Fusion of Classifiers based on Centrality Measures

Ronan A. Silva[1,4], Alceu S. Britto Jr.[1,5], Fabricio Enembreck[1], Robert Sabourin[2], and Luis S. Oliveira[3]

[1]Pontifical Catholic University of Parana (PUCPR), Curitiba, PR, Brazil
[2]École de Technologie Supérieure (ÉTS), Montreal, QC, Canada
[3]Federal University of Parana (UFPR), Curitiba, PR, Brazil
[4]Federal Institute of Parana (IFPR), Telemaco Borba, PR, Brazil
[5]State University of Ponta Grossa (UEPG), Ponta Grossa, PR, Brazil
Email: {ronan.silva}@pucpr.edu.br

***Abstract*—This paper presents the Centrality Based Fusion (CBF) method for ensemble fusion which is based on the centrality measures in the context of complex network theory. Such a concept has been applied in Social Network Analysis to measure the importance of each person inside of a social network. We hypothesized that the centrality of each classifier inside of an ensemble represented as a complex network could be combined with accuracy to provide the weight for its decision during the ensemble fusion. The main idea is to derive the weight considering the classifier importance inside the ensemble network which reflects the classifiers' diversity. A robust experimental protocol based on 30 datasets has confirmed that the notion of prominence provided employing centrality measures is a promising strategy to weight the classifiers of an ensemble. When compared with 9 fusion methods of the literature, the proposed fusion method won in 189 out of 270 experiments (70%), lost in 61 cases (22.59%) and tied in 20 cases (7.41%).**

***Index Terms*—Fusion methods, Diversity, Centrality Measures, Ensemble of Classifiers, Multiple Classifier Systems.**

I. Introduction

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. So, covering the entire problem space is a responsibility divided among the members of a team composed of diverse and accurate classifiers. This diversity derives from the fact that the members should make different errors, and as a result, merging their decisions may lead to an improvement in classification performance.

The literature presents different approaches to generate an ensemble, categorized as heterogeneous or homogeneous. The former uses different base classifiers to achieve diversity. The later uses the same base classifier, but vary the data used for training the elements which will constitute the ensemble. Bagging [1], Boosting [2] and Random Subspaces [3] are classical pool generators.

No matter the generation method used, in an ensemble of classifiers $C = \{c_1, c_2, \ldots, c_T\}$, every 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 $W = \{w_1, w_2, \ldots, w_M\}$. Fusion methods are applied to the decisions of all the classifiers in the ensemble, or on the decisions of a subset of classifiers selected statically or dynamically [4] producing the ensemble's final decision. The literature provides a variety of fusion methods as seen in [5], and in [6]. The most used, and perhaps the simplest fusion method is the majority vote (also known as plurality vote). It considers each classifier as equal concerning their influence and assigns the class label w_i to x if the majority of classifiers support the decision. However, this technique can sometimes perform worse than an individual classifier in the ensemble. The most interesting alternative has been to weight the vote of the classifiers while assuming that they compete with one another in assigning the correct class label [6]. The competition among the classifiers in the ensemble through the use of weights has been shown to be very promising. The literature thus supports a wide variety of static and dynamic strategies for weighting the decision of each classifier in an ensemble. A static weighting strategy estimates the ensemble member's influence during the training phase of a classification system and the weights obtained remain the same during the test phase, while in a dynamic strategy, each classifier receives a different weight for each test instance.

In the present context, the challenge, irrespective of the weighting strategy (static or dynamic), is how to combine diverse classifiers, considering not only their competence based on accuracy individually but also their competence working together. Therefore, a thorough analysis is needed to evaluate the classifiers and their interactions, after which a proper fusion procedure defines how the ensemble vote. In fact, most of the combination methods do not explore the classifier interactions properly. They usually ignore interactions and are only based on the classifiers' performance or confidence. While the fusion methods continue ignoring the relationship between classifiers, it inspired a variety of different classifier selection schemes [4], [7], in which is found, e.g., static selection based on diversity and accuracy information.

This paper presents a novel approach for combining classifiers, a static method for classifier decision fusion based on centrality measures computed on complex networks constructed from a given ensemble. Social Network Analysis (SNA) employs centrality measures to understand a variety of problems, due to the ability to estimate the importance of each member (vertex), by measuring his influence based on the network relations. The hypothesis is that the centrality of each classifier within an ensemble represented as a complex

network may be combined with the classifier's accuracy to weight the classifier decision. The idea here is to derive the weight for each classifier based on its importance for the diversity of the ensemble.

