A Game-Theoretic Approach to Weighted Majority Voting for Combining SVM Classifiers

Harris Georgiou[1], Michael Mavroforakis[1], Sergios Theodoridis[1]

[1] Dept. of Informatics and Telecommunications, Division of Communications and Signal Processing, University of Athens, Greece
http://www.di.uoa.gr/~dsp
157 84, Panepistimioupolis, Ilissia, Athens, Greece
Harris Georgiou, xgeorgio@di.uoa.gr

Abstract. A new approach from the game-theoretic point of view is proposed for the problem of optimally combining classifiers in dichotomous choice situations. The analysis of weighted majority voting under the viewpoint of coalition gaming, leads to the existence of *analytical solutions* to optimal weights for the classifiers based on their prior competencies. The general framework of weighted majority rules (WMR) is tested against common rank-based and simple majority models, as well as two soft-output averaging rules. Experimental results with combined support vector machine (SVM) classifiers on benchmark classification tasks have proven that WMR, employing the theoretically optimal solution for combination weights proposed in this work, outperformed all the other rank-based, simple majority and soft-output averaging methods. It also provides a very generic and theoretically well-defined framework for all hard-output (voting) combination schemes between any type of classifier architecture.

1 Introduction

1.1 Classifier Combination and Game Theory

In the discipline of collective decision-making, a group of N experts with moderate performance levels are combined in an optimal way, in order to produce a collective decision that is better than the best estimate of each individual expert in the group. According to the famous Condorcet Jury Theorem [1], if the experts' individual decisions are independent and their corresponding estimations are more likely to be correct than incorrect ($p_{correct}>0.5$), then an increase in the collective performance, as a group, is guaranteed when the individual estimations are combined. Moreover, this increase in performance continues to increase asymptotically as the size N of the group increases.

In the case where each expert selects only one out of M available options, the collective group decision can be estimated by the majority voting scheme, i.e., the choice

selected is the one gathering the majority of votes. When the simple majority rule is employed, each of the N experts acts with the same common interest of reaching the optimal collective decision. However, their individual choices place them in possibly contradicting estimations, with each expert trying *to impose* its decision to the others and to the group. This is a typical competitive situation, which can be modeled by the well-studied theory of non-zero sum competitive gaming in classic Game Theory [2]. In reality, each subgroup of consentient experts essentially represents an opposing assembly to all the other similar subgroups with different consensus of choice. It is clear that this second type of cooperative, instead of purely competitive (per expert), gaming reflects the problem of collective decision-making in the most generic way. Special sections of Game Theory, namely the Coalitions and Stable Sets in cooperative gaming [2], have studied the effects of introducing "weights" to the choice of each expert according to their competencies, in order to optimize the final decision of the group.

1.2 Weighted Majority Games and Weighted Majority Rules

The case of a dichotomous situation, where there are only two symmetrical choices for each expert (i.e., M=2) to vote for, then this restricted form is known as the weighted majority game (WMG) [2]. It has been proven by Nitzan and Paroush (1982) [3] and Shapley and Grofman (1984) [4], that the optimal decision rules, in terms of collective performance, are the weighted majority rules (WMR); this is in fact a different name for the well-known weighted majority voting schemes [5], which are often used in pattern recognition for combining hard-output classifiers. The same assertion has also been verified by Ben-Yashar and Nitzan [6] as the optimal aggregation rule for committees under the scope of informative voting in Decision Theory. Although there is in fact an exponential number of such WMR for each WMG, only a few of them can be proven to be well-defined or *qualified* combination rules and even fewer can be proven to be unique, i.e., not producing exactly the same decision profile with others [7]. For example, in the 2^{32} possible[1] voting games of five experts, there are exactly 85 qualified WMR if only positive integer weights are permitted, of which only seven are unique in terms of their decision profile [7].

In this paper, the notion of modeling dichotomous choice situations for a group of experts via the theory of WMG and WMR is for the first time applied for combining hard-output classifiers. Under the conditional independence assumption, a *closed form solution* for the voting weights in the WMR formula exists and it is *directly linked to each expert's competency*. This optimal weight profile for the voting experts is the log of the odds of their individual competencies [3], [4], [7], [8].

