Enhancing Difficult Classes in One-vs-One Classifier Fusion Strategy using Restricted Equivalence Functions

Mikel Galar, Edurne Barrenechea
Dept. of Automática y Computación
Universidad Pública de Navarra
Campus Arrosadía, Pamplona, Spain
Email: mikel.galar@unavarra.es,
edurne.barrenchea@unavarra.es

Alberto Fernández
Dept. of Computer Science
University of Jaén, Jaén, Spain
Email: alberto.fernandez@ujaen.es

Francisco Herrera
Dept. of Computer Science and
Artificial Intelligence
University of Granada, Granada, Spain
Email: herrera@decsai.ugr.es

***Abstract*—One-vs-One is a commonly used decomposition strategy to overcome multi-class problems, even when the base classifier supports directly addressing the multi-class problem. This paper analyzes the fact that, in this strategy, less attention is given to the difficult classes, favoring the easier ones. Different evaluation criteria are used, and a novel fusion strategy, which generalizes the weighted voting, is presented to enhance the difficult classes classification. The new methodology is able to increase the recognition of the difficult classes, thus obtaining a more balanced performance over all classes, which is a desirable behavior.**

I. Introduction

Decomposition strategies [1] are often used to overcome multi-class problems. Error Correcting Output Codes (ECOC) [2] framework comprises most of these techniques. Among them, One-vs-One (OVO), which divides the original problem in as many pairs of classes as possible, is a commonly used strategy. The new binary subproblems are faced by independent base classifiers, whose outputs are then combined in order to obtain the final class label for a given instance [3], [4].

In other respects, the characteristics of each class within a problem are usually different, e.g., the number of instances, the inter-class relations and the overlapping with other classes, may vary. As a consequence, some of the classes might be more difficult to distinguish than others. *Difficult classes* can be considered those obtaining a lower classification rate; that is, the number of correctly classified examples from the class divided by the total number of examples from that class (True Positive Rate, TPR).

This contribution focuses on those problems where all classes are equally important, that is, their recognition rate must be as similar and high as possible. Evaluating a balanced data-set with accuracy rate, all the classes have *a priori* the same importance. Nonetheless, it does not reflect the difficult classes problem, since it averages the results over all instances, without taking into account the accuracy over each class independently. As a consequence, difficult classes are present in a data-set, it is usually easier to increase the accuracy rate by improving the classification of the easiest classes, whereas some of the instances from the difficult ones are misclassified.

We intend to explain why OVO strategy weakens when we aim to achieve a good prediction for all the classes in the problem. Besides, we introduce a new aggregation model based on Restricted Equivalence Functions (REFs) [5], which allows one to modify the decision boundaries of the base classifiers to boost the classification of the difficult classes, without changing the underlying base classifiers. To do so, this aggregation performs an optimization stage using the CHC genetic algorithm (GA) [6] to learn the appropriate set of parameters to enhance the difficult classes (while maintaining the global accuracy rate).

The experiments carried out include a set of twenty-eight real-world problems from UCI [7] and the KEEL data-set repository [8]. In addition to the usage of the accuracy rate to evaluate the performance of the classifiers, we include other measures accounting for the problem of difficult classes. The comparisons among the results obtained are contrasted using the proper statistical tests [9], [10]. In order to analyze the capabilities of the new aggregation, we consider Support Vector Machines (SVMs) [11] as base classifiers.

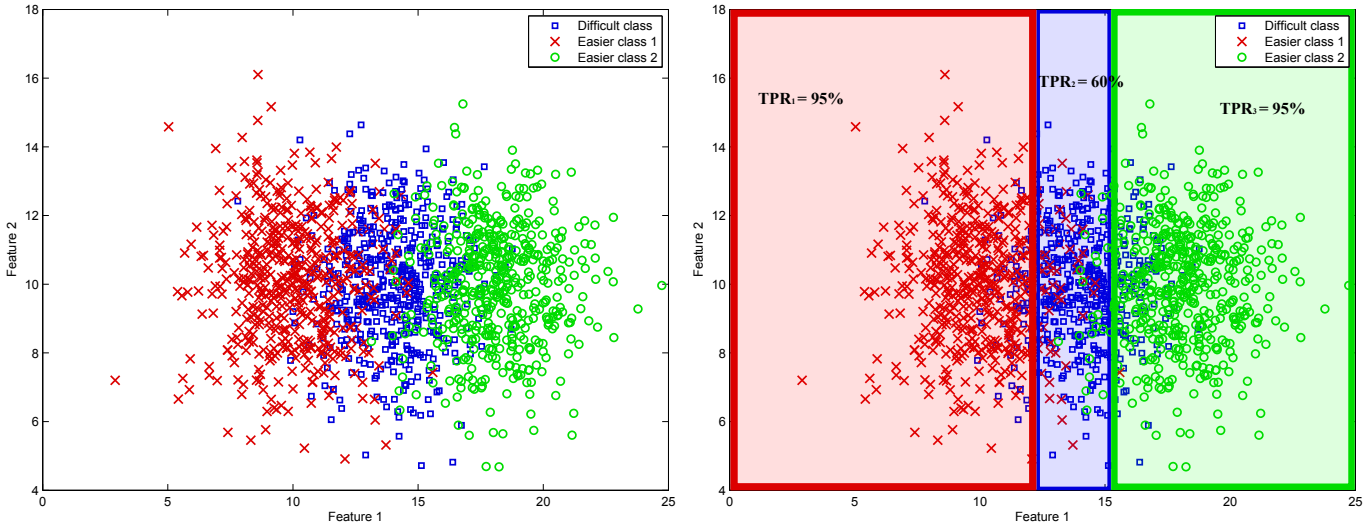
The remainder of this paper is as follows. In Section II, the problem of difficult classes in OVO strategy is analyzed. Next, Section III shows our proposal to enhance the difficult classes. The tuning of the parameters is presented in Section IV. The set-up of the experimental framework is explained in Section V. In Section VI, the experimental analysis is carried out. Finally, Section VII concludes the paper.