This work presents five distinct sections. Section II introduces some essential concepts and definitions related to the proposed fusion method. Section III describes the new approach, while Section IV shows our experimental results and corresponding discussions. Finally, Section V presents our conclusion and future work perspectives.

II. Background Theory

This section presents some important concepts related to the proposed fusion method, such as the pairwise diversity measures and the complex network theory used to design the proposed fusion method.

A. Diversity Measures

Diverse opinions among members are important in an ensemble. Some complementarity is observed when they commit different mistakes. So, the expectation is that by increasing the diversity, the errors committed by the ensemble may decrease.

Diversity can be estimated taking into account pairs of classifiers (also known as pairwise diversity) or the whole ensemble. For the first approach, several measures are compared in [8], as follows: Q statistics [9], Correlation Coefficient [10], Disagreement measure [11] and Double Fault measure [12].

Pairwise measures are based on the relation of incorrect/correct predictions between classifiers. Such a relation between two classifiers, c_i and c_j, is presented in Table I. 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 relation can be estimated using a training set or a validation set.

TABLE I
PAIRWISE RELATION BETWEEN TWO CLASSIFIERS c_i AND c_j.

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

A well-known measure is Q statistics (QS), which is denoted by Equation 1, 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.

$$QS = \frac{N^{11} \times N^{00} - N^{01} \times N^{10}}{N^{11} \times N^{00} + N^{01} \times N^{10}} \quad (1)$$

A similar measure is the Correlation Coefficient (CC), which can be computed as shown in Equation 2.

$$CC = \frac{N^{11} \times N^{00} - N^{01} \times N^{10}}{\sqrt{\Delta}} \quad (2)$$

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

The Disagreement measure (Dis) estimates the number of observations in which one classifier is incorrect, while the other is correct. The diversity of the pair of classifiers is represented by higher scores of Dis in the interval $[0, 1]$. This measure is denoted by Equation 3.

$$Dis = \frac{N^{01} + N^{10}}{N^{00} + N^{01} + N^{10} + N^{11}} \quad (3)$$

Another interesting pairwise measure is the Double Fault (DF) [12]. It is defined as the number of examples that have been misclassified by both classifiers c_i and c_j, as denoted by Equation 4. In contrast to the Dis measure, DF represents diversity by lower scores in the interval $[0, 1]$.

$$DF = \frac{N^{00}}{N^{00} + N^{01} + N^{10} + N^{11}} \quad (4)$$

Despite the high interest of the machine learning community in the use of diversity, the relationship between diversity and the ensemble accuracy is still an open problem [6]. Therefore, in this paper, we have evaluated popular diversity measures to construct a network with the classifiers of an ensemble. The goal is to assess the classifier influence in the ensemble considering its centrality in the network. In the next section, we present some basic concepts and definitions related to complex networks that are needed to understand our approach.

B. Complex Networks

The complex network theory is largely applied in Social Network Analysis (SNA). It allows the estimation of the importance of each member of a social network, taking into account its interaction with other members. An important concept within this theory is the centrality of a member, which is related to its central location in the social network [13]. The centrality of a member may provide some information about its importance or prominence in the network. The simplest way to estimate this centrality is to count the number of adjacent members. However, the literature provides different strategies for estimating centrality, as shown in this section.

As in the SNA, the idea here is to use the "network member importance" or centrality concept. It will be applied to differentiate the classifiers in the fusion process, i.e., to weight their decisions. To that end, we need to represent the ensemble as a complex network. We thus represent an ensemble $C = \{c_1, c_2, \ldots, c_T\}$, composed of T classifiers, as a graph $G(V, E)$, where the set of vertices V corresponds to the ensemble of classifiers C, and the set of edges E denotes the relationship between pairs of classifiers; in our case, this refers to the diversity among them. The centrality measures are described next, already considering an ensemble network.

1) Centrality Measures: Centrality measures are designed to rank the members of a network based on their topological importance. The following four measures are commonly used in the literature: Degree [14], Betweenness [15], Closeness [14] and Eigenvector [16]. These measures focus different

pieces of information, such as a) the number of edges (degree); b) shortest paths (closeness, betweenness), and c) walks, in which vertices and edges are possibly revisited (eigenvector).