In this paper, this particular type of game-theoretic analytical solution for optimal expert combinations in dichotomous choice situations is tested for the first time against other popular combination schemes. The possibility of having a weighted voting scheme that is based only on the prior capabilities of the experts in the group, as well as on the theoretical assertion that this analytical solution is optimal, in terms

[1] For five experts with two choices each there are 2^5=32 decision profiles, each of which can be generally mapped in any of the two possible outputs of the combination rule. See [7].

of collective competency (at least for all non-trained, i.e., iteratively optimized, weights), is extremely attractive as an option of designing very simple yet effective combination models for an arbitrary pool of classifiers.

2 Datasets and Methods

2.1 SVM Classifier Model

The SVM classifier was used as the base model for creating a pool of classifiers for each combination scheme. Specifically, a geometric nearest point algorithm (NPA) [9], based on the notion of reduced convex hulls (RCH) [10], was used for training a standard SVM architecture with radial-basis function (RBF) as the kernel of the non-linear mapping. In previous studies [11] have shown experimental evidence that optimal combinations of SVM classifiers can be achieved through linear combination rules, i.e., the same category of combination rules examined in this study. In the two averaging combination rules that use the soft-output of the individual classifiers, the distances from the decision boundary were used instead of the (thresholded) hard-output of the SVM classifier, as they are indicative of the corresponding classification confidence [12], [13].

2.2 Datasets and Feature Grouping

In order to assess the performance of each classifier combination method, a number of publicly available test datasets [14], with known single-classifier accuracy rates for this specific SVM training model, were used. These datasets are: 1) Diabetis, 2) Flare-Solar, 3) German, 4) Heart and 5) Waveform.

Each base dataset was randomly separated into a base training set and a validation set of samples. In order to make individually trained classifiers as “independent” as possible, the method of training them in different subspaces was employed. As it has been reported previously, e.g., [13], [15], this is an effective approach towards independence among classifiers. To this end, the training set was partitioned into K distinct segments of feature groupings, i.e., containing only some of the features (dimensions) of the initial dataset. Each group of features was created in a way that satisfied two constraints: (a) each group to be distinct, i.e., no feature is included in two or more groups, and (b) each group to contain a subset of features that can describe the classification task equally well as the other feature groups, i.e., employ a “fair” distribution of the available features into K groups. Satisfaction of the second constraint required a method for ranking all the features in terms of discrimination power against the two classes, as well as their statistical independency to all the other features in the initial training set. Thus, the MANOVA method [16] was used to assign a multivariate statistical significance value to each one of the features and then produce a sorted list based on (the log of) this value.

In order to create a "fair" partitioning of this list into equally efficient segments, features were selected in pairs from the top and bottom positions, putting the currently "best" and "worst" features in the same group. Furthermore, the efficiency of each group was measured in terms of summing the log of the statistical significance value, assigned by MANOVA, of all the features contained in this group. The log was employed in order to avoid excessive differences between the values assigned by MANOVA, thus creating more even subset sums of these values. Essentially, every such pair of features was assigned in groups sequentially, in a way that all groups contained features with approximately equal sum of the log of the values assigned by MANOVA. In other words, the MANOVA-sorted list of features was "folded" once in the middle and then "cut" into K subsequent parts of equal sums of log-values, i.e., with every part exhibiting roughly the same sum of the log of the statistical significance values, accompanying each feature included in this part.

Each one of these K distinct feature groups was used for training an individual SVM classifier. Thus, each of these K classifiers used a different, dimensionally reduced, version of the original (full) training set and therefore learns a totally different classification task.

2.3 Classifier Combination Methods

Nine linear combination rules were examined in this study. Specifically, five hard-output combination methods were employed, namely three standard rank-based methods and two voting-based schemes. These rank-based rules are [8], [13]:

- minimum ("min")
- maximum ("max")
- median ("median")

The two majority rules, including the WMR model, are [8], [13]:

- simple majority voting ("majority")
- weighted majority voting, i.e.:

$$O_{wmr}(\vec{x}) = \sum_{i=1}^{K} w_i D_i(\vec{x}) \ . \tag{1}$$

where D_i is the hard-output of each of the K individual classifiers in the pool, w_i is its assigned weight and O_{wmr} the weighted majority sum. The final hard-output decision D_{wmr} of the WMR is taken against a fixed threshold (T) that defines the decision boundary for the combination rule [7], [8]:

$$D_{wmr}(\vec{x}) = sign\left(O_{wmr}(\vec{x}) - T\right) \ . \tag{2}$$

Specifically for the weighted majority voting scheme, three different methods for calculating the weight profile were tested for comparative results:

- "direct" weighting profile for WMR ("wmr/direct") [5], [8]:

$$w_i = p_i \quad , \quad p_i = P_i(\theta = \omega_{correct} \mid \vec{x}) \ . \tag{3}$$

- "odds" weighting profile for WMR ("wmr/odds") [7], [8]:

$$w_i = \frac{p_i}{1-p_i} \quad , \quad p_i = P_i(\theta = \omega_{correct} \mid \vec{x}) \ . \tag{4}$$

- "logodds" weighting profile for WMR ("wmr/logodds") [7], [8]:

$$w_i = \log\left(\frac{p_i}{1-p_i}\right) \quad , \quad p_i = P_i(\theta = \omega_{correct} \mid \vec{x}) \ . \tag{5}$$

where w_i is the combination weight assigned for the i-th classifier, p_i is its prior probability for correct classification, measured in the validation set, and θ, ω are the predicted class labels.

Additionally, two soft-output averaging models were included, a non-weighted and a weighted [8]:

- simple average ("average")
- weighted average ("lsewavg")

The weights in the weighted average rule were calculated as the optimal weighting profile of the individual classifier outputs against the correct classification tag, in terms of a least-squares error (LSE) minimization criterion [15]. Thus, this method can be considered as an example of "trained" weighting rules of soft-output classifiers. In contrast, the WMR approach employs fixed analytical weighting profile and hard-output classifications (votes) as input, that is, no further training is required.

3 Experiments and Results

The evaluation of the combination models consisted of two phases, namely: (a) the design and training of SVM classifiers, trained in distinctly different subspaces, and (b) the application of the various combination schemes to the outputs of the individual classifiers.

Each of the K classifiers was separately trained and optimized, using a different group of features from the full dataset, and subsequently evaluated using the corresponding validation set. This training/validation cycle was applied three times, for each of the five datasets, each time using a new random partitioning of the full dataset into training and validation sets. The mean values and standard deviations of the success rates of all the individual ($3K$) classifiers for each dataset, as well as the details about the size and dimensionality of each (full) training and validation sets, are presented in Table 1.

The K value, i.e., the number of feature groups for each dataset, was determined experimentally in a way that each of the corresponding K training segments would be adequate to produce a well-trained SVM classifier. Thus, the German training set was split in K=5 segments, while the Flare-Solar training set in K=4 segments.

Table 1. Single versus multiple classifier accuracy percentages per dataset and *K* values (number of dataset partitions)

Dataset	Train set	Vali-dat. set	Data Dim.	Single classifier accuracy	*K* value	Individual classifier mean acc%
diabetis	468	300	8	76.5 ± 1.7	5	68.3 ± 3.9
flare-solar	666	400	9	67.6 ± 1.8	4	55.7 ± 3.6
german	700	300	20	76.4 ± 2.1	5	68.9 ± 1.8
heart	170	100	13	84.0 ± 3.3	5	74.3 ± 2.3
waveform	400	4600	21	90.1 ± 0.4	5	81.1 ± 1.2

The classification outputs of the pool of *K* classifiers from each training/validation cycle were fed as input to all nine combination schemes, producing the corresponding combined classification outputs. Since the output of each of the *K* classifiers in the pool was calculated based on the same (dimensionally reduced) validation set, the corresponding outputs and accuracy of the combination rules also refer to this validation set.

Table 2 illustrates the mean accuracy of each combination rule (each cell corresponds to three training/validation cycles), as well as the mean value and standard deviation of the success rates of all nine combination rules, for each dataset and *K* value employed.

Table 2. Mean accuracy percentages of all the nine combination rules, with optimized decision threshold, per dataset and *K* values (number of feature groups and classifiers)

Combination Rule	Diabetis	Flare-Solar	German	Heart	Waveform
	K=5	*K*=4	*K*=5	*K*=5	*K*=5
average	71.67	66.08	70.67	85.33	88.12
lsewavg	76.11	65.58	71.56	85.00	86.79
min	68.56	55.92	70.67	69.00	72.98
max	69.11	60.42	67.33	76.67	85.95
median	69.00	58.33	69.78	80.00	81.17
majority	73.00	63.75	70.67	82.33	86.59
wmr/direct	74.00	66.58	70.67	82.33	86.59
wmr/odds	75.33	66.58	71.33	84.00	86.70
wmr/logodds	75.33	66.42	71.33	84.00	86.64
Mean	72.46	63.30	70.44	80.96	84.62
Stdev	2.99	4.06	1.28	5.25	4.77

In the sequel, the overall relative performance of each combination rule was determined in terms of ranking position for each case, i.e., according to its corresponding accuracy for each dataset and *K* value employed. Specifically, a weighted Borda scheme (wBorda) [17] was employed to attribute 10 points to the top-ranked combi-

nation rule, 9 points to the second, and so on. In case of a "tie" where two combination rules exhibited exactly the same classification accuracy, both got the same wBorda points for the specific ranking position. Using the results from Table 3, regarding the accuracies, Table 4 illustrates the corresponding wBorda ranking points of all nine combination rules, for each dataset and *K* value employed in this study.