II. Difficult Classes Problem in One-vs-One Strategy

This section recalls the basis of OVO strategy and its simplest aggregation (Subsection II-A), which are then used to explain the difficult classes problem in general (Subsection II-B), and more specifically in OVO scheme (Subsection II-C).

A. One-vs-One decomposition

OVO divides a m-class problem into $m(m-1)/2$ independent binary subproblems considering all the possible pairs of classes, which are faced by independent base learners. In order to classify a new instance, it is presented to all the



(a) A three class problem with one difficult class.

(b) Classifier maximizing accuracy rate, even though a low TPR is achieved for the difficult class.

Fig. 1. An example of the difficult classes problem. The class in the center is more difficult to be correctly classified due to its overlapping with the other two classes.

base classifiers. Each classifier distinguishing between a pair of classes $\{C_i, C_j\}$ outputs a confidence degree $r_{ij} \in [0, 1]$ in favor of C_i; thus, the confidence in favor of C_j is computed as $r_{ji} = 1 - r_{ij}$. All the confidence degrees can be organized within a score-matrix:

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

Different aggregations have been presented in the literature to obtain the final output [3]. The simplest aggregation, yet powerful is the voting strategy, where each classifier votes for its predicted class, and the class obtaining the largest number of votes is predicted.

B. Difficult Classes Problem

In a classification problem, the degree of separability of the classes usually vary owing to their different characteristics. The simplest way of showing the difficult classes problem is by the usage of the illustrative example in Figure 1(a). It can be observed that one of the classes is more difficult to distinguish than the other two because of the class overlapping. A classifier aiming at maximizing accuracy rate would define the three regions in Figure 1(b), which shows up the problem we are dealing with in this work.

In this problem, the difficult class have obtained a TPR of 60%, whereas the easier ones have achieved a high TPR (TPR = 95%). In case of the difficult class being at least as important as the rest ones, it would be better to obtain a balanced classification, i.e., a TPR = 83.33% (homogeneous) for all the classes, which would produce the same global accuracy rate. This situation highly differs from the real one and could be more recommendable in many problems requiring an equal recognition of all classes [12], [13]. For this reason, one can observe that the most commonly used metric to assess the performance of classifiers may not properly reflect the problem, and this is why we need to consider other measures [14]. Recall that the accuracy rate is computed as

$$\text{accuracy} = \frac{1}{n_T} \sum_{i=1}^{m} \text{TPR}_i \cdot n_i, \tag{2}$$

where n_i is the number of examples of class i and n_T is the total number of examples evaluated. In fact, the accuracy rate is the weighted mean of the TPRs over each class, where the weights are given by the proportion of examples from each class, which makes it inadequate to evaluate problems with difficult classes. Therefore, we need measures considering the TPR over each class, but they must not take into account the number of examples. The following two measures fulfill the mentioned characteristics:

- The Average Accuracy rate (AvgAcc) [15],

$$\text{AvgAcc} = \frac{1}{m} \sum_{i=1}^{m} \text{TPR}_i. \tag{3}$$

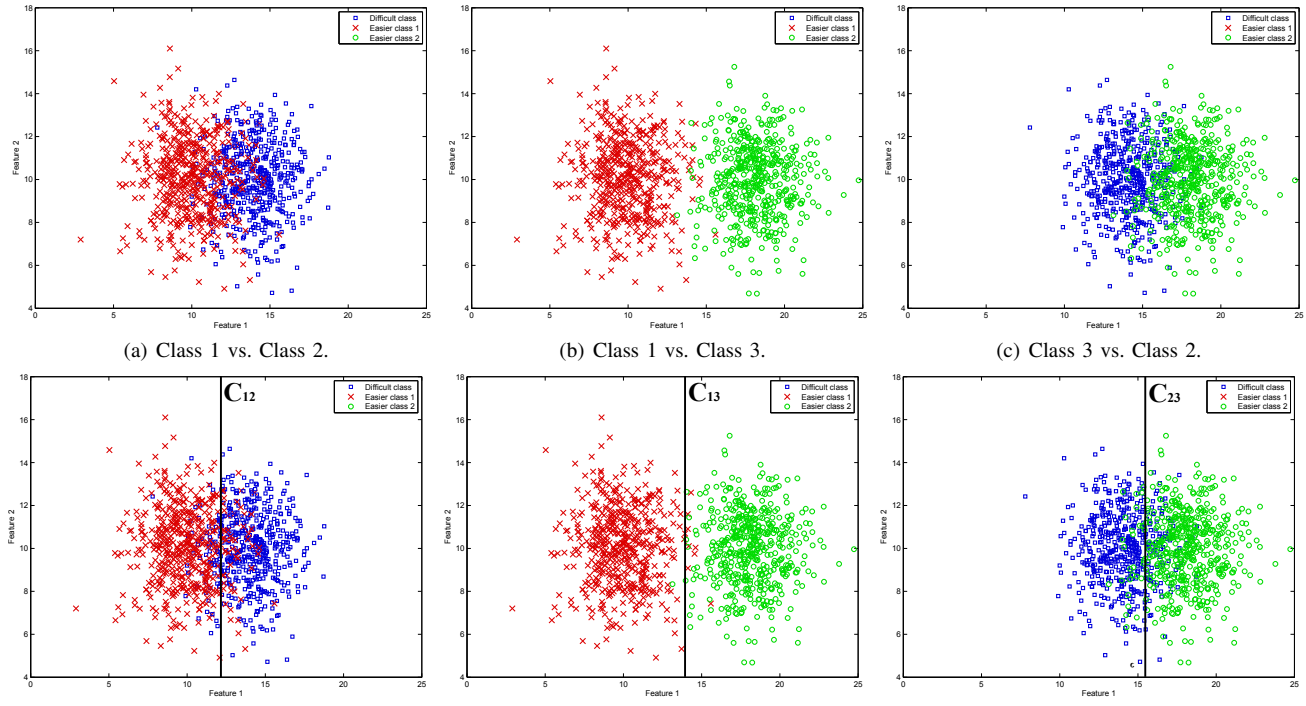
- The Geometric Mean (GM) [16],

$$\text{GM} = \sqrt[m]{\prod_{i=1}^{m} \text{TPR}_i}. \tag{4}$$

The problem with the AvgAcc is that a low rate on one class can be overlooked, partially accounting for the problem explained. Otherwise, the GM strongly penalizes those solutions achieving low a TPR in any of the classes. Along this work, we will show that whereas the GM properly models the difficult classes problem, the AvgAcc could only serve as a complementary measure. Table I represents all these facts, showing the values that would be obtained in each performance measures with two different classification scenarios.

