Evaluation of Fusion Methods for γ-divergence-based Neural Network Ensembles

Uwe Knauer
Fraunhofer IFF
Magdeburg, Germany
Email: uwe.knauer@iff.fraunhofer.de

Andreas Backhaus
Fraunhofer IFF
Magdeburg, Germany
Email: andreas.backhaus@iff.fraunhofer.de

Udo Seiffert
Fraunhofer IFF
Magdeburg, Germany
Email: udo.seiffert@iff.fraunhofer.de

***Abstract*—A significant increase in the accuracy of hyperspectral image classification has been achieved by using ensembles of radial basis function networks trained with different number of neurons and different distance metrics. Best results have been obtained with γ-divergence distance metrics. In this paper, previous work is extended by evaluation of different approaches for the fusion of the multiple real-valued classifier outputs into a crisp ensemble classification result. The evaluation is done by 10-fold cross-validation. The obtained results show that an additional gain in classification accuracy can be achieved by selecting the appropriate fusion algorithm. Second, the SCANN algorithm and Fuzzy Templates are identified as the best performing fusion methods with respect to the complete ensemble of base classifiers. For several subsets of classifiers Majority Voting yields similar results while other simple combiners perform worse. Trainable combiners based on Adaptive Boosting and Random Forest are ranked among the top methods.**

I. Introduction

Hyperspectral imaging is considered a key technology for monitoring the environment, smart farming, quality control in food production, and many others. However, the analysis of hyperspectral data at a high spatial and spectral resolution is a challenging task especially for the design and the application of machine learning algorithms.

Ensembles of artificial neural networks based on non-standard metrics have been found an effective solution to improve classification performance over single classifiers for the classification of hyperspectral data. Especially, the combination of radial basis function networks using the γ metric as similarity measure has shown superior results. This report investigates a number of different combination algorithms to get a better picture of the dependence between datasets, individual base classifiers, and the chosen combination scheme.

In Sec. II we briefly introduce the applied base classifiers, the selected combination algorithms, and the validation procedure. In Sec. III we list the used datasets. The experimental results are presented and discussed in Sec. IV.

II. Methods

A. Base classifiers

The set of base classifiers was not extended compared to previous studies reported in [1], [2]. Classification models were implemented as published in [3], [4], [5]. In the methods GLVQ and SNG no non-linearity in the energy function was used. The distance function between a data vector $\mathbf{v}$ and a prototype vector $\mathbf{w}$ (respectively the hidden neurons in the RBF) was either the squared Euclidean distance defined as

$$d(\mathbf{v},\mathbf{w}) = \sum_i (v_i - w_i)^2 , \tag{1}$$

or the γ-divergence defined as

$$d(\mathbf{v},\mathbf{w},\gamma) = \log\left(\frac{\left(\sum_i v_i^{\gamma+1}\right)^{\frac{1}{\gamma(\gamma+1)}} \left(\sum_i w_i^{\gamma+1}\right)^{\frac{1}{\gamma+1}}}{\left(\sum_i v_i w_i^{\gamma}\right)^{\frac{1}{\gamma}}} \right) . \tag{2}$$

The γ-divergence with $\gamma = 2$ is widely known as the Cauchy-Schwarz distance. All three classifier methods (RBF, SNG, GLVQ) are formulated as energy minimization problem solved usually by stochastic gradient descent. In order to avoid a manually chosen step-size, we used the non-linear conjugate gradient approach with automatic step size from the optimization toolbox 'minFunc' available for Matlab.

The parameter γ was set varying from 1 to 10 in steps of one. Additionally, the generalized Kullback-Leiber divergence [6] was used to investigate the behavior for convergence of γ to zero. Prototype vectors and network weights were initialized randomly. The RBF used a 1-of-N coding scheme at its output to represent discrete class information. In the RBF, SNG, and GLVQ the prototypes were pre-trained using a Neural Gas with the Euclidean distance or γ-divergence as similarity function with an identical setup compared to the later classification model. In the GLVQ and SNG model, separate pre-learning runs for prototypes from identical classes were performed. The dataset was divided into training and test data according to a cross-validation scheme with stratified random sampling.