The Degree centrality K_{c_i} is related to the number of edges of a given network member. It is defined by Equation 5.

$$K_{c_i} = \sum_{c_j=1}^{T} E_{c_i c_j} \tag{5}$$

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. It is worth noting that the weight associated with $E_{c_i c_j}$ can be the original weight of the edge (Weighted Degree) or simply 1.0 (Unweighted Degree), which only indicates the presence of an edge. A classifier with a high degree is very divergent with its direct neighbors.

Another classic centrality measure is Betweenness [15]. This measure considers the number of shortest paths (geodesic paths) from each member of the network to all others that pass through each particular member, as denoted by Equation 6.

$$B_{c_i} = \sum_{c_j c_k} \frac{g_{c_j c_k}^{c_i}}{g_{c_j c_k}} \tag{6}$$

where $g_{c_j c_k}^{c_i}$ 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 must have only one component to calculate this centrality measure; otherwise, it will be impossible to calculate all the distances (paths) between two vertices. Each geodesic is a sub-ensemble of classifiers with high diversity. So, a classifier with high betweenness centrality appears frequently in the most diverse sub-ensembles.

Another commonly used measure that also uses the shortest paths for centrality estimation is the Closeness [13], [14]. 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 7 can be used to compute the Closeness centrality. A classifier with high closeness indicates an important contribution to the team diversity since it obtained by the average of all diversity relationships of a classifier to all others.

$$C_{c_i} = \frac{1}{l_i} = \frac{|T|}{\sum_{c_j} g_{c_i, c_j}} \tag{7}$$

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

One last classic measure is the Eigenvector centrality [16]. It is similar to Degree centrality, but beyond the number of adjacent members, it also considers the centrality of them. Therefore, a member is central if it has a relationship with others that are themselves central. Equation 8 presents the Bonacich Eigenvector.

$$\lambda x = Ax, \lambda x_i = \sum_{j=1}^{n} a_{ij} x_j, i = 1, ..., n. \tag{8}$$

where A is the adjacent matrix, λ is a constant (the eigenvalue), and x is the eigenvector. The centrality of a member is thus proportional to the sum of the centralities of its adjacent members. A classifier with high eigenvector disagrees with other classifiers that highly disagree with their direct neighbors.

A most recent measure called 'Local Centrality' is present in [17]. This centrality measure considers the nearest neighbors and their next neighbors in estimating the prominent position. The centrality L_{c_i} is computed as denoted by Equation 10.

$$Q_{c_j} = \sum_{c_w \in \Gamma_{c_j}} N_{c_w}, \tag{9}$$

$$L_{c_i} = \sum_{c_j \in \Gamma_{c_i}} Q_{c_j}, \tag{10}$$

where Γ_{c_j} is the set of nearest neighbors of the member c_j, and the nearest neighbors and their next nearest neighbors of vertex c_w is N_{c_w}. A high 'local centrality' suggest that the classifier is very divergent with its neighbors, which are themselves very divergent with their neighbors as well.

In this section, we have presented the most common centrality measures appearing in the literature. Each of them focuses on different aspects of the network to evaluate the role of its members. So, the choice of a centrality measure depends on what a network represents, and which questions the network analysis intends to answer. We conducted experiments to drive our decision for the best centrality measure for the proposed fusion method, which is present in Section IV-A.

2) Network Simplification: The simplification process, also known as pruning, can be the removal of specific edges or vertex. In our work, we use edge pruning since it helps emphasize the vertices whose relations of diversity are higher by removing lower diversity edges. Besides, we intend to maintain the network as a connected component to allow the estimation of most centrality measures. The goal is to analyze the ensemble network by the diversity perspective and estimate the key classifiers to the ensemble diversity by the use of centrality measures.

In [18], several edge pruning methods are present, with the simplest one being the Naive Algorithm. However, one drawback with this algorithm is the γ value requirement. This parameter is responsible for determining the number of edges to be removed, so, some knowledge about the network is expected to suggest an adequate value. The algorithm sort the edges in ascending order according to their weight, i.e., the pairwise diversity assessed score. Afterward, it removes the edge from the top if it is not a bridge and this removal could disconnect the graph, generating a new component. The algorithm stops pruning when the total number of edges estimated reaches $n = \gamma \times ((|E| - (|V| - 1))$, where $|E|$ and $|V|$ are the total number of edges and vertices, respectively. A

modified Naive algorithm is proposed in this paper, in which we eliminate the need for the γ parameter (see Section III-B).