C. A Weakness of One-vs-One Strategy

In the case of OVO strategy, the difficult classes problem is accentuated, as we will show following the previous example.



(a) Class 1 vs. Class 2. (b) Class 1 vs. Class 3. (c) Class 3 vs. Class 2.

(d) Classifier 1 vs. 2. (e) Classifier 1 vs. 3. (f) Classifier 2 vs. 3.

Fig. 2. OVO decomposition of the problem in Figure 1(a) and the base classifiers learned for this decomposition.

TABLE I. BEHAVIOR OF DIFFERENT PERFORMANCE MEASURES.

Classifier	TPR_1	TPR_2	TPR_3	accuracy	AvgAcc	GM
Heterogeneous	0.95	0.6	0.95	83.33%	83.33%	0.8151
Homogeneous	0.8333	0.8333	0.8333	83.33%	83.33%	0.8333

Homogeneous classifier refers to that obtaining the same TPR for all classes, whereas Heterogeneous refers to that obtaining a different TPR for each class.

The problem in OVO is that even though the binary classifiers are optimal (in terms of accuracy, AvgAcc, GM and balance between the TPRs over both classes in the subproblem), the resulting combination need not be globally optimal (in terms of all the evaluation measures). The problem in Figure 1(a) is decomposed into three subproblems (Figure 2(a), 2(b), 2(c)), face by independent base classifiers shown in Figures 2(d), 2(e), 2(f). The TPRs obtained in each base classifier can be considered to be equal (in the same classifier), and hence balanced for both classes, i.e., locally optimal in terms of accuracy, AvgAcc, GM. Nevertheless, their combination using the voting strategy would lead to the class separation in Figure 1(b), which is not optimal in terms of GM due to the low TPR over the difficult class, although it is optimal in terms of accuracy.

For this reason, we study the low TPR achieved over the difficult classes in OVO strategy, or rather the non-existent improvement over those classes. Hereafter, we aim to show why OVO strategy tends to improve the accuracy over the easiest classes, without enhancing the classification over the most difficult ones. In order to do so, we consider the simplest scenario: OVO scheme with the voting strategy. Recall that, the True Positive Rate (TPR) of a class is the number of correctly classified examples from the class divided by the total number of examples from that class.

Problem statement and notation.

- m-class problem, $\mathbb{C} = \{C_1, \ldots, C_m\}$.
- There are m_d classes which are much more difficult to classify (for example, due to overlapping, noise, or even imbalance).
- The rest of the classes are easier to be classified.
- Let TPR^i_{ij} be the TPR over class C_i of the classifier distinguishing classes $\{C_i, C_j\}$.

Problem assumptions.

1) Independence of the base classifiers, which is supposed in OVO scheme.
2) An instance is correctly classified if all the competent base classifiers [4] (those considering the real class of the instance in the training phase) correctly classify the instance.
3) Given a difficult class (C_i) and an easier class (C_j) then, $\mathrm{TPR}^i_{ik} \leq \mathrm{TPR}^j_{jt}$ for all $k, t = 1, \ldots, m$, $k \neq i$, $t \neq j$ and there exist $p, q \in \{1, \ldots, m\}$, $p \neq i$, $q \neq j$ such that $\mathrm{TPR}^i_{ip} < \mathrm{TPR}^j_{jq}$.

The first assumption suppose that the outputs given by the classifiers are unrelated as it is assumed in OVO strategy. The second one might be an over-simplification because it is possible to correctly classify an instance even though one of the competent classifier fails. Nonetheless, a total agreement between base classifiers can be required in systems needing a high confidence in the decision, which help us showing the difficult classes problem in OVO. The last assumption suppose that the TPRs obtained in the base classifiers considering a difficult class are always lower or equal than the corresponding ones over the easier classes.

Problem description.

Given an instance $\{\mathbf{x}, y\}$ (where $\mathbf{x}$ are the values of the input attributes and $y = C_i$ with $i \in \{1, \ldots, m\}$), the probability of being correctly classified, denoted as $P(h_{ovo}(\mathbf{x}) = y)$ (where h_{ovo} stands for the OVO classifier), following assumptions 1 and 2, is given by the TPR of each one of the base classifiers that considered instances from this class to be trained (that is, TPR_{ij}^i for all $j = 1, \ldots, m$ with $i \neq j$):

$$P(h_{ovo}(\mathbf{x}) = y) = \prod_{1 \leq j \neq i \leq m} \mathrm{TPR}_{ij}^i \,. \tag{5}$$

Therefore, we consider an instance $\{\mathbf{x}_1, y_1\}$ (belonging to one of the easier classes, i.e., $y_1 = C_i$) and an instance $\{\mathbf{x}_2, y_2\}$ (belonging to one of the difficult classes, i.e., $y_2 = C_j$) to be classified, whose probabilities of being correctly classified are given by Eq. (5). Following assumption 3, we have that

$$P(h_{ovo}(\mathbf{x}_1) = y_1) = \prod_{1 \leq k \neq i \leq m} \mathrm{TPR}_{ik}^i > \prod_{1 \leq t \neq j \leq m} \mathrm{TPR}_{jt}^j = P(h_{ovo}(\mathbf{x}_2) = y_2), \tag{6}$$

which shows that the probability of correctly classifying the instance from the difficult class will always be lower than that of correctly classifying the instance from the easier class because of the differences in the TPRs of the base classifiers.

How can this problem be solved or at least alleviated?

1) Improving the TPR_{ij}^i for each difficult class i ($j = 1, \ldots, m$, $j \neq i$).
2) Developing aggregations accounting for the difficult classes problem, avoiding the modification of the underlying base classifiers.

