International Journal on Artificial Intelligence Tools
Vol. 29, Nos. 3 & 4 (2020) 2060004 (25 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0218213020600040

Selecting and Combining Classifiers Based on Centrality Measures

Ronan Assumpção Silva
Pontifical Catholic University of Parana (PUCPR), Curitiba, Parana, Brazil
Federal Institute of Parana (IFPR), Pinhais, Parana, Brazil
ronan.silva@ifpr.edu.br

Alceu S. Britto, Jr.
Pontifical Catholic University of Parana (PUCPR), Curitiba, Parana, Brazil
State University of Ponta Grossa (UEPG), Ponta Grossa, Parana, Brazil
alceu@ppgia.pucpr.br

Fabricio Enembreck
Pontifical Catholic University of Parana (PUCPR), Curitiba, Parana, Brazil
fabricio@ppgia.pucpr.br

Robert Sabourin
École de Technologie Supérieure (ÉTS), Montreal, Quebec, Canada
robert.sabourin@etsmtl.ca

Luiz S. Oliveira
Federal University of Parana (UFPR), Curitiba, Parana, Brazil
luiz.oliveira@ufpr.br

Received 28 April 2019
Accepted 11 April 2020
Published 18 June 2020

Centrality measures have been helping to explain the behavior of objects, given their relation, in a wide variety of problems, since sociology to chemistry. This work considers these measures to assess the importance of every classifier belonging to an ensemble of classifiers, aiming to improve a Multiple Classifier System (MCS). Assessing the classifier's importance by employing centrality measures, inspired two different approaches: one for selecting classifiers and another for fusion. The selection approach, called Centrality Based Selection (CBS), adopts a trade-off between the classifier's accuracy and their diversity. The sub-optimal selected subset presents good results against selection methods from the literature, being superior in 67.22% of the cases. The second approach, the integration, is named Centrality Based Fusion (CBF). This approach is a weighted combination method, which is superior to literature in 70% of the cases.

Keywords: Multiple classifier systems; fusion; selection; diversity; centrality measures.

1. Introduction

Classification is one of the main tasks in the field of pattern recognition. Every classification problem may present a different number of instances, attributes, and classes, characterizing different levels of difficulty. Sometimes the problem difficult makes unfeasible the use of monolithic classifiers due to the wide range of variability involved. However, the Multiple Classifier System (MCS) is considered attractive in such a scenario. With an MCS, we can avoid the risk of defining a single classifier to work with data showing wide pattern variability. Therefore, the success of an MCS relies upon an ensemble composed of accurate and diverse members, so the problem space is covered by sharing the responsibility among them. Each member of the ensemble of classifiers must be accurate in the sense that its decision is better than random guessing. On the other hand, they must show diversity in their decisions, which concerns making different errors. Thus, their decisions combined may improve classification performance.

An MCS is composed of three distinct phases: Generation, Selection, and Combination. Concerning the generation phase, they can be categorized as heterogeneous when different base classifiers are used to achieve diversity, while in the homogeneous, the same base classifier is used while diversity is obtained by data manipulation. Bagging,[1] boosting[2] and random subspaces[3] are classical approaches for the generation of homogeneous pools. In the selection phase, there are static and dynamic strategies. In the former, one or more classifiers can be part of a subset, defined statically during the ensemble training phase. In the later, the classifiers are dynamically selected in the test phase accordingly to the test pattern characteristics.[4] In the MCS combination phase, the decisions of all classifiers inside the whole ensemble or a subset of them are merged, producing a final decision. Several fusion methods are present in the literature.[5,6]

A challenge in the MCS context is related to how combining the classifiers generated. The whole pool can be combined or just a subset of its classifiers. Testing all possible classifier combinations is often unfeasible, but some heuristic can be used to approximate an optimal subset. The performance of the subset may outperform the performance of the entire pool if the approach used to select the classifiers considers two requirements: (i) accuracy and (ii) diversity. In such a context, our challenge is how to select and combine diverse classifiers, considering not only their accuracy as a team but also their competence working together, covering more properly the problem feature space.

The scheme used to measure the classifier competence in a selection or combination method can vary, but the primary goal is the same, improve the accuracy of the ensemble, usually based on its members' ability and limitations. The combination methods in the literature can be divided into three categories: Non-trainable, trainable, and dynamic weighting.[7] The non-trainable strategy does not require any extra training to determine the influence of the classifiers on voting. A trainable strategy, on the other hand, goes in the opposite direction. This strategy uses the output of every classifier as an input feature for another learning step. In this step, its influence is adapt according to the flaws observed in the training data, using a new learning algorithm for this. Dynamic weighting, similarly to dynamic selection, consider that a test pattern can fall into a region that matches (or not) the region of competence of a classifier.

This work presents two distinct MCS approaches, both inspired by centrality measures of complex network. We used centrality measures to estimate the importance of the classifiers that compose an ensemble, i.e., their influence taking into account their relationship inside the ensemble. In our previous work, a weighted fusion algorithm, named CBF (Centrality Based Fusion),[8] is used to combine the decision of each classifier member considering its importance estimated on centrality measures computed in a network created with the ensemble members. Here, we extend the CBF algorithm by considering the strategy used to compute the pairwise relationship between the classifiers using not only asymmetric but also symmetric diversity measures. Besides, we propose a new algorithm for static selection of classifiers, named Centrality Based Selection (CBS) that uses the importance, or influence, of each classifier inside the ensemble to determine whether an ensemble member will be selected, or not, to compose a subset of most promising classifiers for a given problem.

This paper presents five sections. Section 2 introduces some fundamental concepts and definitions concerning the proposed approaches for classifier fusion and static selection of classifiers. Section 3 describes the two approaches, while Section 4 shows our experiments, results, and corresponding discussions. Finally, Section 5 presents our conclusion and future work perspectives.

2. Definitions

This section presents some important concepts of complex network theory, mainly focused on centrality measures. Those measures help to score classifiers by their importance, regarding the relationship each classifier has with all others represented in the network. This background is the basis of the selection and the combination methods presented in this paper.

2.1. *MCS background*

This section presents the knowledge required to understand how the MCS concepts are related to our approach, based on complex network techniques. In an ensemble of classifiers $C = \{c_1, c_2, \ldots, c_T\}$, every member represents an independent function $c_t : R^n \to W$ that assigns a class label $w_i \in W$ to $x \in R^n$, where $W = \{w_1, w_2, \ldots, w_M\}$. Usually, a training dataset S_{train} is an input to produce an ensemble through classical approaches, such as bagging, boosting, or random subspaces. Then, the whole ensemble C, or a subset of classifiers selected from C, can be used in the classification process.

Table 1. Pairwise relation between two classifiers c_i and c_j.

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

As mentioned before, there are two distinct selection approaches to provide the most promising subset $C' \in C$, named static and dynamic.[4] The static selection defines the subset C' during the training phase of an MCS, using during the testing phase the same selected subset for all the unseen patterns available in S_{test}. Commonly, during the selection process, it uses a validation set S_{val} to evaluate each candidate subset, avoiding the same data used for training (overtuning). The dynamic selection approach, on the other hand, defines a subset C' for each test pattern x_{test} during the testing phase. In this case, the validation set S_{val} is also used in the testing phase to estimate the competence of each classifier for a given unknown pattern (x_{test}). The idea is to select the most promising classifier(s) for each test pattern x_{test}. In this paper, we focus on the weighted fusion of all classifiers in the ensemble and the use of a static selection approach.

Both methods described here have in the initial steps the estimation of the pairwise diversity between classifiers. The well-known diversity measures such as double fault, Q statistics, correlation coefficient, kappa pruning, and disagreement are based on four basic relations. These relations involve correct and incorrect answers of a pair of classifiers, c_i and c_j, regarding to every pattern $x_{val} \in S_{val}$. Table 1 presents these relations.

A detailed description of each diversity measure used in our work is available in Ref. 6, where one may also find an interesting investigation related to the impact of diversity on the ensemble accuracy. The most used diversity measures are asymmetric. However, here we will also investigate the pairwise relations N_{10} and N_{01}, which are symmetric. Next, we present the network representation for the pool, which is capable of dealing with symmetric and asymmetric diversity measures.

2.2. *Ensemble network*

Recently, the authors in Ref. 9 proposed a network representation for the ensemble of classifiers. Every ensemble member c_i has a pairwise relationship with another member c_j. The relationship can represent the difference between them, e.g., a pairwise diversity measure. Figure 1 presents the graph representation of an ensemble network. Figure 1(a) represents an ensemble with symmetric pairwise relations, while Figure 1(b) represents the same ensemble with asymmetric relations. Considering Table 1, N_{11} and N_{00} can be used as the asymmetric relationship, as well as the traditional diversity measures such as Q statistics, correlation coefficient,

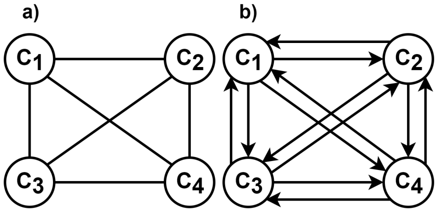

Fig. 1. The ensemble network. The first type (a) is an ensemble with symmetric pairwise relations while the second type (b) have asymmetric relations.

disagreement, double fault, and kappa pruning. However, the relations N_{10} and N_{01} demand a special asymmetric representation.

