319	65.07	53.73	43	5336	65.22	53.90
9	5916	67.54	59.76	44	5885	68.47	59.44
10	5129	64.61	51.81	45	5999	65.07	60.60
11	5064	65.69	51.15	46	5087	64.14	51.38
12	5689	66.46	57.46	47	5323	63.37	53.77
13	6737	68.01	68.05	48	5879	68.47	59.38
14	5420	65.53	54.75	49	5704	66.31	57.62
15	5480	63.37	55.35	50	5638	66.00	56.95
16	6793	66.92	68.62	51	5411	66.77	54.66
17	5968	67.70	60.28	52	7628	68.93	77.05
18	5326	65.53	53.80	53	5931	67.08	59.91
19	6835	67.08	69.04	54	5133	66.62	51.85
20	5573	65.84	56.29	55	5912	65.38	59.72
21	5163	66.15	52.15	56	5981	66.62	60.41
22	6073	66.46	61.34	57	5530	65.53	55.86
23	5115	66.15	51.67	58	6040	68.32	61.01
24	5925	68.01	59.85	59	5658	66.31	57.15
25	6949	67.54	70.19	60	5292	65.53	53.45
26	6250	67.39	63.13	61	5064	64.45	51.15
27	5460	66.15	55.15	62	5499	66.77	55.55
28	6207	67.08	62.70	63	6231	65.53	62.94
29	5668	65.07	57.25	64	6677	68.32	67.44
30	6218	67.70	62.81	65	5599	66.92	56.56
31	5757	65.53	58.15	66	5568	66.92	56.24
32	6015	67.54	60.76	67	6748	67.39	68.16
33	4857	65.69	49.06	68	4912	63.68	49.62
34	5958	65.84	60.18	69	5202	65.07	52.55

(Continues)

TABLE 6 (Continued)

Id	Weight	Accuracy	Avrg div	Id	Weight	Accuracy	Avrg div
70	6419	67.39	64.84	85	5812	67.23	58.71
71	6179	68.32	62.41	86	6061	65.22	61.22
72	6176	67.85	62.38	87	6046	67.23	61.07
73	6142	66.15	62.04	88	5468	67.08	55.23
74	7173	68.16	72.45	89	5970	67.08	60.30
75	6558	69.71	66.24	90	5969	67.39	60.29
76	5493	63.52	55.48	91	5252	63.83	53.05
77	6659	69.55	67.26	92	6099	66.92	61.61
78	5779	66.62	58.37	93	5899	65.07	59.59
79	5587	67.23	56.43	94	5511	65.69	55.67
80	6568	66.46	66.34	95	6436	66.62	65.01
81	5555	66.77	56.11	96	6248	67.70	63.11
82	5549	60.90	56.05	97	5717	65.38	57.75
83	5866	65.22	59.25	98	5048	61.05	50.99
84	6106	66.15	61.68	99	6461	68.62	65.26
Pearson weight/accuracy							0.68
Pearson weight/diversity							1
Pearson accuracy/diversity							0.68
Avrg diversity ensemble							58.82

Abbreviation: avrg div, average diversity regarding the classifier's direct neighbors.

Centrality score-based fusion is also compared with all other approaches in a pairwise fashion using the sign test.[99] Figure 5 presents the number of wins, ties, and losses in a grouped column. The dashed line represents the critical value (cv), which uses the number of experiments (n_{exp}) for its estimation. The null hypothesis H_0 stands for statistically equivalent methods. Otherwise, the rejection of the hypothesis suggests that one approach is better than the other. Equation (24) presents the (cv). CSBF method shows a significantly better result when compared with literature methods. The critical value (cv = 25.20) is obtained from the $n_{\text{exp}} = 40$ with a significance level $\alpha = .05$ ($z_\alpha = 1.645$) using Equation (24[100]). According to cv, all literature methods perform worse than CSBF, except PW, KWMV, and PV, in which the number of wins is not higher than cv. In summary, when considering the whole set of experiments, the CSBF won in 422 out of 600 experiments (70.33%), lost in 142 cases (23.67%), and tied in 36 cases (6.00%).

$$\text{cv} = \frac{n_{\text{exp}}}{2} + z_\alpha \times \frac{\sqrt{n_{\text{exp}}}}{2}. \tag{24}$$

Figure 6 presents a statistical analysis using the Friedman test and the Nemenyi post-hoc test. As can be seen, the proposed method is not statistically different from most of the methods evaluated. A thorough analysis of Figure 6 shows that the proposed method is statistically different

TABLE 7 Average accuracy and SD of each evaluated weighting approach based on 6-fold cross-validation