III. Proposed Method

The proposed approach is described as an Ensemble composed of three distinct phases, showed in Figure 1.

A. Phase a: Pool creation, Accuracy and Diversity Estimation

In the first phase of the proposed method (Figure 1a), an initial pool of classifiers C of size T is created using a pool generation method applied in the training set S_{train}. Such a pool can be created using any of the classical methods available in the literature, such as Bagging [1], Boosting [2] and Random Subspaces [3]. Once the pool is generated, a validation set S_{val} is used to estimate the accuracy of each classifier, as well as the pairwise diversity. The accuracy is the percentage of instances that a classifier predicts correctly, while the pairwise diversity is estimated by using one of the measures as mentioned above. The pairwise estimation is normalized to fit the established range [0.1,1.0], avoiding the value 0 that can be misinterpreted in some cases; for instance, an edge with 0 weight could be interpreted as an absent edge. Another range can be adopted here since some of the centrality measures consider more important low edge weight values while others are the opposite. The Table II clarify this point, presenting the proper values to represent the weight of the edges depending on the centrality to use on the next step.

TABLE II
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	↑	↑	↓	↑

B. Phase b: Construction of the Ensemble Network

The pool of classifiers generated in the previous phase and the diversity of each classifier are used to construct a network (process A in Figure 1b), where the vertices are classifiers, and the edges between them are the score of the chosen pairwise diversity measure. The ensemble network is a complete graph at first, so it is simplified (process B in Figure 1b) using an edge pruning method. The main motivation for this pruning step is to highlight the most important relations in the ensemble, in this case, the most important relations are those with high diversity value scored by one chosen diversity measure. To that end, we propose a modified Naive 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.

Algorithm 1 differs from the original Naive Pruning Algorithm [18] in two main ways. First, 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 increases the number of components of the network, and continues pruning. Second, it does not require the γ parameter used to estimate the number of edges to be removed, which requires some knowledge about the network. The Stop criterion of the proposed algorithm maintains the graph as a single component, and thereby allows an estimation of all classic centrality measures. The resulting network H is the input for centrality estimation (process C in the Figure 1b). It should be noted that the order in which the edges are pruned must preserve higher diversity relations.