The network structure allows the estimation of the importance of the ensemble members, through the centrality measures.[8,10] In the case of the proposed static selection method, the importance of the classifier can be used as criteria to select a subset C' that approximates the best possible subset without the need for an exhaustive search. The equations required to score the ensemble members in a network structure must also respect the symmetric and the asymmetric relations. In the next sections, we present some popular centrality measures.

2.2.1. *Centrality measures*

The centrality measures use topological information about members inside a particular network to suggest a score, according to the member influence or importance. The following measures are classic: Degree,[11] betweenness,[12] closeness,[11] and eigenvector.[13] They focus on different network information, such as (a) the number of edges (degree centrality); (b) geodesics (betweenness centrality, closeness centrality), and (c) walks, in which vertices and edges can be revisited (eigenvector centrality).

The degree centrality K_{c_i} is related to the number of edges of a given network member. It is defined by Eq. (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 number of members (classifiers) of the network. The weight associated with $E_{c_i c_j}$ can be the original weight of the edge, so it is used to estimate the weighted degree. However, to estimate the unweighted degree, the weight is considered to be simply 1.0, which only indicates the presence of an edge, i.e., the score of all edges is the same when it exists. For an ensemble network with diversity representing the classifiers relations, a classifier with a high degree means it is very divergent from its neighbors.

Betweenness[12] is another classic centrality measure. It considers the number of shortest paths (geodesic paths) from each member of the network to all others that pass through a particular member. That member, frequently present on that type of path is considered important. The betweenness centrality measure can be described as the Eq. (2).

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

where $g^{c_i}_{c_j c_k}$ is the number of geodesics between c_j and c_k that pass through c_i. The total of the geodesics between c_j and c_k is $g_{c_j c_k}$. The network may have only one component to calculate this centrality measure with this equation in order to calculate all the distances (paths) between two vertices. Each geodesic represents a sub-ensemble of classifiers with high diversity. Consequently, a classifier with high betweenness centrality is a frequent member of the most diverse sub-ensembles.

Closeness is another classic measure that uses the shortest paths for its estimation.[11,14] This measure considers the average distance (length of the average shortest paths) of a particular member to all others in the network. Members with high closeness centrality are those members closest to all others. Like betweenness, Eq. (3) for closeness centrality also depends on a connected network. A classifier with high closeness centrality indicates a distinguished contribution to the team diversity since it was obtained by the average of all diversity relationships of a particular classifier to all others.

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

where l_i is the average shortest path length of each member to other members. The geodesics of the member c_i to all other members c_j are estimated, and the smaller average of the shortest path length is, higher the centrality.

Eigenvector centrality[13] is the last classic measure considered in this work. A member is considered to be central if it has a relationship with others that are themselves central. Equation (4) presents the Bonacich eigenvector.

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

where A is the adjacent matrix, λ is a constant (the eigenvalue), and x is the eigenvector. The score of centrality is proportional to the sum of the centralities of its adjacent members.

These measures can be used in asymmetric and in symmetric relations. For example, there are two measures derived from degree centrality to deal with asymmetric relations: Indegree and outdegree. The first computes the ties directed to a vertice, while the latter computes the ties a vertice directs to others. Therefore, those variations could lead to different observations that possibly explain complementary collective behaviors.

In this section, we have presented traditional centrality measures, considered in the literature.[14] Each centrality measure scores the influence of every network member but uses different characteristics of the network. Besides, choosing a proper centrality measure depends on what a network represents, and which questions the network analysis intends to answer.

3. Proposed Methods

This section presents the two proposed methods, one for fusion and another for static selection of classifiers, both using centrality measures to estimate the importance of the classifiers in the ensemble. These methods are independent, but they share theoretical aspects and the initial steps. The Centrality Based Fusion (CBF) considers the fusion of all ensemble members, while Centrality Based Selection (CBS) selects a subset of classifiers from the original ensemble using a static approach.

3.1. *Centrality Based Fusion (CBF)*

The CBF fusion approach is described as an ensemble composed of three distinct phases, as depicted in Figure 2.

3.1.1. *Phase a: Pool creation, accuracy and diversity estimation*

The first phase of the method (Figure 2(a)) has an initial pool of classifiers C of size T, created using a pool generation method applied in training set S_{train}. Any method available in the literature can be used to create the ensemble, such as the traditional bagging,[1] boosting,[2] or random subspaces.[3] Then, a validation set S_{val} is used to estimate the accuracy of each classifier created, as well as the pairwise diversity. To estimate the pairwise diversity, any of the diversity measures presented in Section 2.2.1 can be useful, but the approach is not limited to them. The score obtained by the centrality measure must be normalized to fit the established range $[0.1, 1.0]$, avoiding the value 0, which can lead to misinterpretation such an edge

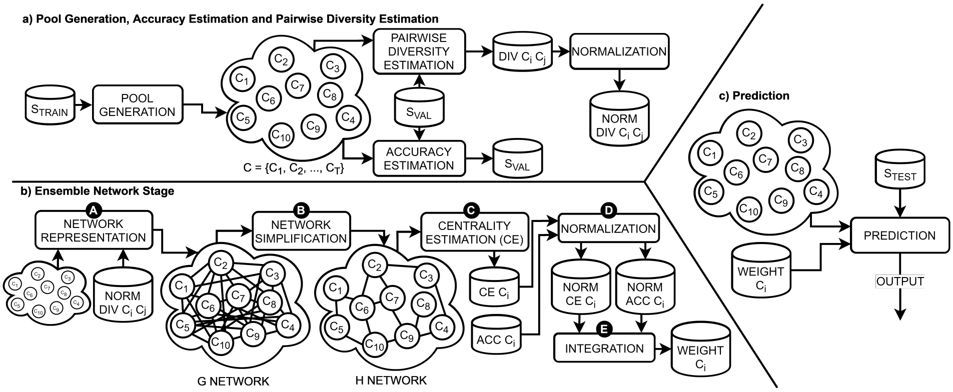

Fig. 2. A general overview of the CBF method.

Table 2. Diversity is expressed as an increasing value (↑) or decreasing value (↓) depending on the centrality measure.

	DF	CC	Dis	QS
Degree	↑	↑	↓	↑
Betweenness	↓	↓	↑	↓
Closeness	↓	↓	↑	↓
Eigenvector	↑	↑	↓	↑
Local centrality	↑	↑	↓	↑

with 0 weight could be interpreted either as an absent edge or as an edge presenting a low score. Adopting another range at this point is possible, but it should respect the meaning of the centrality score, such that the centrality score of a member being high according to a high score or the opposite. Table 2 clarifies this point, presenting values to better represent the weight of the edges depending on the centrality measure to use on the analysis of the classifier's importance.

3.1.2. *Phase b*: *Construction of the ensemble network*

A network represents the ensemble members along with the computed diversity of each pair of classifiers (process A in Figure 2(b)). Both estimated in the previous phase. The vertices in the network are representing classifiers, while the edges between them are representing the score of the particular pairwise diversity measure. The ensemble network is a complete graph before the simplification process (process B in Figure 2(b)) using an edge pruning method, such as those mentioned in Refs. 15 and 16. The reason behind this step is highlighting the most important relations in the ensemble; in this case, the pruning directive preserves the most important relations. To prune, we propose a modified Naïve Pruning Approach, present in Algorithm 1. So, the proposed edge-based pruning technique keeps all vertices, while removing edges that represent a low diversity between pairs of classifiers without increasing the number of components of the network.

Algorithm 1 differs from the original Naïve Pruning Algorithm[15] in two main aspects. First, as mentioned earlier, it always stops pruning upon encountering the first edge whose removal would increase the number of components, while the original algorithm ignores the bridge edge, i.e., an edge whose removal may increase the number of components of the network, and continues pruning. Second, it does not require the γ parameter to estimate the number of edges to be removed, due to some knowledge about the network is required. The graph as a single component allows the estimation of the presented classic centrality measures. So the adopted stop criterion pursues a network presenting only one component while reduces the number of edges by preserving the most diverse relations. As a consequence of the pruning process, the network H is used as the input for centrality estimation (process C in Figure 2(b)).