Base	MAR	MER	PR	SR	DS	DTE	CSBF: I N_{10}
Australian	87.10 ± 1.84	86.67 ± 1.48	86.82 ± 1.37	86.67 ± 1.48	86.24 ± 1.70	86.24 ± 1.27	86.09 ± 1.00
Banana	84.70 ± 0.71	**84.75** ± 0.80	**84.75** ± 0.80	**84.75** ± 0.80	84.75 ± 0.72	84.75 ± 0.72	84.50 ± 0.93
Blood	78.20 ± 2.20	78.61 ± 2.63	**78.74** ± 2.59	78.61 ± 2.63	72.72 ± 3.63	72.72 ± 3.63	78.07 ± 2.22
CMC	51.33 ± 3.66	51.12 ± 2.90	51.12 ± 2.77	51.12 ± 2.90	51.87 ± 3.28	51.80 ± 2.96	**51.53** ± 3.18
CTG	88.57 ± 1.37	88.95 ± 1.13	88.62 ± 1.35	88.95 ± 1.13	89.14 ± 1.12	89.23 ± 1.23	**89.18** ± 1.15
Dermatology	**97.49** ± 2.09	97.49 ± 1.86	97.49 ± 1.86	97.49 ± 1.86	**97.49** ± 1.86	**97.49** ± 1.86	96.94 ± 1.77
Diabetes	76.11 ± 2.47	75.98 ± 3.63	76.11 ± 3.43	75.98 ± 3.63	75.98 ± 2.81	75.85 ± 2.76	**76.24** ± 4.07
Ecoli	85.42 ± 4.42	86.31 ± 3.21	86.31 ± 3.21	86.31 ± 3.21	86.61 ± 3.05	85.72 ± 3.42	**86.91** ± 3.04
Faults	68.42 ± 1.34	**71.56** ± 1.14	69.91 ± 0.90	**71.56** ± 1.14	71.87 ± 0.75	71.61 ± 0.87	71.51 ± 1.19
German	73.60 ± 1.59	74.91 ± 2.01	74.81 ± 1.87	74.91 ± 2.01	74.20 ± 2.45	74.40 ± 2.65	**75.11** ± 1.68
Glass	61.19 ± 4.27	64.93 ± 4.88	62.14 ± 4.77	64.93 ± 4.88	60.24 ± 7.65	59.31 ± 7.31	**64.95** ± 4.77
Haberman	**75.49** ± 3.88	74.84 ± 3.82	75.49 ± 5.03	74.84 ± 3.82	71.90 ± 3.33	71.90 ± 3.33	75.16 ± 3.87
Heart	84.08 ± 3.94	83.71 ± 5.54	83.71 ± 4.57	83.71 ± 5.54	84.08 ± 5.04	83.34 ± 5.84	**84.45** ± 4.63
ILPD	70.81 ± 4.24	**71.34** ± 2.45	**71.34** ± 2.45	**71.34** ± 2.45	67.53 ± 4.70	67.36 ± 4.64	71.16 ± 3.10
Ionosphere	**86.87** ± 5.69	85.45 ± 6.03	86.02 ± 5.56	85.45 ± 6.03	87.72 ± 5.74	88.01 ± 5.43	85.73 ± 6.19
Laryngeal1	80.74 ± 2.03	**83.60** ± 3.63	82.17 ± 2.54	**83.60** ± 3.63	82.67 ± 4.81	82.67 ± 4.81	82.67 ± 4.81
Laryngeal3	71.42 ± 6.66	74.52 ± 5.12	72.55 ± 7.35	74.52 ± 5.12	70.84 ± 5.75	69.99 ± 5.86	**74.82** ± 5.75
Lithuanian	82.25 ± 2.04	83.05 ± 2.38	83.05 ± 2.38	83.05 ± 2.38	83.25 ± 2.62	83.25 ± 2.62	**83.30** ± 2.23
Liver	68.34 ± 4.56	68.93 ± 2.48	68.34 ± 1.88	68.93 ± 2.48	68.63 ± 2.47	68.63 ± 2.47	**69.51** ± 3.80
Magic	79.02 ± 0.63	79.31 ± 0.69	79.31 ± 0.68	79.31 ± 0.69	79.31 ± 0.73	79.31 ± 0.73	**79.39** ± 0.69
Mammo	84.10 ± 1.95	**84.34** ± 2.54	**84.34** ± 2.54	**84.34** ± 2.54	84.22 ± 2.47	84.22 ± 2.47	83.98 ± 2.36
Monk-2	77.32 ± 4.99	**87.73** ± 3.25	85.42 ± 4.59	**87.73** ± 3.25	87.73 ± 3.25	87.73 ± 3.25	86.81 ± 5.05

(Continues)

TABLE 7 (Continued)

Base	MAR	MER	PR	SR	DS	DTE	CSBF: I N_{10}
Optdigits	95.52 ± 0.52	96.01 ± 0.45	95.75 ± 0.56	96.01 ± 0.45	96.03 ± 0.50	96.05 ± 0.54	**96.03** ± 0.58
PageBlocks	94.85 ± 0.48	95.09 ± 0.39	**95.12** ± 0.38	95.09 ± 0.39	95.12 ± 0.36	95.21 ± 0.42	95.12 ± 0.28
Phoneme	76.89 ± 0.86	77.04 ± 0.98	77.04 ± 0.98	77.04 ± 0.98	75.63 ± 1.64	75.63 ± 1.64	**77.35** ± 0.57
Ring	76.11 ± 0.50	76.05 ± 0.64	76.05 ± 0.64	76.05 ± 0.64	75.72 ± 0.43	75.72 ± 0.43	**76.24** ± 0.47
Segmentation	92.38 ± 1.59	92.90 ± 1.20	92.64 ± 1.18	92.90 ± 1.20	92.64 ± 1.16	92.56 ± 1.10	**93.21** ± 1.32
Sonar	74.05 ± 3.66	76.93 ± 6.21	74.54 ± 5.79	76.93 ± 6.21	75.97 ± 5.16	75.49 ± 5.66	**77.41** ± 4.50
ThyroidNew	93.98 ± 3.73	94.44 ± 3.59	**94.91** ± 3.73	94.44 ± 3.59	96.76 ± 1.91	96.76 ± 1.91	**94.91** ± 3.73
Vehicle	77.66 ± 2.74	79.08 ± 2.08	78.49 ± 2.48	79.08 ± 2.08	78.61 ± 2.76	78.84 ± 2.73	**79.31** ± 2.33
Vertebral2C	84.81 ± 4.50	**85.78** ± 3.78	**85.78** ± 3.78	**85.78** ± 3.78	85.45 ± 4.12	85.45 ± 4.12	85.14 ± 2.76
Vertebral3C	**85.50** ± 4.07	85.50 ± 3.03	85.50 ± 3.03	85.50 ± 3.03	86.47 ± 3.80	86.14 ± 3.71	85.18 ± 3.38
WDBC	96.66 ± 1.65	96.84 ± 1.21	96.84 ± 1.21	96.84 ± 1.21	97.01 ± 1.41	97.01 ± 1.41	**97.47** ± 1.07
WDVG	86.24 ± 1.70	86.16 ± 1.65	86.18 ± 1.61	86.16 ± 1.65	86.16 ± 1.70	86.14 ± 1.71	**86.40** ± 1.67
Weaning	78.47 ± 2.16	81.12 ± 3.61	80.79 ± 3.00	81.12 ± 3.61	82.12 ± 3.47	82.12 ± 3.47	**81.45** ± 3.01
Wifi	97.15 ± 0.75	97.20 ± 0.79	97.15 ± 0.90	97.20 ± 0.79	97.15 ± 0.93	97.15 ± 1.05	**97.40** ± 0.93
Wine	96.07 ± 2.99	97.22 ± 3.56	97.78 ± 3.68	97.22 ± 3.56	97.22 ± 3.56	97.22 ± 3.56	**98.33** ± 2.55
WineQRed	57.54 ± 2.14	58.54 ± 1.87	58.41 ± 1.71	58.54 ± 1.87	47.90 ± 3.19	45.96 ± 4.00	**58.73** ± 2.59
WineQWhite	52.63 ± 2.06	53.14 ± 1.96	53.16 ± 1.87	53.14 ± 1.96	36.77 ± 2.95	35.18 ± 2.32	**53.37** ± 1.88
Yeast	54.05 ± 2.32	53.78 ± 2.37	53.85 ± 2.51	53.78 ± 2.37	56.27 ± 1.33	56.54 ± 1.57	**55.93** ± 2.19
Average	79.58 ± 2.62	80.52 ± 2.57	80.21 ± 2.59	80.52 ± 2.57	79.45 ± 2.76	79.27 ± 2.79	**80.67** ± 2.58
SD	12.23 ± 1.53	12.04 ± 1.52	12.21 ± 1.62	12.04 ± 1.52	14.03 ± 1.68	14.30 ± 1.70	11.91 ± 1.55
BR	4	7	7	7	1	1	**26**
WS	+0.00001	0.06576	+0.00424	0.06576	+0.03846	+0.02852	

The best results are in bold.