After training, the predicted labels for the test data with the respective model were collected as well as scalar model outputs. In case of the RBF, the scalar output was the output of the linear output layer. For the GLVQ and SNG we used the distances to the closest prototype of the same class as well as the smallest distance to a prototype of any other class as scalar output. We set 20, 30, or 40 as total number of prototypes/hidden neurons in all three models. In the GLVQ and SNG an identical number of prototypes per class was used. In addition to the Euclidean distance we also used weighted

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.55

Euclidean distance as an alternate distance metric where the weights are automatically adapted in the training phase.

B. Combining algorithms

The selection of combining algorithms is motivated by previous work on the combination of object detection methods presented in [7]. The same framework for fusion and evaluation of classifier outputs was used to investigate fusion of classifier results based on non-standard metrics.

For all selected fusion methods we assume, that measurement level output [8] is available. This implies that each classifier provides a measure of its confidence into the classification result. If a classifier assigns more than a single class, then for each feature vector confidence outputs for all classes are provided or must be calculated.

Often, a set of simple combiners is used. [9] provides a theoretical foundation of sum rule, median rule, maximum rule, and minimum rule which is based on Bayes' theorem. Given the outputs of R classifiers $x_1, \ldots, x_R$, the sumrule assigns those class ω_k for which

$$(1-R)P(\omega_k) + \sum_{i=1}^{R} P(\omega_k|x_i) \tag{3}$$

is maximized. $P(\omega_k)$ and $P(\omega_k|x_i)$ denote the apriori and the aposteriori probability of detection.

The Dempster Shafer theory of evidence as well as different Fuzzy approaches aim to generalize the calculation of probabilites by introducing concepts such as uncertainty, doubt, belief, fuzziness and plausibility. Fuzzy Templates are obtained by averaging the decision profiles of classifiers for all classes separately. A decision profile D is the $M \times R$ matrix whose elements are the R measurement level outputs of the classifiers for the M different classes.

$$D(x) = \begin{bmatrix} d_{1,1}(x) & \cdots & d_{1,R}(x) \\ \vdots & \ddots & \vdots \\ d_{M,1}(x) & \cdots & d_{M,R}(x) \end{bmatrix} \tag{4}$$

Therefore, two $M \times R$ matrices (decision templates) are used for a typical binary classification problem. The decision templates and the decision profiles are treated as fuzzy sets A and B. According to [10] the fuzzy measure S between two vectors of sympathy μ_A and μ_B (representing decision template and profile) is defined as

$$S(A,B) = \frac{\|A \cap B\|}{\|A \cup B\|} = \frac{\sum_{x \in X} \min\left(\mu_A(x), \mu_B(x)\right)}{\sum_{x \in X} \max\left(\mu_A(x), \mu_B(x)\right)}. \tag{5}$$

It is used to calculate the similarity between the so-called decision template and decision profile.

Another common fusion method is voting. We selected weighted majority voting because of its importance. Given a binary decision h_i which assigns 1, if a pattern x is detected and else assigns -1, the combined decision $H(x)$ is

$$H(x) = sign\left(\sum_{i=1}^{R} \alpha_i h_i(x)\right) \tag{6}$$

The different weights α_i are calculated by the following equation.

$$\alpha_i = \frac{1}{2} \ln \frac{1-\epsilon_i}{\epsilon_i} \text{ with } \epsilon_i < 0.5 \tag{7}$$

AdaBoost and Random Forests have been selected as representatives for the learning of fusion rules. They are used to learn a combining rule from the correlated or diverse decisions of the classifiers here. In previous studies both methods provided good results with no need to adapt many of their parameters prior to classifier training. For instance, it is much more time demanding to find the right kernel and its best parameter setting when using support vector machines.

Another interesting approach to classifier fusion is SCANN, which consists of the processing steps stacking, correspondence analysis, and nearest neighbor search [11]. It can be used if only abstract level outputs (class assignments) are available. If measurement level output is available (confidence values) then it must be transformed into an unique class assignment, e.g. by thresholding.

Cascaded reduction and growing of result sets (CRAGORS) changes the operating points of the classification methods and uses set operations for stepwise construction of an improved set class assignments [12]. CRAGORS-ROC is a variant which operates on ROC curves instead of precision/recall curves. CRAGORS is very similar to Boolean combination. Iterative Boolean Combination (IBC) uses all Boolean functions to iteratively include additional classifiers into an existing ensemble and to optimize the ROC curve of the resulting classifier [13].