Algorithm 1: Modified Naive 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 ← E;
SortEdges(F);
i ← 1;
while i <= N do
    if C(c_r, c_s; F\e_i) ≠ −∞ then
        F ← F\e_i;
    else
        return H = (V, F);
    end
    i ← i + 1;
end
```

The centrality estimation (process C in Figure 1b) plays an essential role in our method since it is responsible for highlighting the prominent vertices (classifiers) of the ensemble. A higher centrality score means that the classifier occupies a distinct influential position in the network, i.e., it plays an essential role concerning diversity among ensemble members. It is the meaning of any centrality measure computed in an ensemble network in which the edges represents pairwise diversity relations. However, each centrality measure may rank the classifiers differently, according to its specific concept of influential position. For instance, the classifier that lies in the most of geodesic paths is the most important for betweenness centrality, while the classifier that has the most neighbors is the most important regarding degree centrality. Each centrality measure is evaluated on Section IV-A to discover the one most promising to the ensemble accuracy.

After the centrality measure is computed, each single classifier will receive a score (CE_i), as well as the previously computed accuracy Acc_i. Both measures are normalized in the range of $[0.1, 1.0]$ (process D in Figure 1b) using the min-max normalization process. As a consequence of the normalization process, both measures are expected to be equally influent to estimate the classifiers' importance to the team. This is important due to the differences in the scores suggested by the centrality measures that are in different scales. Also, it is important to normalize accuracy, e.g., the classifiers' accuracy

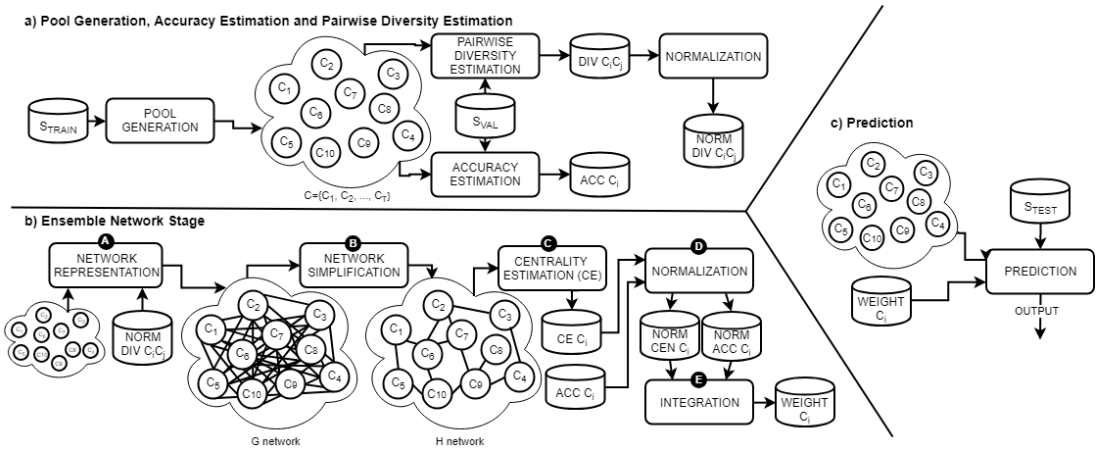

Fig. 1. A general overview of the CBF method.

for some classification problems are low, so, when combined with the normalized centrality score it makes the accuracy score less important. Then, the measures are combined to obtain the final weight (process E in Figure 1b) for each classifier i as denoted in Equation 11:

$$\psi_i = CE_i \times Acc_i \tag{11}$$

where the resulting weight ψ_i is an input of the final phase (Prediction in Figure 1c) of the proposed method. The diversity importance of a classifier and the accuracy are two keys to elevate the precision of an ensemble based classifier. The first measure how distinct is the classifier for the ensemble's generalization ability by the analysis of the pairwise diversity relationship. The accuracy, on the other hand, considers only the individual performance of the classifiers. So, two classifiers in the ensemble can perform equally well in recognition tax, but they possibly recognize different problems, in which one of the classifiers are more useful to the ensemble's generalization than the other. The centrality of classifier assessed by diversity relations as the unique parameter can perform poorly since the prominence is based only on disagreement decisions compared to others. These situations justify the importance in combining both parameters to measure the influence of the classifier in the ensemble's final prediction.

C. Phase c: Prediction

During the final phase of our method, as shown in Figure 1c, each classifier in the pool has its own weight ψ_i, which reflects its importance in the network. After the weighted sum of votes is computed for each class, the class with the highest score is the final ensemble decision.

IV. Experiments

We carried out experiments on 30 classification problems extracted from the following different Machine Learning repositories: UCI Machine Learning Repository [19], KEEL (Knowledge Extraction based on Evolutionary Learning) Repository [20], Ludmila Kuncheva Collection of Real Medical Data [21], STATLOG project [22] and artificial datasets generated with the Matlab PRTools toolbox. These datasets used are present in Table III. They present only numeric features with no missing values, and a varied number of instances, attributes, and classes.

TABLE III
Characteristics of the datasets (DS). Number of Instances (#I). Number of Features (#F). Number of Classes (#C).

#	DS	# I	# F	# C	Repository
1	Adult	690	14	2	UCI
2	Banana	2000	2	2	PRTools
3	Blood	748	4	2	UCI
4	CTG	2126	21	3	UCI
5	Diabetes	766	8	2	UCI