Abbreviations: +, a significant result; BR, the number of best results the method obtained over all datasets; CSBF, centrality score–based fusion; DS, Dempster-Shafer; DTE, decision templates; MAR, maximum rule; MER, medium rule; PR, product rule; SR, sum rule; WS, Wilcoxon signed test, so the values represent the *P* value.

TABLE 8 Average accuracy and SD of each evaluated weighting approach based on 6-fold cross-validation

Base	BC	BWWV	KWMV	MV	PV	PW	QBWWV	RSWV	WMV	CSBF: I N_{10}
Australian	**86.67** ± 1.48	57.83 ± 2.45	86.24 ± 1.46	86.53 ± 1.57	86.09 ± 1.23	85.95 ± 1.06	57.83 ± 2.45	86.09 ± 1.23	86.09 ± 1.59	86.09 ± 1.00
Banana	84.70 ± 0.83	50.00 ± 0.12	**84.85** ± 0.79	**84.85** ± 0.79	84.75 ± 0.80	84.75 ± 0.65	50.00 ± 0.12	84.75 ± 0.80	**84.85** ± 0.79	84.50 ± 0.93
Blood	**78.61** ± 2.63	76.20 ± 0.29	78.21 ± 2.01	78.21 ± 2.01	78.21 ± 2.01	78.34 ± 2.09	76.20 ± 0.29	77.94 ± 1.96	78.21 ± 2.01	78.07 ± 2.22
CMC	51.06 ± 2.90	39.91 ± 2.96	50.99 ± 2.84	50.79 ± 3.00	50.79 ± 3.03	50.85 ± 3.35	39.71 ± 3.04	20.91 ± 2.18	50.72 ± 2.99	**51.53** ± 3.18
CTG	88.90 ± 1.12	80.25 ± 1.67	89.04 ± 1.03	89.04 ± 1.03	88.99 ± 1.10	89.14 ± 0.97	78.55 ± 1.68	89.04 ± 1.03	89.04 ± 1.03	**89.18** ± 1.15
Dermatology	**97.49** ± 1.86	34.90 ± 2.22	96.94 ± 1.77	96.94 ± 1.77	96.94 ± 1.77	96.94 ± 1.77	34.07 ± 1.90	96.94 ± 1.77	96.94 ± 1.77	96.94 ± 1.77
Diabetes	75.98 ± 3.63	65.92 ± 1.53	75.98 ± 3.65	75.85 ± 3.48	75.85 ± 3.79	75.98 ± 4.01	65.92 ± 1.53	67.32 ± 18.43	75.85 ± 3.48	**76.24** ± 4.07
Ecoli	86.31 ± 3.21	47.32 ± 1.71	86.61 ± 2.68	86.61 ± 2.68	**86.91** ± 2.86	**86.91** ± 2.86	44.34 ± 2.17	86.61 ± 2.68	86.61 ± 2.68	86.91 ± 3.04
Faults	**71.61** ± 1.21	23.29 ± 4.18	71.56 ± 1.34	71.31 ± 1.24	71.51 ± 1.42	**71.61** ± 1.35	20.72 ± 4.16	71.56 ± 1.50	71.36 ± 1.20	71.51 ± 1.19
German	74.80 ± 1.82	69.80 ± 2.55	75.10 ± 1.88	74.81 ± 1.89	**75.20** ± 1.61	75.00 ± 1.74	69.80 ± 2.55	25.40 ± 2.09	75.10 ± 2.00	75.11 ± 1.68
Glass	64.93 ± 4.88	28.04 ± 4.57	64.95 ± 4.77	64.47 ± 5.09	**65.42** ± 4.68	65.41 ± 4.50	27.58 ± 5.47	40.44 ± 20.61	64.47 ± 5.09	64.95 ± 4.77
Haberman	75.16 ± 4.03	72.55 ± 1.60	75.49 ± 3.71	75.49 ± 3.71	**75.82** ± 4.33	75.16 ± 3.87	72.55 ± 1.60	74.51 ± 2.26	75.49 ± 3.71	75.16 ± 3.87
Heart	83.71 ± 5.54	60.00 ± 2.87	83.71 ± 5.54	83.71 ± 5.54	**84.45** ± 4.63	83.71 ± 5.54	60.00 ± 2.87	**84.45** ± 4.63	83.71 ± 5.54	**84.45** ± 4.63
ILPD	71.51 ± 2.25	70.64 ± 1.39	71.51 ± 2.87	**71.68** ± 2.79	70.99 ± 2.81	71.34 ± 2.97	70.64 ± 1.39	27.45 ± 3.24	**71.68** ± 2.79	71.16 ± 3.10
Ionosphere	85.45 ± 6.03	66.66 ± 2.48	85.45 ± 6.03	85.45 ± 6.03	**85.73** ± 6.19	**85.73** ± 6.19	66.66 ± 2.48	85.44 ± 6.36	85.45 ± 6.03	**85.73** ± 6.19
Laryngeal1	**84.08** ± 4.04	41.80 ± 5.14	82.66 ± 4.30	83.12 ± 3.88	82.67 ± 4.81	82.20 ± 4.91	41.80 ± 5.14	82.20 ± 5.85	83.12 ± 3.88	82.67 ± 4.81
Laryngeal3	74.52 ± 5.12	35.98 ± 7.94	73.96 ± 4.88	73.95 ± 4.26	74.25 ± 5.32	74.25 ± 5.32	31.18 ± 7.17	74.53 ± 5.28	73.95 ± 4.26	**74.82** ± 5.75
Lithuanian	83.00 ± 2.34	49.30 ± 0.97	**83.35** ± 2.38	83.25 ± 2.54	83.25 ± 2.43	83.25 ± 2.53	49.30 ± 0.97	83.30 ± 2.40	83.20 ± 2.46	83.30 ± 2.23
Liver	69.22 ± 2.70	44.55 ± 8.09	69.23 ± 3.49	68.34 ± 2.37	69.52 ± 4.84	**69.82** ± 4.11	44.55 ± 8.09	32.25 ± 4.30	68.93 ± 2.48	69.51 ± 3.80
Magic	79.31 ± 0.69	67.23 ± 0.77	79.32 ± 0.70	79.31 ± 0.70	79.32 ± 0.70	79.35 ± 0.69	67.23 ± 0.77	79.32 ± 0.69	79.32 ± 0.70	**79.39** ± 0.69
Mammo	**84.34** ± 2.54	55.54 ± 1.50	84.22 ± 2.57	84.22 ± 2.57	84.22 ± 2.57	84.22 ± 2.57	55.54 ± 1.50	84.22 ± 2.57	84.22 ± 2.57	83.98 ± 2.36
Monk-2	87.73 ± 3.25	47.69 ± 2.85	87.96 ± 3.17	87.27 ± 3.62	88.19 ± 3.18	88.43 ± 3.07	47.69 ± 2.85	**88.66** ± 3.04	87.73 ± 3.25	86.81 ± 5.05