The former solution is the straightforward one, but it is rather difficult to carry out. Besides, following the example in Figure 1, we have shown that even though all the base classifiers obtain balanced TPRs for the classes considered, it does not imply that the difficult classes problem would disappear. One way to solve this problem might be to consider the biasing of the base classifiers towards the difficult classes, but in this case the problem is that they are not known a priori. Moreover, in base classifiers considering two difficult classes, the biasing would even be more difficult. On this account, we focus on the latter solution, which can also be combined with the first one. We do not alter the base classifiers, but combine them differently. This approach have the advantage of being independent of the base classifier considered.

We propose to modify the classification of the instances by a flexible aggregation, whose parameters are obtained from the results obtained in the training set. Hence, it could be shown as a post-processing method, where the votes of the classifiers are adapted to the difficulty of each class. In this way, we perform a global optimization with all classifiers at the same time, which is not considered by previous OVO combinations. As a consequence, we are able to obtain globally better solutions (achieving more balanced classifications).

We believe that the score-matrices contain enough information as to obtain significantly different results over the difficult classes only changing the aggregation and properly setting its parameters. For this reason, the score-matrices used in the experiments of this paper are exactly the same for all aggregations, and we aim to learn and deduce the errors committed by each classifier so we can adjust the aggregation to empower the classification of the difficult classes. Therefore, all the differences shown are only due to the aggregation, which is of great importance in order to evaluate the performance of the proposed methodology appropriately.

III. A REF-BASED AGGREGATION

In this section, the new aggregation method for OVO scheme is introduced. First, several preliminary concepts are recalled in Subsection III-A, which are needed to present the new aggregation in Subsection III-B.

A. Restricted Equivalence Functions

In order to introduce the aggregation method, we need to recall several concepts. A negation models the concept of opposite:

Definition 1. *A mapping $n : [0,1] \to [0,1]$ with $n(0) = 1$, $n(1) = 0$, strictly decreasing, and continuous is called strict negation. Moreover, if n is involutive, i.e., if $n(n(a)) = a$ for all $a \in [0,1]$, then n is called a strong negation.*

Restricted Equivalence Functions [5] measure the degree of proximity (equivalence) between two points.

Definition 2. *[5], [17] A function $\mathrm{REF} : [0,1]^2 \to [0,1]$ is called restricted equivalence function associated with the strong negation n, if it satisfies the following conditions*

1) $\mathrm{REF}(a,b) = \mathrm{REF}(b,a)$ *for all* $a, b \in [0,1]$;
2) $\mathrm{REF}(a,b) = 1$ *if and only if* $a = b$;
3) $\mathrm{REF}(a,b) = 0$ *if and only if* $a = 1$ *and* $b = 0$ *or* $a = 0$ *and* $b = 1$;
4) $\mathrm{REF}(a,b) = \mathrm{REF}(n(a), n(b))$ *for all* $a, b \in [0,1]$;
5) *For all* $a, b, c \in [0,1]$, *if* $a \leq b \leq c$, *then* $\mathrm{REF}(a,b) \geq \mathrm{REF}(a,c)$ *and* $\mathrm{REF}(b,c) \geq \mathrm{REF}(a,c)$.

In this work, the interest of this closeness measure resides in the possibility of its parametrization by means of automorphisms as follows.

Definition 3. *A continuous, strictly increasing function $\varphi : [a,b] \to [a,b]$ such that $\varphi(a) = a$ and $\varphi(b) = b$ is called automorphism of the interval $[a,b] \subset \mathbb{R}$.*

Proposition 1. *[5] Let φ_1, φ_2 be two automorphisms of the interval $[0,1]$. Then*

$$\mathrm{REF}(a,b) = \varphi_1^{-1}(1 - |\varphi_2(a) - \varphi_2(b)|)$$

is a restricted equivalence function associated with the strong negation $n(a) = \varphi_2^{-1}(1 - \varphi_2(a))$.

Automorphisms can be easily constructed using a parameter $\lambda \in (0, \infty)$: $\varphi(a) = a^\lambda$, and hence, $\varphi^{-1}(a) = a^{1/\lambda}$.

B. Generalizing the Weighted Voting Method

Our REF-based aggregation is a generalization of the well-known Weighted Voting strategy (WV), whose robustness has

been both theoretically and empirically proved [18]. In WV, the confidences of the base classifiers are used as weights to vote for the classes, giving the class with the largest total confidence as final output:

$$\text{Class} = \arg \max_{i=1,\dots,m} \sum_{1 \le j \ne i \le m} r_{ij}. \quad (7)$$

In our aggregation model, instead of directly adding up the confidences of the classifiers in each row, we first compare these confidences to the certain vote (i.e., 1.0), since it is the case in which the highest vote should be given. Therefore, the more similar r_{ij} to 1.0 is, the more importance the vote has. Both values are compared using a REF. Then, instead of voting using r_{ij}, we consider the vote given by $\text{REF}(r_{ij}, 1)$, indicating how close is r_{ij} from the certain vote. Recalling Proposition 1, the operations and parameters needed for the comparison can be reduced:

$$\text{REF}(a, 1) = (1 - |a^{\lambda_2} - 1^{\lambda_2}|)^{1/\lambda_1} = (a)^{\lambda_2/\lambda_1} = a^{\lambda} \quad (8)$$

Both parameters (λ_1, λ_2) are encoded in a single equivalent one (λ). In Figure 3, the influence of λ in the REF's application to the comparison of a certain value to 1 is plotted. Observe that $\lambda = 1$ does not modify the vote of the classifier, since $\text{REF}(r_{ij}, 1) = r_{ij}$, whereas values below one $(\lambda < 1)$ empowers the weights $(\text{REF}(r_{ij}, 1) > r_{ij})$ and the contrary occurs with $\lambda > 1$ $(\text{REF}(r_{ij}, 1) < r_{ij})$.