6	Ecoli	336	7	8	UCI
7	Faults	1941	27	7	UCI
8	German	1000	24	2	STATLOG
9	Glass	214	9	6	UCI
10	Haberman	306	3	2	UCI
11	Heart	270	13	2	STATLOG
12	ILPD	583	10	6	UCI
13	Ionosphere	350	34	2	UCI
14	Laryngeal1	213	16	2	LKC
15	Laryngeal3	353	16	3	LKC
16	Lithuanian	2000	2	2	PRTools
17	Liver	345	6	2	UCI
18	Magic	19020	10	2	KEEL
19	Mammo	830	5	2	KEEL
20	Monk	432	6	2	KEEL
21	Phoneme	5404	5	2	ELENA
22	Segmentation	2310	19	7	UCI
23	Sonar	208	60	2	UCI
24	Thyroid	692	16	2	LKC
25	Vehicle	847	18	4	STATLOG
26	Vertebral	300	6	2	UCI
27	WBC	569	30	2	UCI
28	WDVG	5000	21	3	UCI
29	Weaning	302	17	2	LKC
30	Wine	178	13	3	UCI

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} ($\cong 32,3\%$) and one ($\cong 16,7\%$) for testing S_{test}). We adopted a stratified sampling to guarantee a class distribution balance in all subsets (S_{train},

S_{val} and S_{test}). The Bagging method [1] was used to create the pool of $T = 100$ classifiers, while the Perceptron (Single Layer) was selected as the base classifier. Each bag used to train a classifier had only 66% of the training samples. This unstable base classifier and small bags were used mainly to generate a pool of weak and diverse classifiers. All classifiers have accuracy higher than 50% estimated on a validation set.

In the first set of experiments, we evaluated the use of different combinations of pairwise diversity and centrality measures. The objective was to find out the best set-up for the proposed fusion method. In the second set of experiments, we have compared the proposed method against a variety of well-known approaches found in the literature.

A. Evaluation of pairwise diversity and centrality measures

We assessed a combination of 4 pairwise diversity measures (DF, CC, Dis and QS) and 7 different centrality measures (Betweenness Weighted (W), Betweenness Unweighted (NW), Closeness, Degree Weighted (W), Degree Unweighted (NW), Eigenvector e Local Centrality).

After performing 28 experiments (4 pairwise diversity measures $\times$ 7 centrality measures) considering the 30 classification problems, we computed the Friedman and the Nemenyi post hoc test. Figure 2 shows the average rank (the value next the parameter names used on CBF method) computed in which we can see that the best set-up was obtained by using the Weighted Degree centrality (Degree W) computed in a network built using the Double Fault (DF) pairwise diversity. Degree centrality, weighted or unweighted, is much easier to compute, as compared to the other centrality measures. The DF measure, in fact, is the choice of the 5 best configurations, but the difference between them is not statistically significant.

It should be mentioned that in these experiments, the values of the QS and the CC diversity measures were sometimes set to 1.0 due to a division by zero. This division may occur when some of the possible relations between incorrect/correct samples described in Table I are not observed.

B. Comparison with state-of-art methods

The best CBF method (CBF:WD-DF) was assessed using Weighted Degree centrality computed in DF pairwise relations. This setup concerning our approach was compared to the following 9 fusion methods appearing in the literature: a) the Majority Vote (MV) also known as Plurality Vote; b) the Weighted Majority Vote by Accuracy (WMV); c) the Performance Weighting (PW) [23]; d) the Kuncheva Weighted Majority Vote (KWMV) [6]; e) the Bayesian Combination (BC) [24]; f) the Max Rule (MAR) [5]; g) the Median Rule (MER) [5]; h) the Sum Rule (SR) [5]; and i) the Product Rule (PR) [5]. Only MV do not score the classifier influence assuming that each classifier has equal influence. The others are divided into two groups: i) static weighted score of the classifier based on individual accuracy and ii) the score of the classifier is based on posterior probability. The exception is BC, which uses the individual performance of the classifier and its posterior probability to estimate the classifier influence. Comparing CBF to different groups concerning how they estimate the classifier weights possibly can lead to more reliable observations about the performance and limitations.

Table IV shows the average accuracy and corresponding standard deviation of the proposed method and the 9 fusion methods in the literature. As can be seen, the CBF method reaches the best possible result in 14 of 30 classification problems, while the best competitors show the best possible result in just 6 classification problems.