(Continues)

TABLE 8 (Continued)

Base	BC	BWWV	KWMV	MV	PV	PW	QBWWV	RSWV	WMV	CSBF: I N_{10}
Optdigits	96.01 ± 0.45	13.33 ± 0.80	95.91 ± 0.55	95.93 ± 0.55	95.92 ± 0.53	96.03 ± 0.52	12.17 ± 0.45	95.91 ± 0.55	95.92 ± 0.53	**96.03** ± 0.58
PageBlocks	95.09 ± 0.39	91.69 ± 0.35	94.76 ± 0.22	94.74 ± 0.25	94.78 ± 0.20	95.00 ± 0.16	91.67 ± 0.35	94.76 ± 0.22	94.76 ± 0.22	**95.12** ± 0.28
Phoneme	77.05 ± 0.96	70.80 ± 0.37	77.05 ± 0.96	77.07 ± 0.99	77.13 ± 0.98	77.09 ± 0.91	70.80 ± 0.37	**77.55** ± 0.74	77.05 ± 0.96	77.35 ± 0.57
Ring	76.04 ± 0.64	50.69 ± 1.12	75.93 ± 0.51	75.97 ± 0.55	75.99 ± 0.52	76.12 ± 0.54	50.69 ± 1.12	**76.38** ± 0.59	75.93 ± 0.51	76.24 ± 0.47
Segmentation	92.90 ± 1.20	16.80 ± 1.63	92.90 ± 1.19	92.90 ± 1.19	92.95 ± 1.23	92.99 ± 1.24	16.28 ± 1.31	92.90 ± 1.19	92.90 ± 1.19	**93.21** ± 1.32
Sonar	76.93 ± 6.21	55.25 ± 5.28	76.93 ± 5.99	77.39 ± 7.20	76.92 ± 5.54	76.93 ± 5.99	55.25 ± 5.28	35.15 ± 24.19	76.92 ± 7.11	**77.41** ± 4.50
ThyroidNew	94.44 ± 3.59	18.59 ± 5.18	**94.91** ± 3.73	**94.91** ± 3.73	**94.91** ± 3.73	**94.91** ± 3.73	17.66 ± 5.43	**94.91** ± 3.73	**94.91** ± 3.73	**94.91** ± 3.73
Vehicle	79.08 ± 2.08	24.35 ± 2.51	79.20 ± 1.95	79.31 ± 2.02	79.08 ± 2.16	**79.31** ± 2.26	22.11 ± 1.66	79.20 ± 1.95	79.20 ± 1.95	79.31 ± 2.33
Vertebral2C	85.78 ± 3.78	68.09 ± 5.02	85.78 ± 3.61	**86.10** ± 3.91	84.82 ± 3.17	84.50 ± 3.08	68.09 ± 5.02	84.50 ± 3.08	85.78 ± 3.61	85.14 ± 2.76
Vertebral3C	85.50 ± 3.03	57.42 ± 2.27	**85.51** ± 3.57	85.50 ± 3.03	84.86 ± 3.68	84.86 ± 3.68	54.52 ± 3.78	**85.51** ± 3.57	**85.51** ± 3.57	85.18 ± 3.38
WDBC	96.84 ± 1.21	62.92 ± 0.72	97.01 ± 1.12	97.01 ± 1.12	97.01 ± 1.12	**97.54** ± 0.99	62.92 ± 0.72	97.19 ± 1.16	97.01 ± 1.12	97.47 ± 1.07
WDVG	86.16 ± 1.65	34.24 ± 1.08	86.20 ± 1.61	86.18 ± 1.60	86.26 ± 1.67	86.30 ± 1.75	32.66 ± 0.81	86.20 ± 1.61	86.20 ± 1.61	**86.40** ± 1.67
Weaning	**81.45** ± 3.62	55.98 ± 3.23	**81.45** ± 3.62	**81.45** ± 3.62	**81.45** ± 3.01	**81.45** ± 3.01	55.98 ± 3.23	75.57 ± 17.29	**81.45** ± 3.62	**81.45** ± 3.01
Wifi	97.20 ± 0.79	27.30 ± 0.90	97.20 ± 0.84	97.20 ± 0.84	97.20 ± 0.84	97.25 ± 0.87	26.60 ± 0.63	97.20 ± 0.84	97.20 ± 0.84	**97.40** ± 0.93
Wine	97.78 ± 3.68	36.51 ± 3.49	96.11 ± 2.99	97.78 ± 2.49	97.78 ± 2.49	97.78 ± 2.49	36.51 ± 3.49	96.11 ± 2.99	97.78 ± 2.49	**98.33** ± 2.55
WineQRed	58.60 ± 1.88	44.90 ± 2.58	**59.04** ± 2.44	58.41 ± 1.99	58.79 ± 2.36	58.79 ± 2.52	40.09 ± 1.71	14.52 ± 16.07	58.41 ± 1.99	58.73 ± 2.59
WineQWhite	53.14 ± 1.96	39.40 ± 1.60	53.21 ± 1.98	53.10 ± 2.06	53.15 ± 2.04	**53.49** ± 2.10	39.32 ± 1.67	11.52 ± 1.65	53.10 ± 2.10	53.37 ± 1.88
Yeast	53.78 ± 2.44	27.76 ± 2.40	54.85 ± 2.32	53.71 ± 2.36	54.79 ± 2.36	55.46 ± 1.82	26.08 ± 1.96	21.57 ± 14.44	53.78 ± 2.41	**55.93** ± 2.19
Average	80.57 ± 2.59	49.54 ± 2.51	80.53 ± 2.58	80.50 ± 2.55	80.57 ± 2.59	80.60 ± 2.59	48.78 ± 2.48	71.50 ± 4.77	80.50 ± 2.55	**80.67** ± 2.58
SD	12.05 ± 1.54	18.86 ± 1.89	11.90 ± 1.52	12.11 ± 1.58	11.97 ± 1.54	11.93 ± 1.59	19.33 ± 1.92	26.15 ± 6.07	12.09 ± 1.57	11.91 ± 1.55
BR	7	0	6	5	8	9	0	6	5	**17**
WS	0.17702	+0.00001	0.09492	0.06288	+0.02144	0.07508	+0.00001	+0.00012	+0.03572	

The best results are in bold.

+, a significant result; BC, Bayesian combination; BR, the number of best results the method obtained over all datasets; BWWV, best-worst weighted vote; KWMV, Kuncheva weighted majority vote; MV, majority vote; PV, power value; PW, performance weighting; QBWWV, quadratic best-worst weighted vote; RSWV, rescaled weighted vote; WMV, weighted majority vote; WS, Wilcoxon signed test, so the values represent the *P* value.