C. Validation approach

According to the 10-fold cross-validation scheme each dataset is divided into ten disjunct subsets. Nine of them are used for training a combiner or to parametrize a combining rule. The remaining subset is used for testing its performance. This is repeated for all ten subsets to obtain average classification accuracy and the corresponding standard deviation.

The datasets consists of different representations. Typically, the real-valued outputs of all neurons of the output-layer as well as the crisp class labels are contained. The fusion algorithms are tested with all applicable inputs (e.g. crisp labels for SCANN and majority voting and real-valued features for Fuzzy templates). The results of those input features which yield the best performance are compared to the other combining algorithms.

So far, the validation approach is similar to whose presented in [1], [2]. Additionally, we applied a holdout validation to test for the influence of the sampling strategy onto the obtained results. The datasets are divided into three partitions of equal size. The first partition is used to train a set of base classifiers. For faster training and as an additional baseline testing for our divergence-based approach, Random Forest classifiers are used. Diversity is induced by randomly choosing a subset of training data from the first partition and by setting different numbers of trees for the individual Random Forest classifiers. The second partition is used to obtain realistic class predictions from the base classifiers for unseen data and to train a combining rule based on these predictions. Finally, the

third partition is used for testing the complete processing chain (apply base detectors, then apply combining method).

In addition to the cross-validation scheme the generalization performance within the combining step is evaluated with holdout testing. We systematically varied the fraction of randomly chosen samples (from base classifier predictions) for the training set and tested the accuracy of the combined predictions with the remaining samples.

III. Datasets

The hyperspectral datasets have been selected from several industrial applications where hyperspectral imaging can be used for the detection of a desired target material or defective objects for a subsequent material sorting. We deliberately chose classification tasks that showed mediocre classification accuracy on single prototype based models.

For the hyperspectral image acquisition, material samples were positioned each with a standard optical PTFE (polytetrafluoroethylene) calibration pad on a translation table. Hyperspectral images were recorded using a HySpex SWIR-320m-e line camera (Norsk Elektro Optikk A/S). Spectra are from the short-wave infra-red range (SWIR) of 970 nm to 2,500 nm at 6 nm resolution yielding a 256 dimensional spectral vector per pixel. The camera line has a spatial resolution of 320px. Radiometric calibration was performed using the vendors software package.

Five binary classification problems were considered for this publication:

1) D_1 - *Detection of aluminum within waste material (D_1):* The dataset contained samples of bulk materials from demolition sides or river excavation, where aluminum is search for recycling purposes among heavy metal and scrap materials.
2) *Classification of mature vs. immature coffee beans (D_2):* For the purpose of coffee quality control, one possible defect in raw coffee are green coffee beans that have to be detected among suitable beans.
3) *Detection of putrid hazelnuts (D_3) or fungi infested (D_4) among healthy hazelnuts:* The challenge in detection defects in hazelnuts is the fact, that defects appear in the inside of a nut. With hyperspectral imaging the change in chemical composition of the outer hull can be measurement as a secondary effect to detect inner defects.
4) *Anomaly detection on the surface of fluffed pulp (D_5):* Here the industrial application is the detection of defects on paper products for an online-inline quality control mechanism.

Each dataset contained a balanced number of 10,000 spectra per material category.

IV. Results and Discussion

Tab. I-III summarize cross-validation ensemble results. Again, as presented in [1] the RBF networks have shown the best performance. The new findings are the improvements in classification accuracy by changing the used combining method. Especially SCANN, Fuzzy Templates, and Voting are capable to further improve results of RBF network based ensembles. For the LVQ and SNG classifier ensembles only SCANN provides a significant increase in classification accuracy compared to the trained combiners AdaBoost, Random Forest, and CRAGORS which were used in previous studies.

Simple combiners such as Voting do not work for LVQ and SNG classifiers as they do for the RBF ensembles.

Tab. IV shows the ranking of each combining algorithm over all datasets and ensemble members. While the so far used combining methods AdaBoost and RandomForest are still among the top ranked, SCANN and Fuzzy Template based fusion outperform them. Among the simple combiners, only Voting performs well. However, its performance depends on the dataset as well as the choice of the base classifiers.