Table 3. wBorda value of all combination rules, with optimized decision threshold, per dataset and *K* values. Each cell value represents the ranking weight according to classification accuracies, with 10 points for top position, 9 points for the second and so on. In cases of equal accuracies, the same ranking weight was assigned to the corresponding combination rules

Combination Rule	Diabetis	Flare-Solar	German	Heart	Waveform
	K=5	*K*=4	*K*=5	*K*=5	*K*=5
average	6	8	8	10	10
lsewavg	10	7	10	9	9
min	3	3	8	4	3
max	5	5	6	5	5
median	4	4	7	6	4
majority	7	6	8	7	6
wmr/direct	8	10	8	7	6
wmr/odds	9	10	9	8	8
wmr/logodds	9	9	9	8	7

Table 4. Overall evaluation of all the combination rules, with optimized decision threshold, using the wBorda results for all datasets and *K* values available. The list is sorted according to the wBorda sum and mean ranking position of each combination rule, from the best to the worst combination rule

Combination Rule	wBorda Sum	wBorda Mean	wBorda Stdev
lsewavg	45	9.0	1.22
wmr/odds	44	8.8	0.84
average	42	8.4	1.67
wmr/logodds	42	8.4	0.89
wmr/direct	39	7.8	1.48
majority	34	6.8	0.84
max	26	5.2	0.45
median	25	5.0	1.41
min	21	4.2	2.17

Table 4 presents a summary of the results shown in Table 3, as well as the list of all the combination rules sorted according to their sum of wBorda points, i.e., their overall efficiency throughout the five original datasets. Tables 2 through 4 present the performance and wBorda results for all the combination rules with optimized decision

threshold (T). The decision threshold employed by each combination rule was in every case optimized against final accuracy, using a typical Newton-Raphson optimization algorithm [18].

4 Discussion

The results from Tables 3 and 4 clearly demonstrate the superior performance of the WMR model. Specifically, the all versions of the WMR model exhibited the best performance amongst all the other hard-output combination rules. As expected, it has been proven better than the simple majority voting, as well as all the other rank-based methods (max, min, median). The "odds" weighting profile has also been proven marginally better than the "direct"- and the "logodds"-based profiles for designing the optimal WMR formula.

Interestingly, the "odds"-based version of WMR exhibited better performance than the simple averaging rule, e.g., a soft-output combination model, losing only from the weighted averaging rule with LSE-trained weights. Thus, the WMR model, especially with the "odds" and "logodds" weighting profiles, performs equally well or better than simple soft-output averaging combination rules. All four weighted combination rules, i.e., the three WMR and the LSE-trained weighted average, have been clearly proven better than all the non-weighted hard-output combination rules.

Table 4 also demonstrates the robustness and stability of the each combination rule. For small values of standard deviation (less than unity) in the corresponding wBorda mean ranks, the relative ranking position of a combination rule against the others remains more or less the same. Thus, the maximum rule exhibits a consistently lower ranking position than the simple majority rule, while the "odds"- and the "logodds"-based versions of the WMR models perform consistently better than the simple majority and the three rank-based rules. Furthermore, the "odds"- and the "logodds"-based versions of WMR exhibit the same consistency and robustness as the simple majority rule but with higher success rates.

With respect to the overall performance of the combination rules, results from Tables 1 and 2 demonstrate that in all cases the best combination rules increased the overall success rates of the classifier pool, from +2% (German dataset) to +11% (Flare-Solar dataset), in many cases very close to or equal to the corresponding reference performance level of the single SVM classifier results.