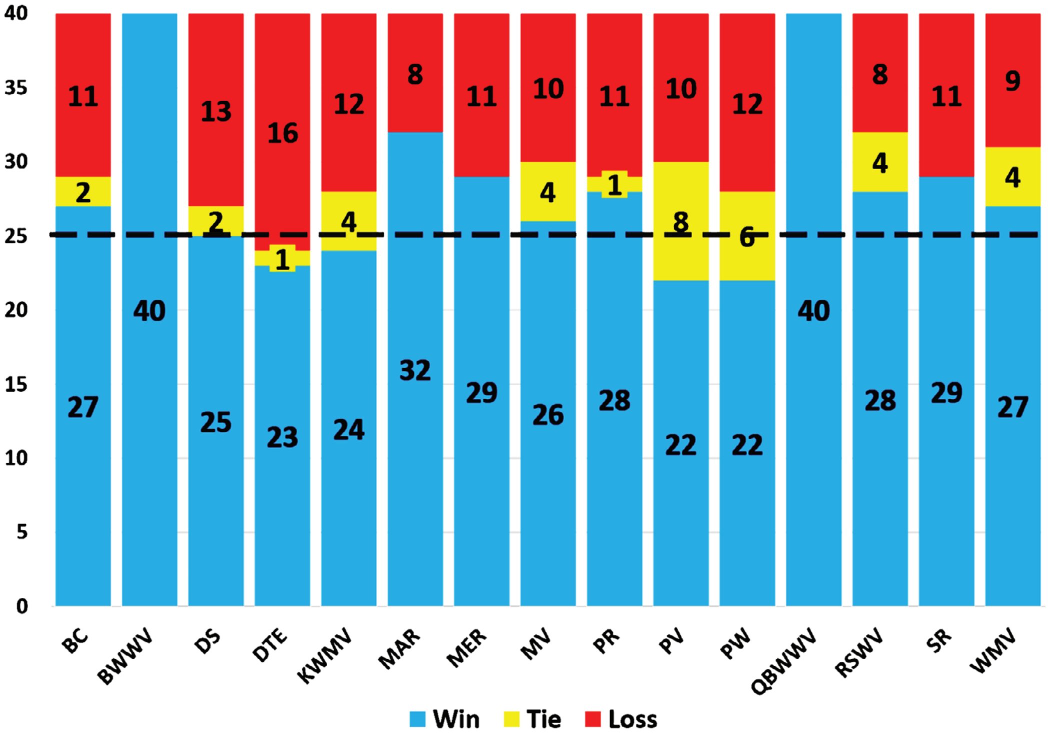

FIGURE 5 Pairwise comparison of the proposed method (CSBF:indegree N_{10}) with 15 literature methods [Color figure can be viewed at wileyonlinelibrary.com]

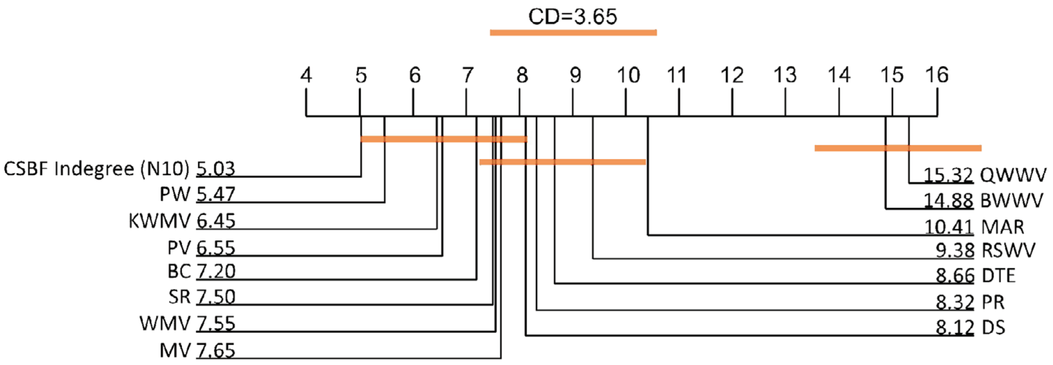

FIGURE 6 Friedman and Nemenyi post-hoc tests comparing the proposed method (CSBF:indegree N_{10}) with the literature [Color figure can be viewed at wileyonlinelibrary.com]

from QWWV, BWWV, MAR, and RSWV (according to CD). The highest score is assigned to CSBF, so it suggests the new approach performs well against the literature.

In another statistical analysis, the Wilcoxon test was performed to compare the proposed method against each of the 15 methods in the literature in a pairwise fashion. The results in Tables 7 and 8 show that the proposed method provides better results compared with the literature methods at $\alpha = .05$ significance level, but that these results are significant for most pairwise comparisons.

6.3 | Discussion

This work aimed to answer some important research questions about the use of centrality measures to provide the importance of each classifier for a static fusion ensemble method. The first research question is related to the improvement of the ensemble accuracy by the use of centrality measures. To answer this question, we represent the ensemble of classifiers in a structure named ensemble network to estimate the importance (centrality) of the classifiers based on the analysis of their relationships. When compared with the literature, it has shown to be an interesting alternative to combine classifiers. Considering the literature approaches (MV, WMV, PW, RSWV, QBWWV, BWWV, DTE, DS, SR, PR, MAR, and MER), the proposed method CSBF:indegree N_{10} was capable of performing better in 422 out of 600 experiments (70.33%), while lost in 142 cases (23.67%).

The ensemble network built to study the CSBF approach involved the study of different pairwise relations, which is the basis to estimate the importance of the classifiers. Diversity measures are some of the most common pairwise relations in ensemble learning. However, there is no consensus on which one is the most related to the ensemble's accuracy.[39] Therefore, we evaluated some well-known diversity measures such as DF, QS, CC, Dis, KS, and the pairwise relation N_{10}, which is used to estimate these diversity measures. These pairwise relations were analyzed by the centrality measures to score the classifiers in the fusion process. The most interesting relations concerning the ensemble's accuracy were N_{10} and DF. The first relation, N_{10}, is the only asymmetric relation analyzed, so indegree centrality was used to score the classifiers according to their importance. It showed to be a promising alternative to weight the classifiers as it identifies and assigns high scores for classifiers that are frequently correct, while its pairs are not. The symmetric pairwise relation DF was the second most interesting relation analyzed, according to the scores provided by eigenvector and degree centrality. So, classifiers that frequently avoid the mutual error (DF) are preferred in comparison with diversity measures concerned on assessing only the difference between classifiers such as Dis, QS, CC, and KS. Product measurement also presents good results, so DF and product measurement, which focus on the mutual error, were the best diversity measures to aim the ensemble's accuracy.

The centrality measure most appropriated to score the importance of the classifiers depends on the pairwise relationship used to represent the ensemble network. Therefore, closeness, eigenvector, and degree centrality performed better concerned symmetric relationships, whereas indegree was the most interesting for the asymmetric measure analyzed. Indegree centrality computed the number of relationships directed to a classifier (N_{10}), which reflects how much that classifier complements the ensemble's errors. Such a classifier received more score (weight), leading to a more influent vote. Degree centrality was computed by summing the weight of the symmetric relationship, so the most influent classifiers observed highly avoids the mutual errors (DF) regarding its pairwise relationship with the other ensemble members. This weighting method lost only for closeness centrality in the context of the symmetric relationships; however, this centrality requires the estimation of geodesics, which increases the method complexity.