The diagrams of Fig. 1 visualize the accuracy gains of the different combination methods (red dots compared to the green cross) as well as the additional gains by adding the γ-divergence based classifiers into the ensembles (left compared to right column). The different starting points for the combination procedure (base classifier performance) are visible by comparing the positions of the green crosses. While classification performance of RBF-networks for dataset D_4 reaches high precision and recall values, for dataset D_3 a large gap to optimal classification exists. The clouds of red dots indicate the performances of all tested combination algorithms. It becomes obvious that the best performing methods optimize both - precision and recall while the less successful methods (especially the simple combining rules) tend to optimize only one of the two measures.

Tab. V-VII summarize the gains in terms of the average accuracy using ensemble decisions by the SCANN algorithm. By using SCANN and γ-divergence based ensembles, LVQ results match RBF results for most of the tested datasets. In all cases, the results highlight the importance of including γ-divergence based classifiers into the ensembles.

The high quality of ensemble decisions requires further investigation. The mean correlation coefficients between the base classifier decisions for datasets D_1-D_5 are 0.35, 0.36, 0.07, 0.66, and 0.18, respectively. Hence both, uncorrelated as well as correlated ensembles contribute to the observed improvements.

We systematically varied the size of the training dataset. Using Random Forests as combination method, we found that training a classifier with approximately 10% (randomly sampled) of the dataset provides a competetive result to 10-fold cross-validation (where 90% of the data are used for training). This indicates a very good generalization performance. Hence, we could expect similar results for unknown data from the same distribution.

	Datasets				
	RBF				
Fusion Method	D_1	D_2	D_3	D_4	D_5
SCANN	**1.00±0.00**	**1.00±0.00**	**0.95±0.01**	**1.00±0.00**	**1.00±0.00**
CRAGORS	0.94±0.01	0.95±0.02	0.77±0.01	0.99±0.01	0.89±0.02
CRAGORS ROC	0.92±0.01	0.93±0.02	0.77±0.02	0.98±0.01	0.88±0.02
Fuzzy Templates	**1.00±0.00**	**1.00±0.00**	**0.95±0.01**	**1.00±0.00**	**1.00±0.00**
IBC	0.97±0.01	0.98±0.01	0.85±0.01	0.99±0.01	0.97±0.01
AdaBoost	**1.00±0.00**	0.99±0.00	0.92±0.02	1.00±0.00	0.98±0.01
Random Forest	**1.00±0.00**	0.99±0.01	0.90±0.02	1.00±0.00	0.98±0.01
Maxrule	0.96±0.01	0.96±0.01	0.76±0.03	1.00±0.00	0.74±0.02
Medianrule	**1.00±0.00**	**1.00±0.00**	0.70±0.03	**1.00±0.00**	0.55±0.03
Minrule	0.96±0.01	0.96±0.01	0.76±0.02	1.00±0.00	0.74±0.03
Prodrule	**1.00±0.00**	**1.00±0.00**	0.71±0.02	**1.00±0.00**	0.69±0.02
Sumrule	**1.00±0.00**	**1.00±0.00**	0.69±0.03	**1.00±0.00**	0.56±0.02
Voting	**1.00±0.00**	**1.00±0.00**	**0.95±0.01**	**1.00±0.00**	**1.00±0.00**

TABLE I. FUSION OF THE OUTPUT LAYER OF RBF-NETWORK ENSEMBLES FOR DATASETS D_1-D_5

	Datasets				
	LVQ				
Fusion Method	D_1	D_2	D_3	D_4	D_5
SCANN	**1.00±0.00**	**0.99±0.01**	**0.82±0.02**	**1.00±0.00**	**0.98±0.01**
CRAGORS	0.87±0.01	0.81±0.02	0.66±0.03	0.90±0.03	0.77±0.01
CRAGORS ROC	0.85±0.01	0.80±0.02	0.64±0.01	0.89±0.03	0.78±0.02
Fuzzy Templates	0.99±0.01	0.97±0.01	0.64±0.02	**1.00±0.01**	**0.98±0.01**
IBC	0.93±0.02	0.94±0.01	0.71±0.03	0.96±0.03	0.89±0.01
AdaBoost	0.97±0.01	0.96±0.01	0.75±0.02	1.00±0.01	0.94±0.01
Random Forest	0.97±0.01	0.96±0.01	0.75±0.03	1.00±0.00	0.94±0.01
Maxrule	0.81±0.01	0.84±0.02	0.59±0.02	0.79±0.05	0.59±0.03
Medianrule	0.75±0.02	0.89±0.01	0.50±0.02	0.90±0.02	0.50±0.02
Minrule	0.81±0.02	0.84±0.02	0.58±0.03	0.79±0.04	0.59±0.03
Prodrule	0.99±0.00	0.99±0.00	0.51±0.02	0.95±0.02	0.52±0.03
Sumrule	0.97±0.01	0.98±0.01	0.50±0.03	0.98±0.01	0.50±0.02
Voting	0.99±0.00	0.97±0.01	0.64±0.02	0.99±0.01	**0.98±0.01**