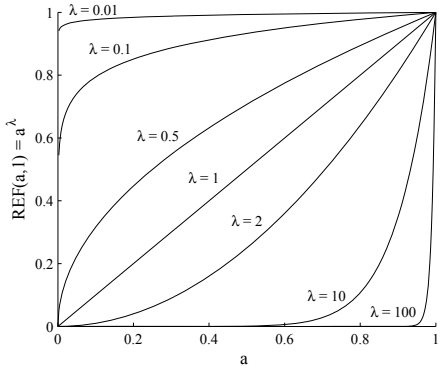


Fig. 3. Influence of the parameter λ in $\text{REF}(a, 1)$.

Remark 1. *Hence, looking at Figure 3, it can be observed that the confidences in favor of the difficult classes should use a REF with a low λ, whereas those of the easier classes might consider a higher value of λ in order to seek for a balance between their predictions. The estimation of these parameters is not trivial; for this reason, Section IV is devoted to their global adjustment.*

We have shown how the confidences given by the base classifiers can be altered; then, similarly to the WV, the votes are aggregated in each row:

$$\begin{aligned} \text{Class} &= \arg \max_{i=1,\dots,m} \sum_{1 \le j \ne i \le m} \text{REF}_{ij}(r_{ij}, 1) \\ &= \arg \max_{i=1,\dots,m} \sum_{1 \le j \ne i \le m} (r_{ij})^{\lambda_{ij}} \end{aligned} \quad (9)$$

where λ_{ij} is the corresponding parameter used in $\text{REF}_{ij}(r_{ij}, 1)$. WV method is recovered when $\lambda_{ij} = 1$ for all $i, j = 1, \dots, m$ and $i \ne j$ (see Proposition 2 in [5]). Since each r_{ij} is directly related with the corresponding r_{ji}, we will consider a single parameter for each base classifier (i.e., the degrees of freedom in the score-matrix). Therefore, $\lambda_{ji} = \frac{1}{\lambda_{ij}}$ for all $i, j = 1, \dots, m$ and $i < j$ (see Section IV).

IV. Adjusting the Aggregation to Enhance the Difficult Classes

The tuning of the parameters is needed in order to adapt the aggregation to each class within each problem. To do so, any optimization algorithm that maximizes the objective function could be used; in our case, we consider a GA, and more specifically, the real-coded CHC algorithm [6], since it has been successfully applied to similarly tune the parameters of fuzzy rule based systems [19]. We should note that this type of algorithm is needed because a complex global tuning is performed. This adjustment is required due to the fact that the locally doing it does not ensures a global improvement

This section is organized as follows: the fitness function designed to empower the difficult classes is presented in Subsection IV-A, whereas the CHC algorithm and the codification of the real parameters are described in Subsection IV-B.

A. Objective Function

The key factor of the parameter adjustment is the objective function. Notice that the standard accuracy rate must not be further optimized, as it is usually done, since it does not account for difficult classes. As a consequence, different classifier evaluation criteria are used to determine the quality of the system obtained with a given set of parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_{\frac{m(m-1)}{2}})$ (a parameter is used for each base classifier, according to the degrees of freedom in the score-matrix). The usage of the following fitness function is proposed:

$$\text{Fitness}(\boldsymbol{\lambda}) = \text{Margin}(\boldsymbol{\lambda}) + \frac{\text{GM}(\boldsymbol{\lambda}) + \text{AvgAcc}(\boldsymbol{\lambda})}{2}, \quad (10)$$

where the Margin quantifies how well is the real class of an instance separated from the second class with the highest sum of votes. The global margin is computed as follows.

$$\text{Margin} = \arg \min_{c=1,\dots,n_c} \frac{V_i^c - V_j^c}{n_{Tr} \cdot m} \quad (11)$$

where C_i is the predicted class and C_j is the second class with the largest value in Eq. (9); V_i^c, V_j^c are the values obtained in Eq. (9) by each class (for instance c), respectively. n_c is the number of correctly classified instances, only these instances are used. The margin is normalized by the number of classes and instances (n_{Tr}) to reduce its influence in the fitness function with respect to the other factors. Among all the margins computed, we take the minimum one, since it is the value better representing how well separated are the most difficult classes.

The most important part of the fitness function, and our main objective, is the GM, since it is the measure which better balances the accuracy over all classes. Nevertheless, the other factors are needed according to the following facts:

1) AvgAcc has *a priori* the same weight, but despite its value is higher, its variations depending upon the correctly classified instances are generally lower,

and hence, it has less influence when comparing different evaluations of the fitness function. It is a very important factor in cases where GM value is low, since it serves as a guide for the GA.

2) Margin has a very low weight in the fitness function (due to its normalization), mainly serving as a stabilization process once the best GM and AvgAcc combination has been found.

B. CHC Algorithm and Parameters' Representation

The real-coded CHC (Cross generational elitist selection, Heterogeneous recombination and Cataclysmic mutation) algorithm [6] was selected to optimize the fitness function (Eq. (10)) due to its successful application in similar tuning approaches [19]. It holds a good trade-off between exploration and exploitation, being a proper metaheuristic for complex search spaces.

In this elitist GA, all the M (population size) parents and their offspring are put together and the M best individuals form the next population. Instead of using a mutation operator as most of the GAs do, an incest prevention mechanism combined with a reinitialization of the population is used to increase diversity. The components needed to design the whole process are: representation of the solutions, initialization of the initial population, crossover operator, incest prevention and restarting mechanism.

1) *Representation of the parameters:* the set of parameters ($\boldsymbol{\lambda}$ of length $m(m-1)/2$) to be optimized are real parameters, so they are the elements (called genes) of a chromosome. Recall that the value of λ ranges from 0 to ∞, which cannot be directly encoded within a chromosome. Therefore, we use the following chromosome ($\Phi(\boldsymbol{\lambda})$) to encode $\boldsymbol{\lambda}$:

$$\begin{aligned}\Phi(\boldsymbol{\lambda}) = &\quad (\phi(\lambda_1), \phi(\lambda_2), \ldots, \phi(\lambda_{\frac{m(m-1)}{2}})) \\ = &\quad (c_{\lambda_1}, c_{\lambda_2}, \ldots, c_{\lambda_{\frac{m(m-1)}{2}}})\end{aligned}$$

where each gene $c_{\lambda_i} \in (0,1)$, $i = 1, \ldots, \frac{m(m-1)}{2}$ and the parameter's value is recovered as follows

$$\lambda_i = \phi^{-1}(c_{\lambda_i}) = \begin{cases} (2 \cdot c_{\lambda_i})^2 & \text{if } c_{\lambda_i} \leq 0.5 \\ \dfrac{1}{(2 \cdot (1 - c_{\lambda_i}))^2} & \text{otherwise.} \end{cases}$$

In this manner, the whole search space can be explored (Figure 3). The square allows us to homogeneously search the whole space, since, as it can be observed in the figure, the nearer λ is to the upper or the lower bounds, the greater change it needs to significantly alter the output of $\mathrm{REF}(a, 1)$.

