ICANN2006

Harris Georgiou, Michael Mavroforakis, Sergios Theodoridis:

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

Athens, 10 – 13 Sept 2006

Overview:

- Introduction
- Material and Methods
  - Classifiers
  - Data partitioning
  - Classifiers' Combination Rules
- Results
- Conclusions

Introduction I: The game-theoretic approach

- In collective decision-making, *N* experts of moderate competencies are to be combined to produce better results.
- *Condorcet Jury Theorem* [1]: "If experts are *independent* and their competency is better than moderate ($P_{correct}$>0.5), collective competency increases asymptotically as *N* increases."
- In *M*-class classification problems, simple majority voting is used.
- "Simplistic" game-theoretic model: each expert acts **on its own** trying to impose its decision to the collective output.

  » non zero-sum N-player purely competitive gaming
- "Realistic" game-theoretic model: experts form **coalitions** according to their consensus and compete as opposing **assemblies** to determine the collective output.

  *» non zero-sum N-player cooperative / **coalition gaming***

Question:

How do the game-theoretic solutions apply to classifier ensembles?

Introduction II: WMG and WMR framework

- In dichotomous choice situations (M=2 classes), coalition games become *"Simple Games"* or Weighted Majority Games (**WMG**).
- Game Theory [3-5]: Optimal decision rules for WMG in terms of collective performance are the Weighted Majority Rules (**WMR**).
- Furthermore, there is **analytical solution** to the **optimal weighting profile** used in WMR, directly linked to the individual experts' competencies: $P_{correct}(j)$
- Optimal weights for the WMR are calculated according to the **log-odds** formula [3-4,7-8]
- Optimality of closed form solution depends on conditions:

  (a) each expert's competency is better than random, i.e., $P_{correct}(j)>0.5$

  (b) experts' decisions are independent to each other

Question:

How efficient is the WMR scheme in designing classifier ensembles?

Material and Methods: Overview

Purpose of this study: Investigate the efficiency of weighted voting schemes in ensembles of SVM classifiers, using the corresponding game-theoretic optimal solutions for WMR.

1. **Task:** 2-class classification problem using benchmark datasets (Raetsch [14]: *Diabetis*, *Flare-Solar*, *German*, *Heart*, *Waveform*).
2. **Base classifier:** Support Vector Machine (**SVM**) with RBF kernel, trained with new geometric nearest point algorithms (NPA) for reduced convex hulls (**RCH**) [10].
3. **Diversity:** apply "guided" feature groupings to create distinct **subspaces** of roughly equal discrimination power.

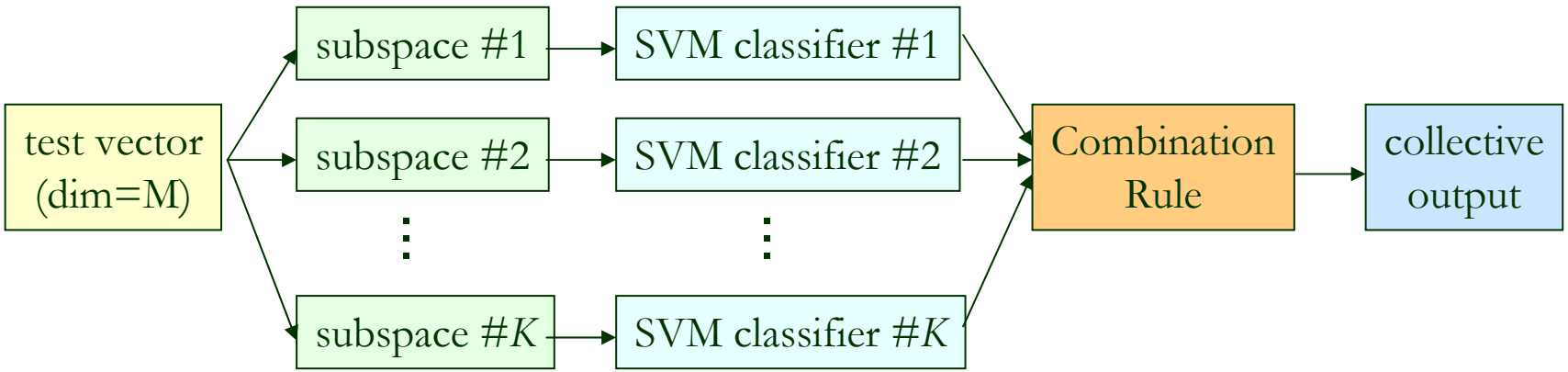
Notes on using SVM:

- Soft-output: distance from decision boundary is used as output
- Hard-output ("vote"): threshold soft-output against boundary

Material and Methods: Feature groupings

- In *random subspaces* methods, commonly used in *Random Forests* [19] and *Rotation Forests* [20], feature selections (groups) are created randomly, in order to produce distinct subspaces.
- Instead, here the features are ranked (sorted) against their individual discrimination power R_m for **non-random groupings**.
- The R_k is the **log of the MANOVA** [16] significance value, calculated for the initial (full) dataset.
- Given the required number K of distinct feature groups, the R_m-sorted list of features is partitioned into K parts.
- The **feature groups** G_k are generated so that the sums of R_m values contained in each one is roughly the same.
- Generation of G_k is deterministic and sequential, selecting pairs of features from the top and the bottom of the R_m-sorted list.
- Using MANOVA as the feature ranking method ensures that statistical dependencies between features are taken into account.