Algorithm 1 Modified Naïve algorithm

```
Input: A weighted graph $G = (V, E)$ such that $E = \{e_1, e_2, ..., e_N\}$ and $V = \{c_1, c_2, ..., c_T\}$
Output: A subgraph $H \subset G$ such that $H = (V, F)$ and $F \subset E$
$F \leftarrow E$
$SortEdges(F)$
$i \leftarrow 1$
while $i <= N$ do
    if $C(c_r, c_s; F \backslash e_i) \neq -\infty$ then
        $F \leftarrow F \backslash e_i$
    else
        return $H = (V, F)$;
    end if
    $i \leftarrow i + 1$
end while
```

The centrality estimation (process C in Figure 2(b)) is responsible for highlighting the prominent vertices (classifiers) of the ensemble. A classifier showing a higher centrality score means an important influent position in the network; therefore, it plays an essential role concerning the diversity of the given group of classifiers. It is the meaning of a centrality measure considering an ensemble network in which the edges represent the diversity relation provided by the score of a computed pairwise diversity measure. Each centrality measure provides different scores accordingly to its philosophy of what is an influential position. Therefore, a centrality measure may suggest a different rank of importance for the classifiers compared to others. For instance, the classifier that has the most neighbors is the most important regarding degree centrality, while the classifier that lies in most of the geodesic paths is the most important for betweenness centrality and so on. Different centrality measures are evaluated on Section 4.1.1 to observe the relation between them to the ensemble accuracy.

After the centrality measure is computed, every single classifier will receive a score (CE_i), as well as the previously computed accuracy Acc_i. These scores are normalized using the min-max normalization process, adopting the range of $[0.1, 1.0]$ (process D in Figure 2(b)). Therefore, after this process, both measures are considered equally important for the estimation of the influence of each ensemble member. It does make a difference considering the scores suggested by the centrality measures are on different scales. Also, normalizing the accuracy may prevent this score of being less or more important than its diversity role. Then, these normalized scores are merged to obtain the final weight for each ensemble member i (process E in Figure 2(b)) as denoted in Eq. (5):

$$\psi_i = CE_i \times Acc_i \tag{5}$$

where the resulting weight ψ_i is an input of the final phase of the presented approach (Prediction in Figure 2(c)). Applying the normalization processes and merging them into a new scoring measure, this proposal suggests that both attributes, individual accuracy and diversity role being equally essential and complementary. It is supported by two different aspects of the classifier, so, considering that the first measures how distinct the classifier is according to its relationship to another ensemble members, on the other hand, accuracy selfishly considers only its individual performance. Even two classifiers in the ensemble performing equally well in individual accuracy, but eventually, one classifier can contribute most for the ensemble's generalization than the other by recognizing different problems. Another doubtful solution is choosing the Centrality of classifier assessed by diversity relations as the unique parameter. Basing the classifier influence only on disagreement decisions may lead the ensemble to poor results due to the lack of accuracy within its members. An attractive solution in this direction is moving forward and mapping when these disagreements are healthy for the ensemble generalization. These few situations already justify the importance of combining both parameters to measure the influence of the classifier in the ensemble's final prediction.

3.1.3. *Phase c: Prediction*

As shown in Figure 2(c), each classifier in the pool has its own weight ψ_i, which reflects its importance for the ensemble. Each classifier manifests its decision by choosing a particular class among the possible ones, and its influence is computed only for that class. The most preferred class accordingly to each classifier influence is the final ensemble decision. It is the closure of the final phase of our method.

3.2. *Centrality Based Selection (CBS)*

This section presents the Centrality Based Selection (CBS) method, which performs the static selection of classifiers from an initial pool. Basically, it shares the main modules of Phases "a" and "b" of the CBF method, in which we have the ensemble generation, the network construction, and estimation of the classifiers' importance based on centrality measures. In the CBS method, we have an additional module in which a subset of classifiers C' is obtained from C, the original ensemble, still in the training phase. For this purpose, the created network structure (graph G) allows the evaluation of the classifiers' importance by using a centrality measure χ. The score Ψ of each classifier, assigned by the centrality measure χ is used as one important criterion in the classifier subset selection. Only the p most important classifiers have the accuracy estimated, so the best performing classifier between the most important is chosen to be part of the subset C'. The subset C' is used to classify every unseen pattern $x_{test} \in S_{test}$ in the Phase 'c' using majority vote as fusion rule.

3.2.1. *The CBS algorithm*

A detailed description of the CBS method is presented in Algorithm 2. The goal is to find a sub-optimal subset $C' \in C$ selected statically. It requires the following inputs: the ensemble of classifiers C, a pairwise measure ϖ, a centrality measure χ, a validation set S_{val}, and p as the proportion of important classifiers to be evaluated.

All the inputs are used to build and analyze the ensemble network, except p. The ensemble C is used to estimate the relationship between pairs of classifiers ϖ, defined by some pairwise diversity measure. A graph called ensemble network represents these inputs, so $G(C, \varpi)$. Any pairwise relation presented in Table 1 or the diversity measures presented in Section 2.1 can be used to estimate ϖ. The number of symmetric pairwise relations is defined by the combination $C_{T,2}$, where T is the size of the ensemble. For the asymmetric relationship, the number of relations is defined by a permutation $P_{T,2}$. Therefore, for an ensemble of $T = 100$ classifiers, there are 4950 symmetric relations, while for asymmetric, the total is 9900 relations. To score the classifiers by their importance, a centrality measure χ is chosen to analyze the ensemble network. Any chosen χ must take into account the type of relations observed in the network, i.e., asymmetric or symmetric. The parameter τ and the inversion function φ are used to adapt the network for the proper analysis, considering the used pairwise relation and the centrality measure.

The pool C is the first input, represented in the network by the set of vertices V (line *1*). A computed pairwise diversity ϖ is represented in the network by the set of edges E (line *2*). In the lines *3–6*, the inversion function is applied in the score of diversity to properly analyze the importance of the network members by the chosen centrality measure χ. Next, a normalization process (lines *7–9*) is adopted to avoid possible null or negative values present in the set E. It is important to consider φ and τ (lines *7–10*) to estimate the centrality measures correctly. These parameters are detailed later in this section.

The next step is to use the set of vertices and the set of edges to build the ensemble network, represented by the graph $G(V, E)$ (line *11*). The score of importance Ψ_i related to every ensemble member is obtained by a given centrality measure χ over the graph G (line *12*). Every classifier from the pool C is added do the set CR, the remaining pool (line *13*), in which the iterations will compose the main subset C', initially empty (line *14*). The CR set is ordered regarding Ψ, and the classifier with the highest score is c_M (lines *15–16*). Each classifier added to C' in the algorithm is removed from the remaining pool CR, as seen in the lines *17–18*.

Afterwards, the most complementary classifier regarding c_M is the one that receives an incoming edge from c_C with the highest weight (line *19*). This classifier is added to C' (line *20*) and removed from set CR.

The lines *22–34* presents the process to compose the subset C'. It is a repetition that allows adding to C' only the most accurate classifier between the p most

Algorithm 2 CBS(C,ϖ,χ,τ, S_{val},φ, p)

```
Input: Ensemble of classifiers $C = \{c_1, c_2, \ldots, c_T\}$
Input: Pairwise Relation ($\varpi$), where $\varpi(c_i, c_j)$
Input: Centrality Measure ($\chi$)
Input: Validation dataset ($S_{val}$)
Input: Normalization ($\tau$)
Input: Inversion Function ($\varphi$)
Input: Proportion of Candidates ($p$)
Output: $C'$
$V \leftarrow C$
$E \leftarrow \varpi$
if $\varphi = TRUE$ then
    for all $E_i \in E$ do
        $E_i = 1/E_i$
    end for
    if $\tau = TRUE$ then
        $E = normalize(E)$
    end if
end if
Build the network $G(V, E)$
$\Psi \Leftarrow$ Estimate Centrality $(\chi, G)$
$CR \leftarrow C$
$C' \leftarrow \emptyset$
Order $CR$ according to $\Psi$ score;
Compute $c_M$ (the classifier with the highest $\Psi$ in $CR$)
$C' \cup c_M$
$CR \leftarrow CR \setminus c_M$
Compute $c_C \leftarrow$ the classifier that most complement (or diverge) from $c_M$
$C' \leftarrow c_C \cup C'$
$CR \leftarrow CR \setminus c_C$
repeat
    Estimate $A_{C'}$ the accuracy of the ensemble $C'$
    $CP \leftarrow$ the $p$ most important/central classifiers in $CR$
    for all $c_i \in CP$ do
        $A \leftarrow$ estimated accuracy $A_i$ from $c_i \cup C'$
    end for
    $AM \leftarrow$ Highest value of accuracy found in set $A$
    $c_M \leftarrow$ The most accurate classifier
    if $AM \geq A_{C'}$ then
        $C' \leftarrow c_M \cup C'$
        $CR \leftarrow CR \setminus c_M$
    end if
until $AM < A_{C'}$
```

important classifiers in the ensemble. The accuracy of the current ensemble C' is estimated (line *23*) to compare with $c_i \in C'$, i.e. a new configuration of C' considering an additional classifier c_i. To obtain a set of the most complementary classifiers, CP receives only $p \times |CR|$ classifiers with the highest score Ψ (line *24*).