A comparison concerning the number of wins, ties, and losses is presented in Figure 3. The CBF method shows a significantly better result when compared to each literature method, except PW. The dashed line illustrates the critical value ($cv = 19.5$). The cv value was obtained from the number of experiments with a significance level $\alpha = 0.05$. In summary, when considering the whole set of experiments, the CBF won in 189 out of 270 experiments (70%), lost in 61 cases (22.59%), and tied in 20 cases (7.41%).

Figure 4 presents a statistical analysis using the Friedman and the Nemenyi post hoc test. As can be seen, our approach is statistically different from most of the approaches in the literature. A thorough analysis of Figure 4 shows that the proposed method is statistically different from MAR and PR. However, CBF provides similar results for most of the literature, 7 out of 9, according to the critical distance (CD). The lowest average rank assigned to our approach suggest that it is good option compared to literature.

In another statistical analysis, the Wilcoxon test was performed to compare CBF:WD-DF against each of the 9 approaches in the literature in a pairwise fashion. The results in Table IV show that the proposed method, at $\alpha = 0.05$ significance level, the result is statistically significant for all pairwise comparisons. PW is the literature approach with the most similar results (0.04884) which can be considered a tie.

C. Discussion

Our discussion focuses on two points: i) the comparison between the centrality measures and ii) the literature comparison. First, as observed in the experiments, the Degree centrality provided the best results compared to other centrality measures assessed regarding the ensemble accuracy. This centrality properly exploited the diversity between the classifiers represented in our network, lending more importance to classifiers with high diversity levels considering only their direct neighbors (adjacent members). A diverse classifier has a large neighborhood (unweighted degree), has strong relations with their neighborhood (weighted degree) or both in the case of weighted degree. The Betweenness measure proved to be an interesting alternative but was not better than the Degree Centrality. Even though a classifier may appear in many geodesics, as the length of the geodesics is usually short (a few classifiers), it gives too much importance to few classifiers, increasing the difference between them and leading to a reduced very influent classifiers. Eigenvector and Local centralities, which estimates the centrality calculation of each classifier based on its relationship with its neighbors

TABLE IV

AVERAGE ACCURACY AND STANDARD DEVIATION OF EACH EVALUATED APPROACH. THE BEST RESULTS ARE IN BOLD. (WS) STANDS FOR WILCOXON SIGNED TEST. THE VALUES REPRESENT THE P-VALUE AND + 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

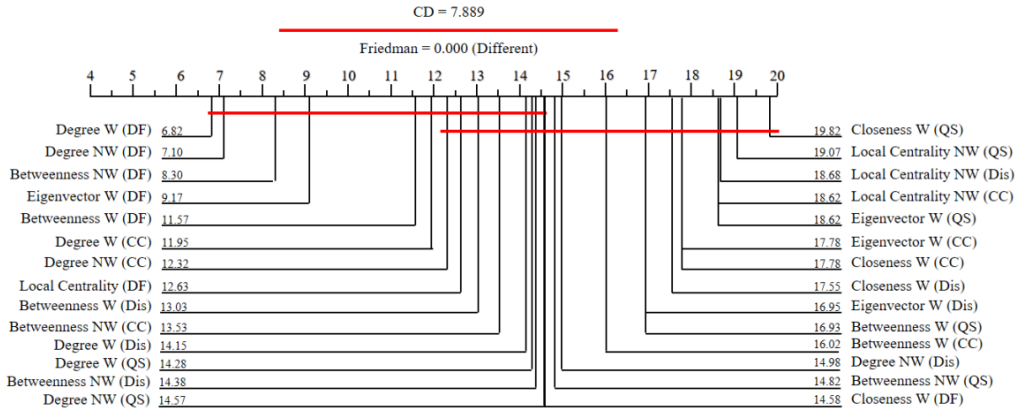

Fig. 2. Nemenyi and Friedman tests to assess pairwise diversity and centrality. The centrality is estimated using the weight of the edges (W) or not (NW).

(adjacent members) and their relationship with their respective neighbors, did not generate better results than Degree Centrality. These centrality measures are less sensitive to the weight of their direct neighbors compared to Degree. Closeness, which like Betweenness, uses geodesic paths for its estimation, did not show promising results. The DF diversity measure provided distinct contribution compared to the others, suggesting that classifiers should avoid common errors instead only be different from others.

Finally, the CBF:WD-DF is compared to the literature. It shows that CBF is usually better than the literature approaches, which ignores how important is the relation between classifiers concerning the ensemble's diversity and accuracy. The critical value (cv) suggested the existence of statistical difference between all methods (except PW) while the critical distance (CD) indicates that some approaches are statistically similar. The lowest average rank suggests that CBF:WD-DF is usually between the best approaches concerning different problems.

V. 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 CBF method, the ensemble is represented as a complex network created to reflect the diversity