Material and Methods: Classifier combination rules



Combination rules used:

- minimum
- maximum
- median
- majority (simple)
- weighted majority (**WMR**)
- average
- weighted average (LSE-opt)

WMR weighting profiles tested:

$$p_i = P_i(\theta = \omega_{correct} \mid \vec{x})$$

- "direct": $w_i = p_i$
- "odds": $w_i = \dfrac{p_i}{1 - p_i}$
- "logodds": $w_i = \log\left(\dfrac{p_i}{1 - p_i}\right)$

Results: fullspace vs subspace classifier accuracies

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

Dataset	Train set	Validat. 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

Results: Combination rules accuracies

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

Results: Combination rules ranking positions

Table 3. Weighted-Borda (wBorda) [17] 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

Results: Summary of ranking positions

Table 4. Overall evaluation of all the combination rules, with optimized decision threshold, using the wBorda [17] 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

Discussion: Comparative performances (1)

1. WMR model exhibited the best performance over all other hard-output combination schemes, including simple majority voting.
2. WMR with "odds" and "logodds" weighting profiles have been proven better than the "direct" scheme, thus validating the theoretical solution.
3. WMR with "odds" weighting profile has been proven better than simple averaging scheme, despite the loss of information due to the hard-output thresholding of classifier outputs.
4. All weighted combination rules (i.e., WMR and LSE-weighted average) performed significantly better than the non-weighted rules, as expected.
5. According to the *mean* and *stdev* values in Table-4, WMR schemes are significantly more efficient and stable, in terms of ranking position (wBorda mean/stdev), against all other rank-based (max, min, median) and simple majority models.

Discussion: Comparative performances (2)

6. According to Table-2 and Table-3, in all cases the WMR combination schemes resulted to an increase over the average accuracy of the classifier pool, from +2% (German) to +11% (Flare-solar).
7. Although, in all cases, the hard-output combination rules employed explicit optimization of the decision threshold value *T*, analytical results have shown that the optimal *T* was always very close to the default *T* (middle-value) and data-adaptive optimization resulted in only marginal improvement to the combination rules' performance.
8. Overall results show that using WMR ensembles of SVM classifiers, each using a subspace of roughly 1/*K* portion of the full feature space, produces results competitive to the ones produced by the corresponding single SVM classifier.
9. Results validate related experimental evidence [11] that optimal ensemble schemes of SVM classifiers need not be more complex than (weighted) linear combination rules of their outputs.

Advantages of using WMR

- They are based on a **solid theoretical framework** with insights to analytical solutions and application to classifier ensembles.
- Employing a linear model, similar to a weighted average of independent factors, indicates that WMR can be a fully **parallelizable** model for ensembles of classifiers.
- Since the optimum weighting profile depends solely on each classifier's individual competency (which is trivial to calculate it directly), WMR is a very prominent technique for **iterative** and **on-line** applications of classifiers ensembles.

Updated work

- More recent work includes "adaptive" versions of WMR
- Use WMR for combining various types classifiers (e.g. OBTC)
- Compare WMR to more robust combination schemes (e.g. BSR)
- Extend and study diversity via non-random subspaces

Conclusions

- Game-theoretic modeling of combining multi-experts' decisions in dichotomous choice situations leads to Coalition Gaming and WMG in particular.
- For such WMG setups, WMR schemes are the theoretically optimal combination rules, with closed-form solution for their corresponding optimal weighting profile for the experts' "votes".
- The performance of such WMR was evaluated in benchmark datasets using multiple SVM classifiers as the experts pool, each trained in a different subspace (non-random feature group).
- Experimental comparative results were evaluated against various typical voting, hard-output and soft-output combination schemes, as well as the reference performance of the corresponding single SVM classifier.
- Although the conditional independence assumption is moderately satisfied by employing subspaces, results have shown that the simple analytical WMR solutions to weighting profiles are valid.

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 Situa-tions. 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 Com-putation, 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 Geome-try. 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)
19. Breiman, L.: Random Forests. Machine Learning, 45 (2001), 5–32
20. Rodriguez, J., Kuncheva, L., Alonso, C.: Rotation Forest – A New Classifier Ensemble Method. IEEE Trans.PAMI, 28 (2006) 1619–1630