The accuracy A_i estimated in line *26* is related to $c_i \in CP$. Using A, the most accurate union is selected. It is assigned to CM (line *29*) and its accuracy value is stored in AM (line *28*). If the accuracy AM of the candidate union $c_i \in CP$ is at least equal to the accuracy $A_{C'}$ of the previously set C' (line *30*), then the candidate classifier c_i is really added to the set C' (line *31*). Otherwise, the process stop (line *34*).

4. Experiments

Our experimental protocol included 30 classification problems, presenting only numeric features, no missing values, and a varied number of instances, attributes, classes, and imbalance ratio (the proportion between the instances in the majority class and the instances in the minority class). All these datasets are presented in Table 3.

The statistical method to evaluate the performance of the approach in these experiments is the cross-validation with $k = 6$. The 6-fold cross-validation was divided as follows: three for training S_{train} (= 50% of the original database), two for validation S_{val} ($\cong 32,3\%$) and one ($\cong 16,7\%$) for testing S_{test}. Stratified sampling was adopted to guarantee a class distribution balance in all subsets from the original dataset S (S_{train}, S_{val} and S_{test}). A pool of $T = 100$ classifiers was created with bagging, and the perceptron with minimum squared error is the base classifier. Every bag used to train a classifier presents only 66% of the training samples. An unstable base classifier, along with small bags, may enforce a pool of weak and diverse classifiers. Each classifier presented accuracy higher than 50% (estimated on S_{val}).

4.1. *CBF evaluation*

CBF experiments begin with the evaluation of different setups, which requires to consider the possible combination of different pairwise diversity and centrality measures. The best setup is then compared to the literature. A final discussion ends the section, which inspires some decisions concerning the CBS approach.

4.1.1. *Evaluation of pairwise diversity and centrality measures*

A combination of 4 pairwise diversity measures (double fault-DF, correlation coefficient-CC, disagreement-Dis, and Q statistics-QS) and seven different centrality measures (betweenness unweighted (NW), betweenness weighted (W), closeness, degree unweighted (NW), degree weighted (W), eigenvector e local centrality) were assessed to estimate the effect of parameters for the proposed CBF.

Table 3. Main characteristics of the classification problems: Number of instances (# I), number of attributes (# A), number of classes (# C), and Imbalance Ratio (I.R.).

Base	# I	# A	# C	I.R.	Repository
Australian	690	14	2	1.25	UCI[17]
Banana	2000	2	2	1.00	PRTools[18]
Blood	748	4	2	3.20	UCI[17]
CTG	2126	21	3	9.40	UCI[17]
Diabetes	766	8	2	1.86	UCI[17]
Ecoli	336	7	8	71.50	UCI[17]
Faults	1941	27	7	12.24	UCI[17]
German	1000	24	2	2.33	STATLOG[19]
Glass	214	9	6	8.44	UCI[17]
Haberman	306	3	2	2.78	UCI[17]
Heart	270	13	2	1.25	STATLOG[19]
ILPD	583	10	6	2.49	UCI[17]
Ionosphere	351	34	2	1.79	UCI[17]
Laryngeal1	213	16	2	1.63	LKC[20]
Laryngeal3	353	16	3	4.11	LKC[20]
Lithuanian	2000	2	2	1.00	PRTools[18]
Liver	345	6	2	1.38	UCI[17]
Magic	19 020	10	2	1.84	KEEL[21]
Mammo	830	5	2	1.06	KEEL[21]
Monk	432	6	2	1.12	KEEL[21]
Phoneme	5404	5	2	2.41	KEEL[21]
Segmentation	2310	19	7	1.00	UCI[17]
Sonar	208	60	2	1.14	UCI[17]
Thyroid	692	16	2	12.06	LKC[20]
Vehicle	847	18	4	1.10	STATLOG[19]
Vertebral	300	6	2	2.13	UCI[17]
WBC	569	30	2	1.68	UCI[17]
WDVG	5000	21	3	1.03	UCI[17]
Weaning	302	17	2	1.00	LKC[20]
Wine	178	13	3	1.48	UCI[17]

After performing 28 experiments (4 diversity measures × 7 centrality measures) and considering all the 30 classification problems, we computed these statistics: Friedman and the Nemenyi post hoc test. Figure 3 shows the average rank, presented as the score next the names used on parameters for CBF method, suggesting that the best setup is the Weighted Degree Centrality (Degree W) estimated over Double Fault (DF) as the estimated pairwise diversity of the ensemble. Interestingly, degree centrality, weighted or unweighted, is much easier to compute compared to other centrality measures, and also requires less computational effort. The DF measure is the choice of the five best configurations, but the difference between them is not statistically significant. A comparison between the five best scores also confirms Degree W (DF) as the best approach; however, the critical distance still suggests these approaches being similar.

In these experiments, QS and the CC diversity measures can present an undefined value due to a division by zero. It occurs when some relations between

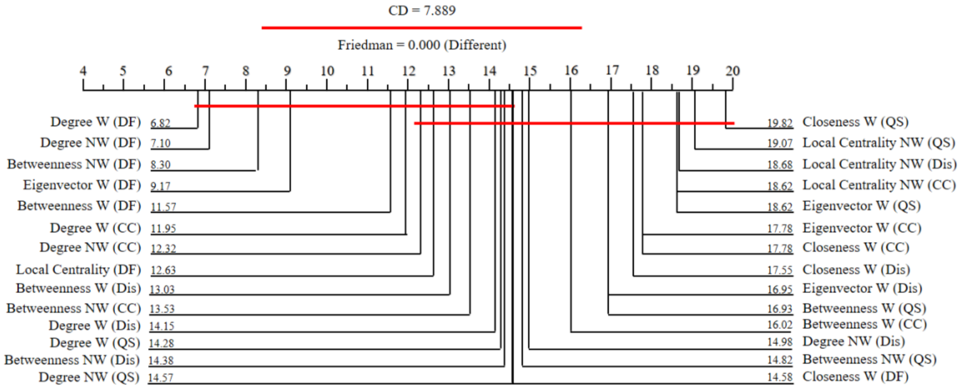

Fig. 3. Nemenyi and Friedman test to assess pairwise diversity and centrality on CBF. The centrality estimation may consider the weight of the edges (W) or not (NW).

incorrect/correct samples described in Table 1 are not observed. Therefore, a new value of 1.0 is assumed when this problem is detected.

4.1.2. *Comparison with state-of-art methods*

The most suitable parameters for CBF were observed using weighted degree centrality computed in double fault pairwise relations.The mentioned setup was the same for comparing the new method to these nine fusion methods of the literature: (a) the Majority Vote (MV); (b) the Weighted Majority Vote by Accuracy (WMV); (c) the Performance Weighting (PW);[22] (d) the Kuncheva Weighted Majority Vote (KWMV);[6] (e) the Bayesian Combination (BC);[23] (f) the Max Rule (MAR);[5] (g) the Median Rule (MER);[5] (h) the Sum Rule (SR);[5] (i) the Product Rule (PR).[5] Only MV disregards the classifier influence, assuming that influence is the same for every classifier. The others may fall into two groups: (i) static weighted score of the classifier based on individual accuracy or (ii) the score of the classifier is based on posterior probability. BC is the only exception because it estimates classifier influence using the accuracy of the classifier and combines with its posterior probability. CBF is compared to competitors, aiming to inform how the new estimation of the classifier weights can reveal issues concerned with performance and limitations.

Average accuracy, along with the corresponding standard deviation of the 10 methods (CBF and the 9 fusion methods in the literature) are presented in Table 4. As shown, the CBF method presents the best possible result in 14 of 30 classification problems, while the best literature competitor achieves the best result in just 6 classification problems.

A comparison of the number of wins, ties, and losses is presented in Figure 4. The new method shows consistent better results compared to every literature method, except PW. The dashed line illustrates the critical value, in this case, $cv = 19.5$.

Table 4. Average accuracy and standard deviation of each evaluated approach. The best results are in bold, and WS stands for Wilcoxon signed test. The values regards the p-value while + is for a significant result.