The ensemble of these classifiers clearly demonstrates that the combination of multiple simpler models, each using a $1/K$ portion of the feature space of the dataset, instead of a single classifier for the complete feature space, can be used to reduce the overall training effort. Specifically for the SVM model, kernel evaluation employs inner product between vectors, i.e., its complexity is directly proportional to the dimensionality (number of features) in the input vectors. If this feature space reduction, from F to F/K features, results in a proportional increase in the complexity of the new (reduced) input space in terms of new class distributions, then it is expected that the training of each of the K SVM classifiers may be completed up to K times faster on average. A similar approach has also been examined in other studies [11], using an ensemble of SVM classifiers trained in small training sets, instead of one large train-

ing set for a single SVM classifier. Furthermore, there is evidence that such ensembles of kernel machines are more stable than the equivalent kernel machines [11]. This reduction in training time, of course, has to be compared to the additional overhead of calculating a combination rule for every output vector from the classifier pool. Consequently, if the optimal design of this combination rule is simple (linear) and efficient, and its weighting profile can be determined analytically with no need for iterative weight optimization, the WMR approach could prove very prominent for this role in classification tasks of high dimensionality and/or dataset sizes.

5 Conclusions

The game-theoretic modeling of combining classifiers in dichotomous choice problems leads to cooperative gaming approaches, specifically coalition gaming in the form of WMG. Theoretically optimal solutions for this type of games are the WMR schemes, often referred to as weighted majority voting. Under the conditional independence assumption for the experts, there exists a closed solution for the optimal weighting profiles for the WMR formula.

In this paper, experimental comparative results have shown that such simple combination models for ensembles of classifiers can be more efficient than all typical rank-based and simple majority schemes, as well as simple soft-output averaging schemes in some cases. Although the conditional independence assumption was moderately satisfied by using distinct partitions of the feature space, results have shown that the theoretical solution is still valid to a considerable extent. Therefore, the WMR can be asserted as a simple yet effective option for combining almost any type of classifier with others in an optimal and theoretically well-defined framework.

References

1. Condorcet, Marquis de: An Essay on the Application of Probability Theory to Plurality Decision Making: An Election Between Three Candidates. In: Sommerlad and Mclean (1989) 66-80
2. Owen, G.: Game Theory. 3rd edn. Academic Press, San Diego USA (1995)
3. Nitzan, S., Paroush, J.: Optimal Decision Rules in Uncertain Dichotomous Choice Situations. Int. Econ. Rev., 23 (1982) 289–297
4. Shapley, L., Grofman, B.: Optimizing Group Judgemental Accuracy in the Presence of Independence. Public Choice, 43 (1984) 329–343
5. Littlestone, N., Warmuth, M. K.: The Weighted Majority Algorithm. Information and Computation, 108 (1994) 212–261
6. Ben-Yashar, R., Nitzan, S.: The Optimal Decision Rule for Fixed-Size Committees in Dichotomous Choice Situations: The General Result. International Economic Review, 38 (1997) 175–186
7. Karotkin, D.: The Network of Weighted Majority Rules and Weighted Majority Games. Games and Econ. Beh., 22 (1998) 299–315
8. Kuncheva, L. I.: Combining Pattern Classifiers. John Wiley and Sons, New Jersey USA (2004)

9. Mavroforakis, M., Theodoridis, S.: Support Vector Machine Classification through Geometry. Proc. XII Eur. Sig. Proc. Conf. (EUSIPCO2005), Antalya, Turkey, Sep. 2005
10. Mavroforakis, M., Theodoridis, S.: A Geometric Approach to Support Vector Machine (SVM) Classification. IEEE Trans. NN, 2006 (in press)
11. Evgeniou, T., Pontil, M., Elisseeff, A.: Leave One Out Error, Stability, and Generalization of Voting Combinations of Classifiers. Machine Learning, 55 (2004) 71–97
12. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press UK (2000)
13. Jain, A., Duin, R., Mao J.: Statistical pattern recognition: a review. IEEE Trans. PAMI, 22 (2000) 4–37
14. Mavroforakis, M., Sdralis, M., Theodoridis, S.: A novel SVM Geometric Algorithm based on Reduced Convex Hulls. 18th International Conference on Pattern Recognition (ICPR 2006), Hong Kong, Aug. 2006
15. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. 4th edn. Academic Press, San Diego USA (2006)
16. Cooley, W., Lohnes, P.: Multivariate data analysis. John Willey & Sons, New York USA (1971)
17. Parker, J. R.: Rank and Response Combination from Confusion Matrix Data. Information Fusion, 2 (2001) 113–120
18. Forsythe, G., Malcom, M., Moler, C.: Computer Methods for Mathematical Computations. Prentice-Hall, New Jersey USA (1977)