2) *Initialization:* All the chromosomes are randomly initialized in $(0, 1)$ except for the first one, which is initialized with 0.5 in all each genes. In this manner, the search is started with an individual representing the original WV, i.e., the proposed aggregation with $\boldsymbol{\lambda} = \mathbf{1}$ (following Eq. (12)).

3) *Crossover operator:* We use the Parent Centric BLX operator [20].

The incest prevention and restarting mechanisms are performed as usual [6], [19]. There are two criteria to end the optimization process: the maximum number of evaluations and the number of restarting procedures without improvements (their set-up is shown in Subsection V-A).

V. Experimental framework

A. Base classifiers and parameters

We consider SVMs [11] as base classifiers to study the validity of the new aggregation methodology. The confidences used in the score-matrices are obtained from the probability estimates given by the SVM logistic model [21]. The configuration parameters considered are shown in Table II, along with the parameters used in the CHC algorithm. These values are common for all problems, which is the default parameters' setting included in KEEL software [22] used to develop the experiments. We considered two configurations, varying the parameter C and the kernel function to study the behavior of the aggregation with different set-ups, which address for the robustness of the proposal. We treat nominal attributes in SVM as scalars to fit the data into the systems using a polynomial kernel.

TABLE II. Parameter specification for the base learners and the CHC algorithm employed in the experimentation.

Algorithm	*Parameters*
SVM_{Poly}	C = 1.0, Tolerance = 0.001, Epsilon = 1.0E-12 Kernel = Polynomial, Polynomial Degree = 1
SVM_{Puk}	C = 100.0, Tolerance = 0.001, Epsilon = 1.0E-12 Kernel = Puk, PukKernel ω = 1.0, PukKernel σ = 1.0
CHC	Population size = 50 individuals, Evaluations = $1000 \cdot m^2$ BITSGENE = 30 Restarting procedures without improvement = 3

Tuning the parameters of each method on each particular problem could lead to better results. However, we are not comparing base classifiers among them; hence, our hypothesis is that the methods wining on average on all problems would also perform better if a more optimal setting would be performed. Moreover, in a framework where no method is tuned, the best methods tend to correspond to the most robust ones, which is also a desirable characteristic.

We consider the probability estimates method by Wu *et al.* [23] (PE) as an aggregation for the comparison against our methodology, since its usage is widely extended, and it has been proved to be an accurate aggregation [3].

B. Data-sets and classifiers' evaluation

We have used twenty-eight data-sets from UCI [7] and KEEL data-set repository [8]. Table III summarizes the properties of these data-sets. They comprise a number of situations, from totally balanced data-sets to highly imbalanced ones, besides the different number of classes. Some of the largest data-sets (nursery, page-blocks, penbased, satimage, shuttle and led7digit) were stratified sampled at 10% in order to reduce the computational time required for training. In the case of missing values (autos, cleveland and dermatology), we removed those instances from the data-set before doing the partitions. As we have previously stated, we consider the accuracy rate, GM and AvgAcc to evaluate the performance of the classifiers, which were estimated by means of a 5-fold

cross-validation. The data partitions used in this paper can be found in KEEL-dataset repository [8] and in the website associated with [3] (http://sci2s.ugr.es/ovo-ova/).

TABLE III. SUMMARY DESCRIPTION OF DATA-SETS.

Data-set	**#Ex.**	**#Atts.**	**#Num.**	**#Nom.**	**#Cl.**
Balance	625	4	4	0	3
Contraceptive	1473	9	9	0	3
Hayes-roth	132	4	4	0	3
Iris	150	4	4	0	3
NewThyroid	215	5	5	0	3
Splice	319	60	0	60	3
Tae	151	5	5	0	3
Thyroid	720	21	21	0	3
Wine	178	13	13	0	3
Car	1728	6	0	6	4
Lymphography	148	18	3	15	4
Vehicle	846	18	18	0	4
Cleveland	297	13	13	0	5
Nursery	1296	8	0	8	5
Page-blocks	548	10	10	0	5
Shuttle	2175	9	9	0	5
Autos	159	25	15	10	6
Dermatology	358	34	1	33	6
Flare	1066	11	0	11	6
Glass	214	9	9	0	7
Satimage	643	36	36	0	7
Segment	2310	19	19	0	7
Zoo	101	16	0	16	7
Ecoli	336	7	7	0	8
Led7digit	500	7	0	7	10
Penbased	1100	16	16	0	10
Yeast	1484	8	8	0	10
Vowel	990	13	13	0	11

In order to carry out the comparison of the classifiers appropriately, non-parametric tests should be considered, according to the recommendations made in [9], [10]. In this contribution, we consider the Wilcoxon paired signed-rank test [24] as a non-parametric statistical procedure to perform comparisons between two algorithms. Any interested reader can find additional information on the thematic website http://sci2s.ugr.es/sicidm/, where software for the application of the statistical tests is provided.

VI. EXPERIMENTAL STUDY

We aim to demonstrate the validity of our aggregation proposal based on REFs (from this point denoted as RA) to enhance the classification of difficult classes in OVO strategy. To do so, we consider two different configurations of SVMs as explained in Subsection V-A. The results obtained with SVM_{Poly} as base classifier are shown in Table IV (note that, accuracy and AvgAcc are presented as percentages, as usual). It can be observed that accuracy has been maintained, whereas GM and AvgAcc have been highly enhanced using RA, being GM improvement remarkable. Anyway, in order to extract meaningful conclusions, these facts must be contrasted with the proper statistical analysis via Wilcoxon tests, whose results are presented in Table V.