DS	MV	WMV	MAR	MER	SR	PR	PW	KWMV	BC	CBF:WD-DF
1	**87.83** ± 3.33	**87.83** ± 3.33	86.96 ± 3.05	87.68 ± 3.45	87.68 ± 3.45	87.68 ± 3.63	87.54 ± 3.64	87.68 ± 3.45	87.54 ± 3.45	87.25 ± 3.20
2	85.10 ± 2.35	85.10 ± 2.36	84.85 ± 2.32	**85.15** ± 2.28	**85.15** ± 2.28	**85.15** ± 2.28	85.00 ± 2.32	85.10 ± 2.36	**85.15** ± 2.36	84.85 ± 2.28
3	78.07 ± 1.21	78.07 ± 1.21	**78.74** ± 1.55	78.21 ± 1.19	78.21 ± 1.19	78.07 ± 1.37	77.94 ± 1.29	77.94 ± 1.29	78.07 ± 1.29	78.21 ± 1.43
4	89.56 ± 0.66	**89.65** ± 0.69	88.48 ± 1.02	89.32 ± 1.05	89.32 ± 1.05	89.13 ± 1.10	89.60 ± 0.92	89.60 ± 0.69	89.27 ± 0.69	89.61 ± 0.95
5	76.64 ± 5.17	76.77 ± 5.37	76.76 ± 5.41	**77.03** ± 5.63	**77.03** ± 5.63	**77.03** ± 5.63	**77.03** ± 5.17	76.90 ± 5.39	**77.03** ± 5.39	76.89 ± 5.20
6	86.01 ± 5.50	86.01 ± 5.50	85.71 ± 3.72	**86.61** ± 5.33	**86.61** ± 5.33	86.31 ± 5.71	86.31 ± 4.69	86.31 ± 5.32	**86.61** ± 5.32	86.31 ± 4.21
7	70.84 ± 0.96	70.84 ± 1.02	65.95 ± 1.79	70.07 ± 1.85	70.07 ± 1.85	68.01 ± 1.84	70.84 ± 1.02	70.84 ± 1.10	70.07 ± 1.10	**71.31** ± 0.76
8	75.50 ± 2.51	75.60 ± 2.59	76.30 ± 1.85	75.20 ± 2.43	75.20 ± 2.43	75.60 ± 2.56	75.90 ± 2.58	75.70 ± 2.51	75.40 ± 2.51	**76.40** ± 2.27
9	61.18 ± 5.46	61.17 ± 6.83	60.66 ± 9.64	62.57 ± 7.40	62.57 ± 7.40	62.09 ± 8.45	61.19 ± 6.50	62.12 ± 7.00	62.09 ± 7.00	**64.47** ± 8.67
10	74.18 ± 4.44	74.18 ± 4.44	**76.14** ± 7.28	73.86 ± 4.76	73.86 ± 4.76	73.86 ± 5.02	74.18 ± 4.44	74.18 ± 4.44	73.86 ± 4.44	74.18 ± 4.86
11	83.70 ± 6.75	83.33 ± 6.12	82.22 ± 4.80	**84.08** ± 6.84	**84.08** ± 6.84	82.59 ± 5.80	83.70 ± 6.24	83.70 ± 6.24	83.70 ± 6.24	82.96 ± 5.54
12	71.36 ± 1.98	71.18 ± 2.04	70.50 ± 3.01	**71.70** ± 2.69	**71.70** ± 2.69	71.53 ± 2.33	71.36 ± 2.82	71.53 ± 2.48	71.53 ± 2.48	71.18 ± 2.67
13	85.76 ± 3.36	85.76 ± 3.36	84.92 ± 5.09	85.47 ± 3.76	85.47 ± 3.76	85.20 ± 3.97	85.76 ± 3.36	85.76 ± 3.36	85.76 ± 3.36	**86.32** ± 4.40
14	80.30 ± 5.54	79.36 ± 5.66	**82.20** ± 8.55	79.36 ± 5.66	79.36 ± 5.66	77.95 ± 6.92	79.83 ± 4.89	79.83 ± 4.89	78.90 ± 4.89	80.77 ± 4.33
15	73.95 ± 3.12	**74.24** ± 3.50	69.14 ± 3.59	73.67 ± 3.82	73.67 ± 3.82	71.97 ± 4.48	**74.24** ± 3.50	73.96 ± 3.91	73.38 ± 3.91	73.95 ± 2.43
16	83.10 ± 2.10	83.10 ± 2.10	82.40 ± 1.88	82.80 ± 1.82	82.80 ± 1.82	82.75 ± 1.79	**83.20** ± 2.19	83.10 ± 2.10	82.80 ± 2.10	83.00 ± 2.13
17	68.70 ± 3.77	68.12 ± 4.19	68.42 ± 4.83	68.70 ± 3.77	68.70 ± 3.77	68.70 ± 2.60	68.99 ± 4.06	68.70 ± 4.23	68.70 ± 4.23	**69.27** ± 5.22
18	79.29 ± 0.49	79.28 ± 0.45	79.24 ± 0.55	79.27 ± 0.50	79.27 ± 0.50	79.28 ± 0.50	79.38 ± 0.56	79.28 ± 0.45	79.27 ± 0.45	**79.44** ± 0.66
19	83.37 ± 2.31	83.49 ± 2.36	82.77 ± 2.14	83.61 ± 2.07	83.61 ± 2.07	83.61 ± 2.31	84.10 ± 2.36	83.73 ± 2.39	83.86 ± 2.39	**84.46** ± 2.20
20	81.48 ± 2.96	82.41 ± 2.49	77.32 ± 5.76	82.18 ± 2.46	82.18 ± 2.46	79.86 ± 3.28	83.10 ± 2.82	82.87 ± 2.85	82.41 ± 2.85	**85.19** ± 3.98
21	77.17 ± 0.86	77.18 ± 0.89	77.07 ± 1.16	77.24 ± 0.89	77.24 ± 0.89	77.24 ± 0.89	**77.54** ± 0.97	77.22 ± 0.92	77.24 ± 0.92	77.46 ± 1.16
22	92.64 ± 1.69	92.60 ± 1.67	92.21 ± 0.96	92.64 ± 1.45	92.64 ± 1.45	92.38 ± 1.39	92.82 ± 1.53	92.60 ± 1.67	92.69 ± 1.67	**92.86** ± 1.36
23	79.34 ± 6.67	79.34 ± 6.67	73.59 ± 10.62	78.85 ± 6.31	78.85 ± 6.31	78.39 ± 6.29	79.36 ± 7.57	79.83 ± 6.54	78.85 ± 6.54	**79.85** ± 7.13
24	96.39 ± 1.17	96.39 ± 1.17	95.23 ± 1.65	96.24 ± 1.39	96.24 ± 1.39	96.24 ± 1.39	96.53 ± 1.13	96.39 ± 1.17	96.24 ± 1.17	**96.96** ± 0.98
25	76.95 ± 2.15	76.83 ± 1.77	75.53 ± 1.77	76.48 ± 2.10	76.48 ± 2.10	76.12 ± 2.19	76.95 ± 1.95	76.83 ± 1.77	76.48 ± 1.77	**77.42** ± 2.79
26	**87.00** ± 4.12	**87.00** ± 4.12	84.00 ± 5.03	86.00 ± 4.32	86.00 ± 4.32	86.00 ± 4.32	86.67 ± 3.94	86.67 ± 3.94	86.00 ± 3.94	86.33 ± 4.96
27	**96.67** ± 1.65	96.49 ± 1.79	96.49 ± 1.16	96.14 ± 1.44	96.14 ± 1.44	96.32 ± 1.32	**96.67** ± 1.65	96.49 ± 1.79	96.14 ± 1.79	**96.67** ± 1.86
28	86.32 ± 0.74	86.30 ± 0.73	86.12 ± 0.67	**86.40** ± 0.87	**86.40** ± 0.87	86.38 ± 0.84	86.36 ± 0.68	86.30 ± 0.73	86.38 ± 0.73	86.34 ± 0.73
29	82.10 ± 3.32	82.10 ± 3.32	80.43 ± 4.18	82.10 ± 3.32	82.10 ± 3.32	82.10 ± 3.32	81.77 ± 3.96	81.77 ± 3.22	82.10 ± 3.22	**82.76** ± 3.47
30	98.85 ± 1.63	98.85 ± 1.63	**98.87** ± 1.60	98.85 ± 1.63	98.85 ± 1.63	98.30 ± 1.71	98.85 ± 1.63	96.07 ± 2.28	98.85 ± 2.28	98.85 ± 1.63
WS	+0.01468	+0.00672	+0.0003	+0.00288	+0.00288	+0.00008	+0.04884	+0.00714	+0.00108	n/a

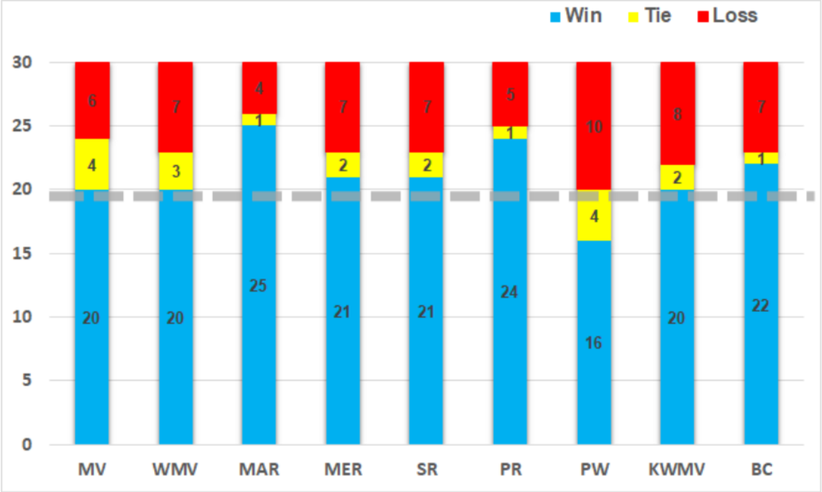

Fig. 4. Pairwise comparison of CBF:WD-DF with literature methods for combination.