TABLE II. FUSION OF THE OUTPUT LAYER OF LVQ-NETWORK ENSEMBLES FOR DATASETS D_1-D_5

	Datasets				
	SNG				
Fusion Method	D_1	D_2	D_3	D_4	D_5
SCANN	**0.99±0.01**	**0.96±0.01**	**0.67±0.02**	**1.00±0.00**	**0.94±0.01**
CRAGORS	0.78±0.03	0.76±0.02	0.58±0.03	0.90±0.04	0.74±0.02
CRAGORS ROC	0.78±0.02	0.73±0.02	0.56±0.02	0.85±0.05	0.72±0.01
Fuzzy Templates	0.98±0.01	0.94±0.01	0.67±0.02	1.00±0.00	0.69±0.03
IBC	0.79±0.02	0.87±0.01	0.58±0.02	0.92±0.03	0.90±0.02
AdaBoost	0.95±0.01	0.90±0.01	0.60±0.02	0.99±0.01	0.90±0.01
Random Forest	0.95±0.01	0.90±0.02	0.58±0.02	1.00±0.01	0.88±0.02
Maxrule	0.53±0.02	0.61±0.03	0.51±0.03	0.60±0.04	0.51±0.04
Medianrule	0.50±0.01	0.50±0.03	0.50±0.03	0.50±0.05	0.50±0.02
Minrule	0.53±0.02	0.61±0.02	0.51±0.02	0.60±0.04	0.51±0.02
Prodrule	0.52±0.03	0.52±0.03	0.50±0.03	0.57±0.04	0.51±0.02
Sumrule	0.50±0.02	0.50±0.02	0.50±0.02	0.50±0.03	0.50±0.02
Voting	**0.99±0.01**	0.94±0.02	0.67±0.01	**1.00±0.00**	0.69±0.03

TABLE III. FUSION OF THE OUTPUT LAYER OF SNG ENSEMBLES FOR DATASETS D_1-D_5

As described in Sec. II the datasets have been collected from the outputs of cross-validation. Hence, the output of a single base classifier is sampled from 5 different classifiers (according to 5-fold cross-validation) with the same parameters but from different training folds. As a rapid test, we implemented a similar classification and fusion framework only based on Random Forest classifiers (which are fast to train) and using both, holdout testing as well as cross-validation for the data generation step prior to learning the combining rules. We can not report any significant difference in the classification accuracy given these two sampling strategies. However, ensembles based on Random Forest base classifiers show only minor improvements in classification accuracy compared to the reported results of RBF, LVQ, and SNG classifiers. The results of the different Random Forests are much more correlated and hence do not lead to improved decisions.

V. CONCLUSION

The good results of previous tests with the selected hyperspectral datasets have been generally replicated. While it was not possible to achieve significant gains in classification accuracy by changing the used metric or adapting the model size, these two actions seem to induce a sufficient level of diversity into an ensemble of multiple neural network classifiers.