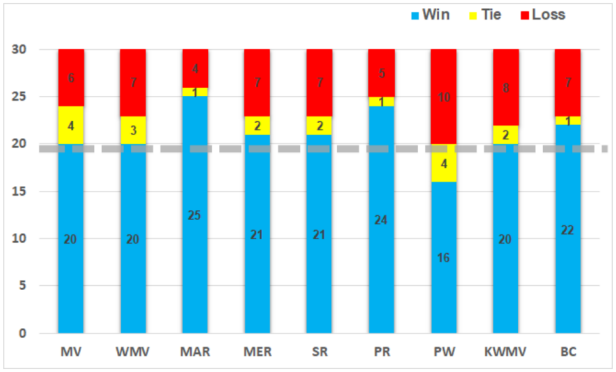

Fig. 3. Pairwise comparison of CBF:WD-DF with literature methods.

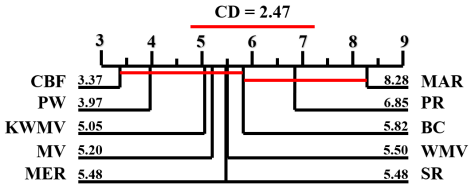

Fig. 4. Friedman and Nemenyi post hoc tests comparing the approaches.

between the classifiers. The importance of each classifier in that network is estimated employing centrality measures, which combined with accuracy, provide the weight used in the fusion process.

The experimental results on 30 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 diversity and centrality measures were evaluated to find out the best set-up for the proposed method. The best results were achieved by using the Double Fault pairwise diversity measure to generate the ensemble network and the Weighted 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 14 of 30 classification problems when compared to 9 different fusion methods in the literature, while the second best method in that comparison presented the best accuracy in just 6 cases. Among a total of 270 comparisons the proposed method was able to prevail in 189 out of 270 experiments (70%), and lost in 61 cases (22.59%).

Further work is necessary to investigate the behavior of the proposed method when using classifiers created with different ensemble learning techniques and different base classifiers.

Acknowledgement

This research has been supported by the Brazilian National Council for Scientific and Technological Development (CNPq) and by the Coordination for the Improvement of Higher Education Personnel (CAPES).

References

[1] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[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*, vol. 26, no. 5, pp. 1651–1686, 1998.

[3] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[4] A. d. S. Britto Jr., R. Sabourin, and L. E. S. Oliveira, "Dynamic selection of classifiers - A comprehensive review," *Pattern Recognition*, vol. 47, no. 11, pp. 3665–3680, 2014.

[5] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[6] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms: Second Edition*. John Wiley & Sons, 2014.

[7] R. M. O. Cruz, R. Sabourin, and G. D. C. Cavalcanti, "Dynamic classifier selection: Recent advances and perspectives," *Information Fusion*, vol. 41, pp. 195–216, 2018.

[8] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[9] G. U. Yule, "On the Association of Attributes in Statistics: With Illustrations from the Material of the Childhood Society," *Philosophical Transactions of the Royal Society of London. 194(A):*, vol. A, no. 194, pp. 257–319, 1990.

[10] A. Seath and R. Sokal, "Numerical Taxonomy. The Principles and Practice of Numerical Classification," *Systematic Zoology*, vol. 24, no. 2, pp. 263–268, 1973.

[11] D. Skalak, "The sources of increased accuracy for two proposed boosting algorithms," *In Proc. American Association for Arti Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop*, pp. 120–125, 1996.

[12] G. Giacinto, "Design of effective neural network ensembles for image classification purposes," *Image and Vision Computing*, vol. 19, pp. 699–707, 2001.

[13] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[14] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[15] J. M. Anthonisse, "The rush in a directed graph," *Stichting Mathematisch Centrum. Mathematische Besliskunde*, vol. BN 9/71, p. 10, 1971.

[16] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology*, vol. 2, pp. 113–120, 1972.

[17] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 4, pp. 1777–1787, 2012.

[18] F. Zhou, S. Mahler, and H. Toivonen, "Simplification of networks by edge pruning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7250, no. Icdm, pp. 179–198, 2012.

[19] M. Lichman, "{UCI} Machine Learning Repository," Irvine, CA, 2013. Available: http://archive.ics.uci.edu/ml

[20] 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*, vol. 17, no. 2-3, pp. 255–287, 2011.

[21] L. Kuncheva, "Ludmila Kuncheva Collection LKC," p. Available: http://pages.bangor.ac.uk/mas00a/activi, 2004.

[22] R. D. King, C. Feng, and A. Sutherland, "Statlog: Comparison of classification algorithms on large real-world problems," *Applied Artificial Intelligence*, vol. 9, no. 3, pp. 289–333, 1995.

[23] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010.

[24] W. L. Buntine, "A Theory Of Learning Classification Rules," Ph.D. dissertation, University of Technology, Sydney, 1992.