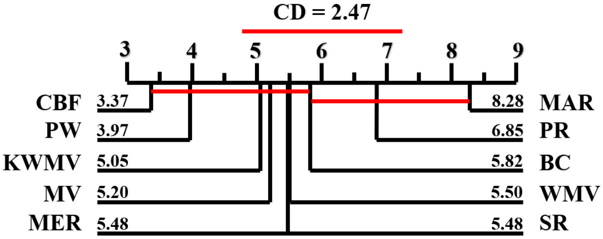

Fig. 5. Friedman and Nemenyi post hoc tests comparing the combination approaches.

The *cv* value estimation takes into account the number of experiments with a significance level of $\alpha = 0.05$. The whole set of experiments suggests CBF as a promising strategy. The new approach prevails in 189 out of 270 experiments (70%), tied in 20 cases (7.41%), and lost in 61 cases (22.59%).

Friedman and the Nemenyi post hoc test are present in Figure 5. CBF is considered statistically different from most of the approaches in the literature. One may observe that the proposed method is statistically different from MAR and PR. However, it is similar to 7 out of 9 literature approaches according to the critical distance (CD). The approach presented the lowest average rank, suggesting CBF as a good competitor, given the literature methods. Wilcoxon test was also performed to compare CBF:WD-DF against each of the 9 approaches in a pairwise fashion. The results present in Table 4 show that the proposed method is statistically significant for all pairwise comparisons, considering $\alpha = 0.05$ significance level. A tie is considered concerning the PW value (0.04884).

4.1.3. *Discussion*

Our discussion considers two important points: (i) the comparison between the centrality measures as parameters for CBF and (ii) the comparison between the approach with the literature. First, the experiments suggest the degree centrality as the best result regarding the relation of centrality measures with the ensemble accuracy. The centrality exploited the diversity between the classifiers lending more importance to classifiers with high diversity relationships; in this case, only the direct neighbors (adjacent members) are considered. A diverse classifier may have a large neighborhood (unweighted degree), strong relations with their neighborhood (weighted degree), or both. The betweenness measure was not better than the degree centrality but presented interesting results. For this measure, a classifier may appear in many geodesics; however, as the length of the geodesics is usually short (a few classifiers), resulting in just a few classifiers being very important. Eigenvector and local centralities did not generate better results compared to degree centrality, even being more complex, because both estimates the centrality score based on the relationship a classifier has with its neighbors (adjacent members) and also their relationship with their respective neighbors. These measures are less sensitive to their direct neighbors compared to degree. Closeness did not show promising results too. The *DF* pairwise diversity measure provided a distinct contribution compared to the others, suggesting that classifiers should avoid common errors instead of only one being different from others.

The comparison between approaches suggests that CBF is, in most cases, superior to literature approaches. Therefore, the ensemble's diversity and accuracy are essential, and it must also be considered to design fusion approaches. The critical value (cv) reveals the existence of a statistical difference between methods, except PW. Critical distance (CD) indicates some approaches being statistically similar. The lowest average rank observed in CBF:WD-DF may suggest the approach being an interesting competitor when approaches are applied to a wide range of classification problems.

4.2. *CBS evaluation*

The same best setup found for the CBF method is used here. The Double-Fault (DF) diversity measure is used to represent the relationship between the classifiers in the network. The Weighted Degree (WD) centrality is used to analyze the pairwise diversity relations and score each classifier according to its importance to the ensemble.

The following literature approaches are compared with the CBS method: the Aggregation ordering in Bagging (AGOB),[24] Diversity Regularized Ensemble Pruning (DREP),[25] Pruning in Ordered Bagging Ensembles (POBE),[26] and Kappa Pruning.[27] The latter has the size of the subset fixed in $T/2$. For DREP, the parameter p is 0.5, the same used in CBS. For comparison purposes, we also used the best performing classifier (*Single Best* — *SB*) according to the estimation in S_{val}, and the fusion of all classifiers using Majority Vote (MV), i.e., no selection or pruning adopted. Both strategies, SB and MV, work as a baseline to identify the problems in which a simple selection or the lack of selection is better than the complex process of selecting a sub-optimal subset.

Table 5 presents the average accuracy of the approaches considering the different classification problems. The best results are in bold, and *#Best* is the number of best results for each approach. CBS had the best results in 12 over 30 problems. According to the Wilcoxon signed test, CBS is statistically different (+) from some strategies, as suggested by the p-value ($p \leq 0.05$) with 95% of confidence.

The CBS approach, compared to the others, presented a high number (on average) of selected classifiers for the wine dataset (see Table 5). It is related to the small difference between classifiers. So, a pool with very similar classifiers can be an issue to the selection process performed in the CBS method.

Another comparison to the literature is present in Figure 6. The pairwise comparison considers CBS and each approach from literature identified below the column. The dashed line represents the critical value (cv), which for 30 classification problems is $cv = 19.5$. Therefore, CBS is statistically different from the greedy selection literature, except for Kappa Pruning, which CBS did not reach 20 victories (rounding up the cv).

All approaches were compared statistically with the Friedman test and the post hoc Nemenyi. Figure 7 presents the scores and critical distance. The score is based on the rank average, in which the lower the value, the better. The order is related to the scores, and the best-ranked methods are on the left. CBS, Kappa, and DREP are considered similar according to the critical distance, while CBS is different from AGOB and POBE.

Figure 8 presents the average accuracy of each approach with the respective average number of selected classifiers (estimated over the 30 classification problems). The new approach presents the best recognition rate and an attractive amount of selected ensemble members. It suggests an attractive performance of CBS by a low number of classifiers.

4.3. *Discussion*

The literature of greedy selection compared to the new approach suggests that CBS is an attractive alternative. The number of selected classifiers is usually less than 20% given an initial pool. The wine dataset, which is a small dataset with too similar classifiers, showed the non-practical result of 97 classifiers selected on average. Despite the low number of selected classifiers in the datasets glass, haberman, ionosphere, and sonar, CBS seems to not perform well in small datasets.

The stop criteria is inspired in conventional greedy selection approaches, so, the algorithm stops selecting a candidate classifier $c_i \in CR$ when it does not maintain or increase the accuracy obtained by the subset C'. It was observed that usually

Table 5. Comparison of CBS with literature approaches. The present values are the average of the recognition tax obtained by a 6-fold cross-validation. ($|C'|$) refers to the size of the selected subset (on average). Best results of the classification problems are in bold. #B is the number of the best results for each approach. WS stands for the p-value of the Wilcoxon signed test, and (+) stands for statistically different comparisons considering $\alpha = 0.05$.