Similar conclusions are drawn from the statistical tests. Both methods (PE and RA) achieve equivalent accuracies, but RA behavior in terms of GM and AvgAcc excels, rejecting the null hypotheses of equivalence with very low p-values.

Regarding the second configuration SVM_{Puk}, the results (shown in Table VI) are similar, but not so large differences are shown at first glance. In this case, PE achieves a slightly higher accuracy, whereas RA excels in the other two performance measures. The statistical analysis of these results is shown in Table VII.

TABLE IV. RESULTS USING SVM_{Poly} AS BASE CLASSIFIER.

	Accuracy		GM		AvgAcc	
Data-set	PE	RA	PE	RA	PE	RA
Autos	74.80	**75.38**	.5479	**.5624**	**72.69**	71.99
Balance	90.40	**91.68**	.8310	**.9156**	85.35	**91.79**
Car	92.71	**93.34**	.8651	**.9364**	87.18	**93.71**
Cleveland	**58.25**	51.16	.0000	**.0756**	30.88	**34.52**
Contraceptive	49.83	**50.71**	.4604	**.5102**	47.34	**51.54**
Dermatology	**94.13**	93.85	**.9408**	.9362	**94.58**	94.30
Ecoli	**77.69**	76.49	**.1544**	.1517	**68.18**	67.77
Flare	**74.67**	72.79	.4517	**.5914**	61.02	**65.54**
Glass	**61.26**	59.81	.2045	**.4596**	55.40	**61.78**
Hayes-Roth	52.22	**71.14**	.4985	**.7069**	55.05	**72.30**
Iris	**96.00**	**96.00**	.9580	**.9583**	**96.00**	**96.00**
Led7digit	**73.00**	71.80	**.7110**	.7014	**73.01**	71.90
Lymphography	81.68	**83.77**	**.3348**	.3325	64.87	**73.13**
NewThyroid	**97.21**	95.81	.9599	**.9621**	96.16	**96.38**
Nursery	**91.90**	91.43	.6529	**.6990**	82.22	**85.39**
Pageblocks	**94.70**	86.49	.3042	**.6658**	68.23	**78.89**
Penbased	95.27	**95.64**	.9513	**.9554**	95.29	**95.66**
Satimage	**84.14**	83.67	.7703	**.8015**	79.55	**81.36**
Segment	92.55	**93.85**	.9197	**.9359**	92.55	**93.85**
Shuttle	96.37	**96.92**	.3477	**.3631**	80.67	**83.30**
Splice	79.59	**80.22**	.8325	**.8374**	84.29	**84.69**
Tae	51.72	**55.72**	.4869	**.5407**	51.91	**55.57**
Thyroid	95.69	**96.94**	.4445	**.8817**	67.88	**89.29**
Vehicle	72.46	**73.05**	**.6970**	.6892	72.82	**73.49**
Vowel	69.90	**72.22**	.6822	**.7050**	69.90	**72.22**
Wine	**97.16**	**97.16**	**.9684**	**.9684**	**96.99**	**96.99**
Yeast	**59.10**	54.58	.0000	**.4088**	**56.74**	56.69
Zoo	**95.05**	**95.05**	**.0000**	**.0000**	**85.24**	**85.24**
Average	80.34	**80.60**	.5706	**.6519**	74.00	**77.69**

TABLE V. WILCOXON TESTS FOR SVM_{Poly} AS BASE CLASSIFIER.

Comparison	Measure	R^+	R^-	Hypothesis	p-value
RA vs. PE	Accuracy	219.0	187.0	Not rejected	0.756995
	GM	364.5	41.5	Rejected for RA at 95%	0.000220
	AvgAcc	361.0	45.0	Rejected for RA at 95%	0.000266

R^+ are ranks in favor of RA and R^- in favor of PE.

TABLE VI. RESULTS USING SVM_{Puk} AS BASE CLASSIFIER.

	Accuracy		GM		AvgAcc	
Data-set	PE	RA	PE	RA	PE	RA
Autos	**68.53**	61.51	**.2544**	.2156	**65.06**	59.78
Balance	**88.00**	87.84	**.8660**	.8497	**86.93**	85.71
Car	63.60	**71.18**	.7452	**.7763**	77.58	**80.37**
Cleveland	**45.09**	44.75	**.0000**	**.0000**	**29.78**	29.41
Contraceptive	**48.41**	45.01	.4406	**.4555**	45.70	**46.31**
Dermatology	**96.09**	95.26	**.9574**	.9478	**96.03**	95.29
Ecoli	**75.31**	75.01	.1381	**.1550**	67.35	**67.64**
Flare	**69.42**	64.35	.3277	**.5188**	59.43	**60.10**
Glass	70.60	**70.61**	.5372	**.5533**	68.04	**68.59**
Hayes-Roth	79.54	**81.05**	.8072	**.8163**	82.30	**83.58**
Iris	94.00	**94.67**	.9375	**.9442**	94.00	**94.67**
Led7digit	70.20	**70.80**	.6840	**.6928**	70.32	**71.01**
Lymphography	80.34	**81.01**	.1557	**.3374**	54.98	**61.65**
NewThyroid	**97.67**	**97.67**	**.9811**	**.9811**	**98.16**	**98.16**
Nursery	81.33	**83.33**	.6793	**.6902**	82.28	**83.72**
Pageblocks	**94.16**	93.43	**.2757**	.2666	**67.40**	65.41
Penbased	**97.82**	**97.82**	**.9781**	**.9781**	**97.85**	**97.85**
Satimage	84.92	**85.23**	.8315	**.8434**	84.16	**85.08**
Segment	97.10	**97.23**	.9704	**.9717**	97.10	**97.23**
Shuttle	**99.72**	99.22	.7650	**.9648**	93.14	**97.17**
Splice	64.56	**72.10**	.3787	**.7575**	51.44	**78.68**
Tae	56.30	**57.63**	.5513	**.5649**	56.24	**57.51**
Thyroid	**92.64**	92.50	.4971	**.5364**	62.44	**66.66**
Vehicle	80.49	**80.61**	.7873	**.7887**	80.71	**80.83**
Vowel	**99.39**	**99.39**	**.9936**	**.9936**	**99.39**	**99.39**
Wine	**98.30**	**98.30**	**.9857**	**.9857**	**98.60**	**98.60**
Yeast	**56.54**	54.18	.0000	**.0954**	**55.37**	55.14
Zoo	84.19	**93.05**	.0000	**.2000**	64.05	**80.00**
Average	79.80	**80.17**	.5902	**.6386**	74.49	**76.63**

TABLE VII. WILCOXON TESTS FOR SVM_{Puk} AS BASE CLASSIFIER.