The answer for the proposed research questions was obtained by a robust experimental protocol using 40 different classification problems, leading to 240 different ensembles (considering the number of datasets and k-fold cross-validation with $k = 6$), to evaluate 29 combinations of pairwise relation and centrality measure per fold, resulting in 174 different scored ensembles per problem and 6960 in total.

7 | CONCLUSION AND FUTURE WORK PERSPECTIVES

We have presented a novel ensemble fusion method based on the concept of centrality in the context of complex network theory. In the proposed CSBF method, the ensemble is represented as a complex network created to reflect the relationship between the classifiers. The importance of each classifier in that network is estimated employing centrality measures, which combines with accuracy, provide the weight used in the fusion process.

The experimental results on 40 classification problems confirmed our main hypothesis. The centrality concept used to represent the importance of classifiers within the ensemble network is a promising strategy for weighting the decisions of the classifiers in the fusion method. Different pairwise relationship and centrality measures were evaluated to find out the best setup for the proposed method. The best results were achieved using the DF pairwise diversity measure to generate the ensemble network and the degree centrality measure to estimate the importance of each classifier. The experimental results showed that the proposed fusion method was able to present the best accuracy on 17 over 40 classification problems when compared with nine different weighting methods in the trainable class label literature (Table 8), whereas the second best method in that comparison presented the best accuracy in just 9 cases. Such comparison was also conducted concerning combination rules. The CSBF method presents the best accuracy on 26 over 40 classification problems when compared with six different nontrainable support functions methods in the literature (Table 7). Among a total of 600 comparisons, the proposed method was able to prevail in 422(70.33%), tie in 36(6.00%), and losing in only 142 cases (23.67%).

Further work is necessary to investigate the best parameters to compose the pairwise relationship between classifiers. The concept of pairwise relationship of classifiers is directly related to edges in complex network. The use of pairwise diversity measures in ensembles is found in several works in the literature.[101-104] However, in this work, the only asymmetric relation tested shows more relationship with ensemble accuracy by the point of view of centrality measures. Thus, another type of asymmetric relationships could be investigated. Another direction for future research is the use of other centrality measures, such as PageRank,[105] Katz,[106] and β-centrality.[107]

ACKNOWLEDGMENTS

This research has been supported by the Brazilian National Council for Scientific and Technological Development (CNPq), grant 306684/2018-2, and by the Coordination for the Improvement of Higher Education Personnel (CAPES), Finance Code 88881.131663/2016-01.

ORCID

Ronan Assumpção Silva https://orcid.org/0000-0001-8190-5474

REFERENCES

1. Latora V, Nicosia V, Russo G. *Complex Networks: Principles, Methods and Applications*. Cambridge, UK: Cambridge University Press; 2017.
2. Moreno JL. *Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama*. Beacon House: Washington, DC; 1934.
3. Cornell J. Using social network analysis to reveal unseen relationships in medieval Scotland. *Digi Scholar Humanit*. 2017;32(2):336-343.
4. Enembreck F, Barthès J-PA. A social approach for learning agents. *Expert Syst Appl*. 2013;40(5):1902-1916.
5. Gomes HM, Enembreck F. *SAE2: Advances On The Social Adaptive Ensemble Classifier for Data Streams*. New York, NY: ACM Press; 2014:798-804.

6. Barddal JP, Gomes HM, Enembreck F. *SFNClassifier*. New York, NY: ACM Press; 2014:786-791.
7. Ma Hao, King Irwin, Lyu Michael R. Learning to recommend with social trust ensemble. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval—SIGIR '09*;*2009;203.
8. Trawinski Krzysztof, Cordon Oscar. A network-based approach for diversity visualization of fuzzy classifier ensembles. Paper presented at: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE); 2016:923-930.
9. Freeman LC. Centrality in social networks conceptual clarification. *Soc Netw*. 1978;1(3):215-239.
10. Anthonisse J. M. The rush in a directed graph. *Stichting Mathematisch Centrum Mathematische Besliskunde*, No. BN 9/71. 1971.
11. Bonacich P. Factoring and weighting approaches to status scores and clique identification. *J Math Sociol*. 1972;2:113-120.
12. Lü L, Chen D, Ren X-L, Zhang Q-M, Zhang YC, Zhou T. Vital nodes identification in complex networks. *Phys Rep*. 2016;650:1-63.
13. Zhou F, Mahler S, Toivonen H. Simplification of networks by edge pruning. *Bisociative Knowledge Discovery. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 7250, Berlin, Heidelberg: Springer Berlin Heidelberg; 2012:179-198.
14. Dianati N. Unwinding the hairball graph: pruning algorithms for weighted complex networks. *Phys Rev E*. 2016;93(1):8.
15. Brandes U. On variants of shortest-path betweenness centrality and their generic computation. *Soc Netw*. 2008;30(2):136-145.
16. Wasserman S, Faust K. *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press; 1994.
17. Boldi P, Vigna S. Axioms for centrality. *Internet Math*. 2014;10(3–4):222–262.
18. Borgatti SP. Centrality and network flow. *Soc Netw*. 2005;27(1):55-71.
19. Bonacich P. Some unique properties of eigenvector centrality. *Soc Netw*. 2007;29(4):555-564.
20. Ruhnau B. Eigenvector-centrality—a node-centrality? *Soc Netw*. 2000;22:357-365.
21. Valente TW, Coronges K, Lakon C, Costenbader E. How correlated are network centrality measures? *Connections*. 2008;28(1):16-26.
22. Zenil H, Kiani Narsis A, Tegnér J. A review of graph and network complexity from an algorithmic information perspective. *Entropy*. 2018;20(8):1-15.
23. Krawczyk B, Minku Leandro L, Gama J, Stefanowski J, Woźniak M. Ensemble learning for data stream analysis: a survey. *Inf Fusion*. 2017;37:132-156.
24. Gomes HM, Barddal JP, Enembreck F, Bifet A. A survey on ensemble learning for data stream classification. *ACM Comput Surv*. 2017;50(2):1-36.
25. Wang S, Minku LL, Yao X. A systematic study of online class imbalance learning with concept drift. *IEEE Trans Neural Netw Learning Syst*. 2018;29(10):4802-4821.
26. Guo H, Li Y, Jennifer S, Gu M, Huang Y, Gong B. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst Appl*. 2017;73:220-239.
27. Deniz A, Yusuf Y. A comparison study on ensemble strategies and feature sets for sentiment analysis. *Lecture Notes Electric Eng*. 2016;363:359-370.
28. Cruz RMO, Robert S, Cavalcanti GDC. Dynamic classifier selection: Recent advances and perspectives. *Inf Fusion*. 2018;41:195-216.
29. Britto Ad S Jr, Robert S, Oliveira LES. Dynamic selection of classifiers - A comprehensive review. *Pattern Recognit*. 2014;47(11):3665-3680.
30. Burduk Robert. Integration base classifiers based on their decision boundary. In: Rutkowski L, Korytkowski M, Scherer R, Tadeusiewicz R, Zadeh L, Zurada J, eds. Paper presented at: Proceedings of the 16th International Conference on Artificial Intelligence and Soft Computing, Part II, ICAISC; June 11–15, 2017; Zakopane, Poland: Springer International Publishing: 3–20.
31. Bryan E, Caitlin VS, Cahill ND, Narayan DA. A comprehensive comparison of graph theory metrics for social networks. *Soc Netw Anal Min*. 2015;5(1):37.

32. Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. *IEEE Trans Pattern Anal Mach Intell*. 1998;20(3):226-239.
33. Ponti MP Jr. Combining classifiers: from the creation of ensembles to the decision fusion. Paper presented at: 2011 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials; 2011; Alagoas, Brazil.
34. Kuncheva Ludmila I. *Combining Pattern Classifiers: Methods and Algorithms*. *2nd*. Hoboken, NJ John Wiley & Sons, 2014; 2014.
35. Kuncheva LI, Whitaker CJ, Shipp CA, Duin RPW. Limits on the majority vote accuracy in classifier fusion. *Pattern Anal Appl*. 2003;6(1):22-31.
36. Li Ye, Xu Li, Wang Ya Gang, Xu Xiao Ming. A new diversity measure for classifier fusion. In: Wang FL, Lei J, Lau RWH, Zhang J, eds. *Multimedia and Signal Processing. Communications in Computer and Information Science*, Springer: Berlin/Heidelberg; 2012:396-403.
37. Giacinto G, Roli F. An approach to the automatic design of multiple classifier systems. *Pattern Recogn Lett*. 2001;22(1):25-33.
38. Ulas C, Koroglu B, Bekar C, Burcak O, Agin OA. Supervised learning approach for the fusion of multiple classifier outputs. *Int J Signal Process Syst*. 2016;4(3):198-203.
39. Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach Learn*. 2003;51(2):181-207.
40. Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP. Ensemble diversity measures and their application to thinning. *Inf Fusion*. 2005;6(1):49-62.
41. Yin X-C, Huang K, Hao HW, Iqbal K, Wang Z-B. A novel classifier ensemble method with sparsity and diversity. *Neurocomputing*. 2014;134:214-221.
42. Gargiulo F, Mazzariello C, Sansone C. Multiple classifier systems: theory, applications and tools. In: Bianchini M, Maggini M, Jain LC, eds. *Handbook on Neural Information Processing, Intelligent Systems Reference Library*. Vol 49, Berlin/Heidelberg: Springer; 2013:335-378.
43. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123-140.
44. Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: Lorenza S, ed. *Proceedings of the International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann; 1996:148-156.
45. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell*. 1998;20(8):832-844.
46. Yaman MA, Subasi A, Rattay F. Comparison of random subspace and voting ensemble machine learning methods for face recognition. *Symmetry*. 2018;10(11):651.
47. Cheplygina V, Tax DMJ, Loog M. Dissimilarity-based ensembles for multiple instance learning. *IEEE Trans Neural Netw Learn Syst*. 2016;27(6):1379-1391.
48. Yule GU. On the association of attributes in statistics: with illustrations from the material of the childhood society. *Philos Trans R Soc Lond*. 1990;194(A):257-319.
49. Seath A, Sokal RR. Numerical taxonomy: the principles and practice of numerical classification. *Syst Zool*. 1973;24(2):263-268.
50. Skalak DB. The sources of increased accuracy for two proposed boosting algorithms. *Proc Am Assoc Artif Intell*. 1996;120-125.
51. Giacinto G. Design of effective neural network ensembles for image classification purposes. *Image Vis Comput*. 2001;19:699-707.
52. Margineantu DD, Dietterich TG. *Pruning adaptive boosting*. San Francisco, CA: Morgan Kaufmann; 1997:211-218.
53. Kohavi Ron, Wolpert David H. Bias plus variance decomposition for zero-one loss functions. Paper presented at: Proceedings of the 13th International Conference on Machine Learning; 1996:275–283.
54. Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Mach Learn*. 2000;40:139-157.
55. Fleiss JL, Levin B, Paik MC. *Statistical methods for rates and proportions. Wiley Series in Probability and Statistics*. Hoboken, NJ: John Wiley & Sons; 2003.
56. Cunningham P, Carney J. Diversity versus quality in classification ensembles based on feature selection. Paper presented at: Proceedings of the 11th European Conference on. *Machine Learning*. 2000;1810:109-116.
57. Hansen LK, Salamon P. Neural network ensembles. *IEEE Trans Pattern Anal Mach*. 1990;12(10):993-1001.