	Kappa	$\|C'\|$	DREP	$\|C'\|$	POBE	$\|C'\|$	AGOB	$\|C'\|$	SB	$\|C'\|$	MV	$\|C'\|$	CBS	$\|C'\|$
1	87.25 ± 2.87	50	84.21 ± 2.48	4	86.65 ± 2.67	13	**87.84** ± 2.64	5	85.66 ± 3.05	1	87.83 ± 2.84	100	85.37 ± 1.97	4
2	84.35 ± 1.17	50	**84.80** ± 1.04	18	84.35 ± 1.10	26	83.89 ± 1.09	15	84.50 ± 0.65	1	84.70 ± 1.12	100	84.20 ± 0.71	33
3	76.88 ± 2.45	50	77.55 ± 2.75	6	76.88 ± 2.60	23	76.42 ± 2.63	22	77.82 ± 3.47	1	77.55 ± 2.52	100	**77.82** ± 3.95	12
4	88.52 ± 1.32	50	89.05 ± 0.92	3	88.52 ± 1.12	16	88.02 ± 1.08	5	88.48 ± 0.34	1	**89.32** ± 1.16	100	**89.32** ± 0.41	5
5	75.85 ± 3.31	50	77.16 ± 2.78	5	75.25 ± 3.05	29	75.41 ± 2.99	17	76.25 ± 4.67	1	76.12 ± 2.91	100	**77.55** ± 2.38	8
6	85.42 ± 2.81	50	84.23 ± 3.49	5	85.42 ± 3.15	25	84.93 ± 3.21	27	83.33 ± 4.33	1	**86.01** ± 2.99	100	**86.01** ± 3.48	13
7	69.19 ± 2.78	50	68.99 ± 2.15	3	60.99 ± 2.46	14	69.08 ± 2.40	2	69.40 ± 1.91	1	**69.81** ± 2.98	100	69.60 ± 1.67	5
8	76.00 ± 2.58	50	75.10 ± 2.34	4	74.20 ± 2.46	8	75.62 ± 2.44	4	72.50 ± 2.08	1	**76.20** ± 3.57	100	75.30 ± 2.21	8
9	65.36 ± 11.42	50	59.78 ± 8.46	3	65.96 ± 9.94	9	65.03 ± 9.64	10	57.90 ± 6.35	1	**67.22** ± 12.60	100	60.73 ± 7.95	7
10	**77.78** ± 4.18	50	76.47 ± 5.66	23	76.78 ± 4.92	45	77.45 ± 5.07	41	73.86 ± 7.82	1	76.47 ± 5.43	100	72.88 ± 4.72	25
11	81.85 ± 7.85	50	82.22 ± 8.51	5	82.05 ± 8.18	14	81.29 ± 8.24	20	80.37 ± 7.53	1	82.96 ± 6.24	100	**83.33** ± 4.76	11
12	71.01 ± 1.71	50	71.19 ± 2.51	4	70.11 ± 2.11	23	70.52 ± 2.19	19	71.19 ± 2.34	1	71.53 ± 2.50	100	**72.73** ± 1.89	7
13	86.04 ± 1.80	50	81.20 ± 2.20	2	**86.84** ± 2.00	32	85.70 ± 2.04	29	84.33 ± 1.58	1	85.76 ± 2.88	100	80.63 ± 3.35	2
14	81.18 ± 6.88	50	82.58 ± 7.05	4	78.88 ± 6.96	26	80.77 ± 6.98	23	82.12 ± 5.71	1	**83.06** ± 5.48	100	82.13 ± 6.06	10
15	70.85 ± 5.20	50	71.68 ± 2.08	6	69.65 ± 3.64	10	70.60 ± 3.33	7	70.27 ± 3.68	1	71.98 ± 4.12	100	**73.68** ± 4.14	6
16	**84.80** ± 1.83	50	84.25 ± 1.61	4	83.60 ± 1.72	5	84.58 ± 1.70	9	82.20 ± 2.47	1	83.45 ± 1.80	100	84.20 ± 1.97	3
17	70.19 ± 5.83	50	69.33 ± 6.22	4	69.99 ± 6.02	31	69.63 ± 6.06	26	66.37 ± 2.26	1	68.15 ± 5.74	100	**70.48** ± 8.77	5
18	79.35 ± 0.34	50	**79.63** ± 0.32	6	79.25 ± 0.33	26	79.02 ± 0.33	2	79.43 ± 0.27	1	79.43 ± 0.29	100	79.49 ± 0.37	8
19	82.76 ± 3.25	50	82.17 ± 1.87	3	82.36 ± 2.56	23	82.21 ± 2.42	28	83.73 ± 1.44	1	83.01 ± 2.39	100	**84.22** ± 2.02	16
20	80.56 ± 3.49	50	86.11 ± 4.88	3	82.86 ± 4.19	17	79.94 ± 4.32	4	**87.50** ± 3.31	1	82.87 ± 3.73	100	86.80 ± 3.82	34
21	77.31 ± 1.09	50	**77.98** ± 1.11	6	69.71 ± 1.10	23	76.93 ± 1.10	28	77.74 ± 1.07	1	77.42 ± 1.03	100	77.79 ± 1.16	9
22	92.73 ± 0.67	50	92.77 ± 0.57	4	91.67 ± 0.62	38	**94.13** ± 0.61	35	92.77 ± 0.66	1	92.51 ± 0.59	100	93.08 ± 0.32	8
23	78.87 ± 3.50	50	73.53 ± 4.37	7	**81.28** ± 3.94	26	78.07 ± 4.02	22	70.14 ± 7.41	1	80.31 ± 3.83	100	77.87 ± 3.27	12
24	95.66 ± 2.01	50	95.66 ± 1.59	5	93.36 ± 1.80	35	95.16 ± 1.76	13	**96.68** ± 1.45	1	95.80 ± 1.38	100	96.24 ± 1.48	16
25	77.90 ± 4.14	50	76.83 ± 2.67	4	76.70 ± 3.41	9	77.20 ± 3.26	5	77.42 ± 4.84	1	77.90 ± 3.46	100	**78.25** ± 2.83	5
26	86.33 ± 4.82	50	86.33 ± 5.22	3	86.23 ± 5.02	12	85.13 ± 5.06	13	**87.00** ± 4.86	1	86.33 ± 4.38	100	85.67 ± 6.05	6
27	96.84 ± 1.61	50	95.60 ± 1.89	6	97.04 ± 1.75	11	96.24 ± 1.78	8	96.48 ± 1.59	1	96.66 ± 1.54	100	**97.19** ± 1.45	47
28	86.56 ± 0.89	50	85.98 ± 1.06	3	85.01 ± 0.97	5	**87.36** ± 0.99	4	86.26 ± 1.11	1	86.70 ± 1.11	100	86.32 ± 1.03	5
29	80.80 ± 6.92	50	79.48 ± 3.28	4	81.00 ± 5.10	28	80.90 ± 4.74	24	80.14 ± 7.49	1	81.14 ± 7.37	100	**81.46** ± 5.82	8
30	**97.78** ± 2.49	50	97.76 ± 1.58	75	**97.78** ± 2.03	23	**97.78** ± 1.94	17	96.63 ± 2.77	1	97.20 ± 2.30	100	97.20 ± 2.30	97
#B	2		3		3		4		3		6		**12**	
WS	0.3576		+0.0060		+0.03572		0.1010		+0.0080		0.9442			

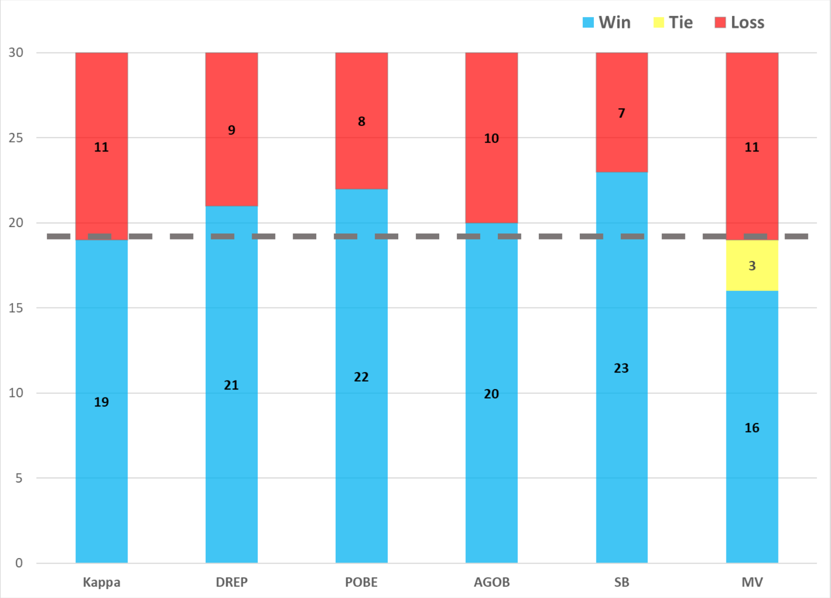

Fig. 6. Pairwise comparison between selection approaches.

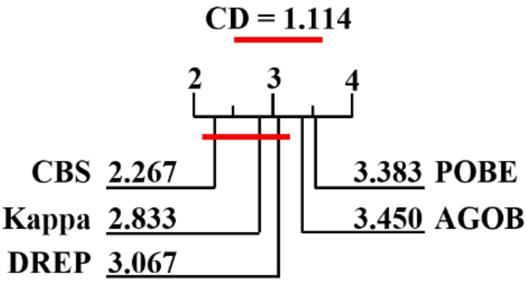

Fig. 7. Friedman test and the post hoc Nemenyi (CBS).

less than 20% of the original pool is maintained in the subset C'. The exception was observed in the wine dataset, due to the accurate but low diverse classifiers.

The pairwise comparison of CBS with the literature approaches suggested that the new approach is superior in 82 cases (68.33%), and is worse in 38 cases (31.67%). It also presented a statistical difference between DREP, POBE, and AGOB. The approach performed better than Kappa, SB, and MV, but without sufficient statistical difference. It is also superior to SB and MV.