Comparison	Measure	R^+	R^-	Hypothesis	p-value
RA vs. PE	Accuracy	220.0	186.0	Not rejected	0.710304
	GM	336.5	69.5	Rejected for RA at 95%	0.003502
	AvgAcc	307.0	99.0	Rejected for RA at 95%	0.022264

R^+ are ranks in favor of RA and R^- in favor of PE.

Observing the results of the tests, the superiority of RA outstands. Whereas the accuracy remains similar, both GM and AvgAcc are improved, rejecting the null hypotheses of equivalence with low p-values.

Interestingly, the classifiers giving better confidence degrees (the case of SVM_{Poly}) have more margin for improvement, since they provide more information to the classification process. In the case of SVM_{Puk}, the configuration (parameter C) used causes to produce too borderline (close to 0 or 1) values in the probability estimates, which are not as useful as those given by SVM_{Poly}. This is the main reason of the differences in their results. Furthermore, having an initially lower GM and AvgAcc values (with PE), SVM_{Poly} has achieved higher results than SVM_{Puk} using RA.

VII. CONCLUDING REMARKS

We have put forward the difficult classes problem in OVO strategy, which has not been previously addressed. In order to improve the classification accuracy over the difficult classes, we have proposed a new aggregation methodology, generalizing the weighted voting strategy.

This methodology is able to properly learn the parameters for the REFs used in the aggregation, yielding to statistical differences in terms of GM (which was our main objective) with respect to the aggregations which do not take into account such a problem. Recall that the differences shown between OVO methods are only due to the aggregation itself, since the score-matrices are exactly the same. Moreover, the GM improvement with respect to the previous OVO aggregation has not been at the expenses of accuracy. Hence, we have shown that the base classifiers can be managed in such a way that different objectives can be obtained, without needing to alter them. The results obtained have shown that there is much margin for improvement in terms of GM and AvgAcc, which could be more important than accuracy in many applications.

ACKNOWLEDGMENT

This work was partially supported by the Spanish Ministry of Science and Technology under projects TIN2011-28488 and TIN-2012-33856 (with FEDER founds) and the andalusian regional projects P10-TIC-06858 and P11-TIC-7765.

REFERENCES

[1] A. C. Lorena, A. C. Carvalho, and J. M. Gama, "A review on the combination of binary classifiers in multiclass problems," *Artificial Intelligence Review*, vol. 30, no. 1-4, pp. 19–37, 2008.

[2] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *Journal of Machine Learning Research*, vol. 1, pp. 113–141, 2000.

[3] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes," *Pattern Recognition*, vol. 44, no. 8, pp. 1761 – 1776, 2011.

[4] ——, "Dynamic classifier selection for one-vs-one strategy: Avoiding non-competent classifiers," *Pattern Recognition*, vol. 2013 in press doi: 10.1016/j.patcog.2013.04.018, 2013.

[5] H. Bustince, E. Barrenechea, and M. Pagola, "Restricted equivalence functions," *Fuzzy Sets and Systems*, vol. 157, no. 17, pp. 2333–2346, 2006.

[6] L. J. Eshelman and J. D. Schaffer, "Real-coded genetic algorithms and interval-schemata," in *Foundation of Genetic Algorithms 2*, D. L. Whitley, Ed. San Mateo, CA: Morgan Kaufmann, 1993, pp. 187–202.

[7] A. Asuncion and D. J. Newman, "UCI machine learning repository," 2007, http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[8] 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, pp. 255–287, 2010.

[9] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[10] S. García and F. Herrera, "An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons," *Journal of Machine Learning Research*, vol. 9, pp. 2677–2694, 2008.

[11] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[12] O. Pujol, P. Radeva, and J. Vitria, "Discriminant ECOC: a heuristic method for application dependent design of error correcting output codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 1007–1012, 2006.

[13] T. K. Paul and H. Iba, "Prediction of cancer class with majority voting genetic programming classifier using gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 2, pp. 353–367, 2009.

[14] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[15] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, no. 1, pp. 27–38, 2009.

[16] R. Barandela, J. S. Sánchez, V. García, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognition*, vol. 36, no. 3, pp. 849–851, 2003.

[17] H. Bustince, E. Barrenechea, and M. Pagola, "Image thresholding using restricted equivalence functions and maximizing the measures of similarity," *Fuzzy Sets and Systems*, vol. 158, no. 5, pp. 496–516, 2007.

[18] E. Hüllermeier and S. Vanderlooy, "Combining predictions in pairwise classification: An optimal adaptive voting strategy and its relation to weighted voting," *Pattern Recognition*, vol. 43, no. 1, pp. 128–142, 2010.

[19] J. Sanz, A. Fernández, H. Bustince, and F. Herrera, "A genetic tuning to improve the performance of fuzzy rule-based classification systems with interval-valued fuzzy sets: Degree of ignorance and lateral position," *International Journal of Approximate Reasoning*, vol. 52, no. 6, pp. 751–766, 2011.

[20] M. Lozano, F. Herrera, N. Krasnogor, and D. Molina, "Real-coded memetic algorithms with crossover hill-climbing," *Evolutionary Computation*, vol. 12, pp. 273–302, 2004.

[21] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.

[22] J. Alcalá-Fdez, L. Sánchez, S. García, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, J. Fernández, and F. Herrera, "KEEL: a software tool to assess evolutionary algorithms for data mining problems," *Soft Computing*, vol. 13, no. 3, pp. 307–318, 2008.

[23] T. F. Wu, C. J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[24] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.