58. Partridge D, Krzanowski W. Software diversity: practical statistics for its measurement and exploitation. *Inf Softw Technol*. 1997;39(10):707-717.
59. Sharkey AJC, Sharkey NE. Combining diverse neural nets. *Knowl Eng Rev*. 1997;12(3):231-247.
60. Ruta D, Bogdan G. Classifier selection for majority voting. *Inf Fusion*. 2005;6(1):63-81.
61. Dai Q, Ye R, Liu Z. Considering diversity and accuracy simultaneously for ensemble pruning. *Appl Soft Comput J*. 2017;58:75-91.
62. Partalas I, Tsoumakas G, Vlahavas I. An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. *Mach Learn*. 2010;81(3):257-282.
63. Tsymbal A, Puuronen S, Patterson DW. Ensemble feature selection with the simple Bayesian classification. *Inf Fusion*. 2003;4(2):87-100.
64. Tsoumakas G, Partalas I, Vlahavas I. An ensemble pruning primer. *Applications of Supervised and Unsupervised Ensemble Methods*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009:1-13.
65. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. *IEEE Trans Pattern Anal Mach Intell*. 2000;22(1):4-37.
66. Rokach L. Ensemble-based classifiers. *Artif Intell Rev*. 2010;33(1-2):1-39.
67. Zhou Z-H. *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL: CRC Press; 2012.
68. Woźniak M, Graña M, Corchado E. A survey of multiple classifier systems as hybrid systems. *Inf Fusion*. 2014;16(1):3-17.
69. Mohandes M, Deriche M, Aliyu SO. Classifiers combination techniques: a comprehensive review. *IEEE Access*. 2018;6:19626-19639.
70. Singh PK, Sarkar R, Nasipuri M. Correlation-based classifier combination in the field of pattern recognition. *Comput Intell*. 2018;34(3):839-874.
71. Mukhopadhyay A, Singh P, Sarkar R, Nasipuri M. Handwritten Indic Script Recognition Based on the Dempster–Shafer Theory of Evidence. *J Intell Syst*. 2018. https://doi.org/10.1515/jisys-2017-0431.
72. Krawczyk B, Woźniak M. Untrained weighted classifier combination with embedded ensemble pruning. *Neurocomputing*. 2016;196:14-22.
73. Hassan Mohammed Falih, Abdel-Qader Ikhlas. Analysis of multiple classifier system using product and modified product rules. Paper presented at: IEEE International Conference on Electro/Information Technology (EIT); 2015:152-157.
74. Kuncheva LI. *Combining Pattern Classifiers*. Hoboken, NJ: John Wiley & Sons; 2004.
75. Michal W. *Hybrid Classifiers: Methods of Data, Knowledge, and Classifier Combination Studies in Computational Intelligence*. Vol 519. Berlin/Heidelberg: Springer; 2014.
76. Moreno-Seco Francisco, Iñesta José M, León Pedro J Ponce, Micó Luisa. Comparison of classifier fusion methods for classification in pattern recognition tasks. In: Yeung Dit-Yan, Kwok James T, Fred ALN, Roli F, de Ridder D, eds. *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, SSPR 2006 and SPR 2006, Hong Kong, China, August 17–19, 2006. Proceedings*; Springer: Berlin, Germany; 2006:705-713.
77. Rokach L. *Ensemble Methods for Classifiers*. New York, NY: Springer-Verlag; 2010.
78. Buntine Wray Lindsay. A Theory of Learning Classification Rules. [PhD thesis]. Sydney, Australia: University of Technology, 1992.
79. Li Wenxing, Hou Jian, Yin Lizhi. A classifier fusion method based on classifier accuracy. Paper presented at: 2014 International Conference on Mechatronics and Control (ICMC) 2014;32:2119-2122.
80. He H, Cao Y. SSC : a classifier combination method based on signal strength. *IEEE Trans Neural Netw Learning Syst*. 2012;23(7):1100-1117.
81. De Stefano C, Fontanella F, di Freca AS. A Novel Naive Bayes Voting Strategy for Combining Classifiers. Paper presented at: 2012 International Conference on Frontiers in Handwriting Recognition; Bari, Italy; 2012:467-472.
82. Nguyen TT, Dang MT, Liew AWC, Bezdek JC. A weighted multiple classifier framework based on random projection. *Inform Sci*. 2019;490:36-58.
83. Mukhopadhyay A, Singh PK, Sarkar R, Nasipuri M. A study of different classifier combination approaches for handwritten indic script recognition. *J Imag*. 2018;4:39.
84. Liu Z, Pan Q, Dezert J, Martin A. Combination of classifiers with optimal weight based on evidential reasoning. *IEEE Trans Fuzzy Syst*. 2018;26(3):1217-1230.

85. Skurichina M, Duin Robert PW. Bagging for linear classifiers. *Pattern Recogn*. 1998;31(7):909-930.

86. Dua D, Graff C. *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science; 2017.

87. Alcalá-Fdez J, Fernández A, Luengo J, et al. KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *J Multi Valued Logic Soft Comput*. 2011;17(2-3):255-287.

88. Kuncheva Ludmila. *Ludmila Kuncheva Collection LKC*. 2004. http://pages.bangor.ac.uk/~mas00a/activities/real_data.htm.

89. King RD, Feng C, Sutherland AS. Comparison of classification algorithms on large real-world problems. *Appl Artif Intell*. 1995;9(3):289-333.

90. Lei B, Xu G, Feng M, et al. *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*. John Wiley & Sons; 2017:480.

91. Martínez-Muñoz G, Suárez A. Out-of-bag estimation of the optimal sample size in bagging. *Pattern Recognit*. 2010;43(1):143-152.

92. Ko AHR, Sabourin R, de Souza Britto A, Oliveira L. Pairwise fusion matrix for combining classifiers. *Pattern Recognit*. 2007;40(8):2198-2210.

93. Cruz RMO, Sabourin R, Cavalcanti GDC. META-DES.Oracle: meta-learning and feature selection for dynamic ensemble selection. *Inf Fusion*. 2017;38:84-103.

94. Juang BH, Katagiri S. Discriminative learning for minimum error classification. *IEEE Trans Signal Process*. 1992;40(12):3043-3054.

95. Frank E, Hall MA, Witten IH. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques."* 4th ed. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 2016.

96. Amancio DR, Comin CH, Casanova D, et al. A systematic comparison of supervised classifiers. *PLoS One*. 2014;9(4):e94137.

97. Dutot Antoine, Guinand Frédéric, Olivier Damien, Pigné Yoann. GraphStream: a tool for bridging the gap between complex systems and dynamic graphs. Emergent Properties in Natural and Artificial Complex Systems. Satellite Conference within the 4th European Conference on Complex Systems (ECCS'2007);2007.

98. Tsymbal Alexey, Pechenizkiy Mykola, Cunningham Pádraig. Diversity in Ensemble Feature Selection. Department of Computer Science, Trinity College Dublin; 2003:1–38.

99. Demšar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res*. 2006;7:1-30.

100. Cruz RMO, Sabourin R, Cavalcanti GDC. Analyzing different prototype selection techniques for dynamic classifier and ensemble selection. Paper presented at: 2017 International Joint Conference on Neural Networks (IJCNN); 2017; Anchorage, AK.

101. Kapp MN, Sabourin R, Maupin P. An empirical study on diversity measures and margin theory for ensembles of classifiers. Paper presented at: 2007 10th International Conference on Information Fusion (IEEE); 2017; Québec City, Canada.

102. Cavalcanti GDC, Oliveira LS, Moura TJM, Carvalho GV. Combining diversity measures for ensemble pruning. *Pattern Recognit Lett*. 2016;74:38-45.

103. Kuncheva LI, Whitaker CJ. Ten measures of diversity in classifier ensembles: limits for two classifiers. Paper presented at: In A DERA/IEE Workshop on Intelligent Sensor Processing (Ref. No. 2001/050), 2001; Birmingham, UK.

104. Johansson U, Löfström T, Boström H. Paper presented at: 2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL); 2013; Singapore.

105. Page Lawrence, Brin Sergey, Motwani Rajeev, Winograd Terry. The PageRank Citation Ranking: Bringing Order to the Web; 1999.

106. Katz L. A new status index derived from sociometric analysis. *Psychometrika*. 1953;18(1):39-43.

107. Bonacich P. Power and centrality: a family of measures. *Am J Sociol*. 1987;92(5):1170-1182.

How to cite this article: Silva RA, Britto Jr AS, Enembreck F, Sabourin R, Oliveira LES. CSBF: A static ensemble fusion method based on the centrality score of complex networks. *Computational Intelligence*. 2020;36:522–556. https://doi.org/10.1111/coin.12249