The greedy approaches usually are based on different aspects when selecting the classifiers to compose the ensemble, such as the measure to ordering the classifier candidates (usually individual accuracy), the criteria to select them (increase ensemble accuracy or diversity), and the number of classifiers to be selected (fixed

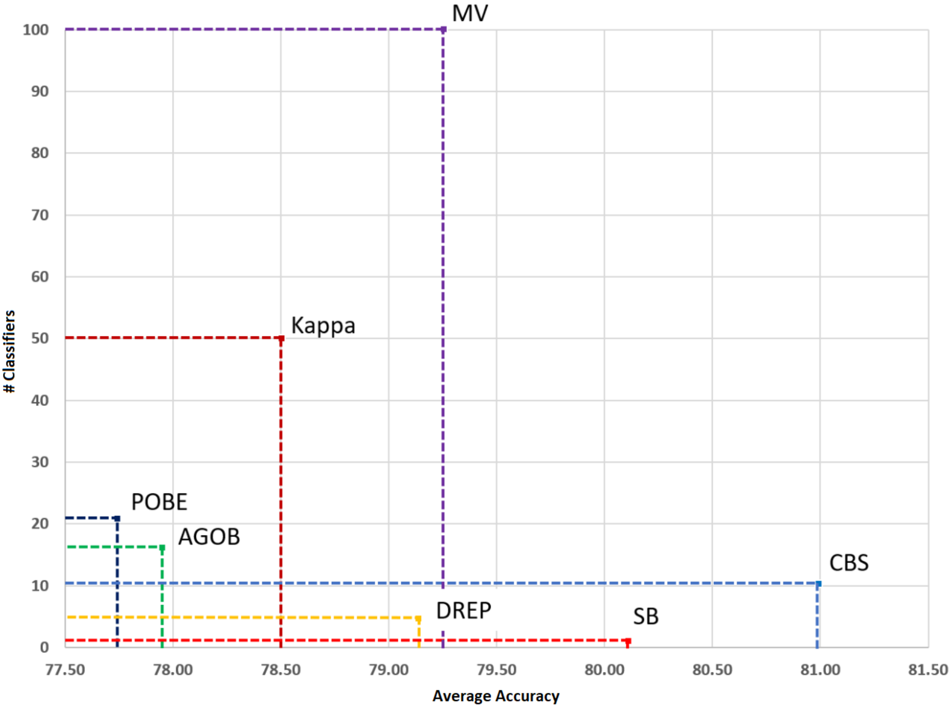

Fig. 8. Accuracy and the subset size of different selection approaches.

or flexible). The CBS approach considers each one of these primary aspects. First, the new method ranked the classifiers according to the importance of each one. To this end, the pairwise relation DF is analyzed by the degree centrality to point out higher scores for the essential ensemble members. It is similar to the complementariness measures adopted by the literature. Another aspect of the complementariness measure is that for some approaches, it measures the gain in relation with the subset C' while others consider it in relation to its direct pairs, individually. In the selection criteria, CBS considers the accuracy contribution of the candidate c_i, similarly to DREP. Both approaches choose the most accurate classifier between the most promising classifiers, i.e., the most complementary classifiers to the subset C'. Kappa only selects the n-best classifiers, avoiding the additional test that CBS and DREP perform.

5. Conclusion

This paper presents a novel ensemble fusion method along with a selection method. Both strategies are based on the concept of centrality in the context of complex network theory. The proposed CBF method considers the ensemble of classifiers as a complex network to analyze the diversity between the classifiers. Each classifier in that network has its influence estimated using centrality measures. The score of such measures combined with accuracy provided the weight used in the fusion and the selection process.

The experimental results on 30 classification problems confirmed our hypothesis. The centrality concept is a promising strategy for weighting the decisions of the classifiers within the ensemble for the fusion method. Different measures, such as pairwise diversity and centrality measures, were evaluated to find out the best setup for the proposed method. Double fault pairwise diversity measure revealed better results concerning the ensemble network relationship while the weighted degree is the distinct centrality measure closer to ensemble accuracy used to estimate the importance of every classifier. The experimental results showed the proposed fusion method presents the best accuracy compared to nine different fusion methods in the literature. It prevailed on 14 of 30 classification problems. The second-best method in that comparison presented the best accuracy in just six cases. A total of 270 comparisons were made, and CBF won in 189 out of 270 experiments (70%), while lost in 61 cases (22.59%).

In the proposed static selection method based on the complex network theory, the CBS method, the ensemble network is analyzed, taking into account the centrality information. Then, the analysis of the most distinct classifiers based on accuracy defines which ones should be added to a sub-optimal subset.

The robust experimental protocol confirms our hypothesis. Therefore, a centrality measure to scoring the classifiers' importance is an attractive searching strategy for selecting diverse and accurate classifiers. The results suggested that CBS was able to find the best result in 12 out of 30 problems, which corresponds to the best performance compared to other selection approaches. A pairwise comparison between approaches with a critical distance measurement also suggests that the new approach is an interesting alternative. In comparison with selection methods in the literature and the baselines single best and majority vote, CBS showed the best results on 67.22% of the experiments. According to the Nemenyi post hoc test, CBS is distinct from POBE and AGOB. One last comparison suggests that the new method performs better than literature at a reduced number of classifiers, i.e., 10% of the original pool size. Further work is suggested to track the behavior of CBS concerning other pool generator strategies and different base classifiers. Besides, the centrality was analyzed in a static network context to define a static subset; therefore, it could also be used to select classifiers dynamically.

Acknowledgment

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

References

1. L. Breiman, Bagging predictors, *Machine Learning* **24**(2) (1996) 123–140.
2. R. E. Schapire, Y. Freund, P. Bartlett and W. S. Lee, Boosting the margin: A new explanation for the effectiveness of voting methods, *Annals of Statistics* **26**(5) (1998) 1651–1686.
3. T. K. Ho, The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8) (1998) 832–844.
4. A. S. Britto, Jr., R. Sabourin and L. E. S. Oliveira, Dynamic selection of classifiers — A comprehensive review, *Pattern Recognition* **47**(11) (2014) 3665–3680.
5. J. Kittler, M. Hatef, R. Duin and J. Matas, On combining classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (March 1998) 226–239.
6. L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, 2nd edn. (John Wiley & Sons, 2014).
7. L. I. Kuncheva, *Combining Pattern Classifiers* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2004).
8. R. A. Silva, A. S. Britto, F. Enembreck, R. Sabourin and L. S. Oliveira, Fusion of classifiers based on centrality measures, in *2018 IEEE 30th Int. Conf. on Tools with Artificial Intelligence* (*ICTAI*) (2018), pp. 363–370.
9. K. Trawinski and O. Cordon, A network-based approach for diversity visualization of fuzzy classifier ensembles, in *2016 IEEE Int. Conf. on Fuzzy Systems* (*FUZZ-IEEE*) (IEEE, 2016), pp. 923–930.
10. R. A. Silva, A. S. Britto, Jr, F. Enembreck, R. Sabourin and L. E. S. de Oliveira, CSBF: A static ensemble fusion method based on the centrality score of complex networks, *Computational Intelligence* **36**(2) (2020) 522–556.
11. L. C. Freeman, Centrality in social networks conceptual clarification, *Social Networks* **1**(3) (1978) 215–239.
12. J. M. Anthonisse, The rush in a directed graph, *Stichting Mathematisch Centrum. Mathematische Besliskunde* **BN 9/71**, p. 10.
13. P. Bonacich, Factoring and weighting approaches to status scores and clique identification, *Journal of Mathematical Sociology* **2** (1972) 113–120.
14. S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, 1994).
15. F. Zhou, S. Mahler and H. Toivonen, Simplification of networks by edge pruning, *Bisociative Knowledge Discovery*, Vol. 7250 of *Lecture Notes in Computer Science* (2012), pp. 179–198 [Modified version of article "Network simplification with minimal loss of connectivity" in *ICDM 2010*].
16. F. Zhou, S. Malher and H. Toivonen, Network simplification with minimal loss of connectivity, *2010 IEEE Int. Conf. on Data Mining* (2010), pp. 659–668.
17. M. Lichman, UCI Machine Learning Repository (2013).
18. R. P. W. Duin, P. Juszczak, D. de Ridder, P. Paclik, E. Pekalska and D. M. Tax, PRTOOLS, a Matlab Toolbox for Pattern Recognition (2004).
19. R. D. King, C. Feng and A. Sutherland, STATLOG: Comparison of classification algorithms on large real-world problems, *Applied Artificial Intelligence* **9**(3) (1995) 289–333.
20. L. Kuncheva, Ludmila Kuncheva Collection LKC (2018).
21. J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez and F. Herrera, KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework, *Journal of Multiple-Valued Logic and Soft Computing* **17**(2–3) (2011) 255–287.

22. L. Rokach, Ensemble-based classifiers, *Artificial Intelligence Review* **33**(1–2) (2010) 1–39.
23. W. L. Buntine, A Theory of Learning Classification Rules, PhD thesis, University of Technology, Sydney (1992).
24. G. Martínez-Muñoz, A. Suárez, G. Martínez-Muñoz and A. Suárez, Aggregation ordering in bagging, in *Proc. of the IASTED Int. Conf. on Artificial Intelligence and Applications* (2004), pp. 258–263.
25. N. Li, Y. Yu and Z. H. Zhou, Diversity regularized ensemble pruning, *Machine Learning and Knowledge Discovery in Databases*, Vol. 7523 of *Lecture Notes in Computer Science* (2012), pp. 330–345.
26. G. Martínez-Muñoz and A. Suárez, Pruning in ordered bagging ensembles, in *Proc. of the 23rd Int. Conf. on Machine Learning* (*ICML '06*) (2006), pp. 609–616.
27. D. D. Margineantu and T. G. Dietterich, Pruning adaptive boosting, in *Proc. of the Fourteenth Int. Conf. on Machine Learning* (*ICML '97*) (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997), pp. 211–218.