In principle 10-fold cross-validation should avoid overfitting and provide a realistic measure for the generalization performance of the classifiers. However, the obtained perfect classification results for nearly all datasets raise questions. These questions could be answered by using independent sets of spectral data for training of base classifiers, training of a combiner based on classifier outputs for an independent test

Fusion Method	Rank
SCANN	1
CRAGORS	7
CRAGORS ROC	9
Fuzzy Templates	2
IBC	6
AdaBoost	3
Random Forest	4
Maxrule	10
Medianrule	13
Minrule	12
Prodrule	8
Sumrule	11
Voting	5

TABLE IV. RANKING OF FUSION ALGORITHMS BASED ON AVERAGE RANK OVER ALL DATASETS AND FEATURE SUBSETS

	Feature Set		
	RBF		
Dataset	L_2	γ	$L_2 + \gamma$
D_1	0.95±0.01	1.00±0.00	1.00±0.00
D_2	0.96±0.01	1.00±0.00	1.00±0.00
D_3	0.76±0.02	0.92±0.02	0.94±0.01
D_4	1.00±0.00	1.00±0.00	1.00±0.00
D_5	0.87±0.01	0.99±0.00	1.00±0.00

TABLE V. FUSION OF RBF CLASSIFIERS WITH SCANN FOR DATASETS D_1-D_5

	Feature Set		
	LVQ		
Dataset	L_2	γ	$L_2 + \gamma$
D_1	0.82±0.01	0.99±0.01	0.99±0.00
D_2	0.79±0.02	0.98±0.01	0.99±0.00
D_3	0.61±0.03	0.78±0.02	0.80±0.02
D_4	0.92±0.03	1.00±0.00	1.00±0.00
D_5	0.72±0.02	0.98±0.01	0.98±0.00

TABLE VI. FUSION OF LVQ CLASSIFIERS WITH SCANN FOR DATASETS D_1-D_5

	Feature Set		
	SNG		
Dataset	L_2	γ	$L_2 + \gamma$
D_1	0.77±0.02	0.98±0.01	0.99±0.01
D_2	0.79±0.02	0.91±0.01	0.95±0.01
D_3	0.55±0.02	0.64±0.02	0.66±0.02
D_4	0.88±0.02	1.00±0.00	1.00±0.00
D_5	0.69±0.03	0.91±0.01	0.93±0.01

TABLE VII. FUSION OF SNG CLASSIFIERS WITH SCANN FOR DATASETS D_1-D_5

set, and testing its performance with another independent set of spectral data. Here, we presented first results of validation experiments which show the validity of the approach, while preparing new independent datasets and implementing an ensemble based classification module into our existing classification frameworks for further validation.

Among the evaluated algorithms for multiple classifier fusion, SCANN performs best. Consistently, it provides the best results for all datasets and all variants of ensembles.

REFERENCES

[1] U. Knauer, A. Backhaus, and U. Seiffert, "Beyond standard metrics - on the selection and combination of distance metrics for an improved classification of hyperspectral data," in *Workshop on Self-Organizing Maps (WSOM 2014)*, ser. Advances in Intelligent Systems and Computing, 2014.

[2] U. Knauer, A. Backhaus, and U. Seiffert, "Fusion trees for fast and accurate classification of hyperspectral data with ensembles of γ-divergence-based RBF networks," *Neural Computing and Applications*, vol. 26, no. 2, pp. 253–263, 2015.

[3] B. Hammer and T. Villmann, "Generalized relevance learning vector quantization," *Neural Networks*, vol. 15, pp. 1059–1068, 2002.

[4] B. Hammer, M. Strickert, and T. Villmann, "Supervised Neural Gas with general similarity measure," *Neural Processing Letters*, vol. 21, pp. 21–44, 2005.

[5] A. Backhaus, F. Bollenbeck, and U. Seiffert, "Robust classification of the nutrition state in crop plants by hyperspectral imaging and artificial neural networks," in *In Proc. 3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, Lisboa, Portugal, 2011.

[6] T. Geweniger, M. Kästner, and T. Villmann, "Optimization of parametrized divergences in fuzzy c-means," in *European Symposium on Artificial Neural Networks*, 2011.

[7] U. Knauer and U. Seiffert, "A Comparison of Late Fusion Methods for Object Detection," in *IEEE International Conference on Image Processing*, 2013, pp. 1–8.

[8] L. Xu, A. Kryzak, and C. V. Suen, "Methods of Combining Multiple Classifiers and Their Application to Handwriting Recognition," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.

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

[10] L. I. Kuncheva, "'fuzzy' vs 'non-fuzzy' in combining classifiers designed by boosting," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 6, pp. 729–741, 2003.

[11] C. J. Merz, "Using correspondence analysis to combine classifiers," *Machine Learning*, vol. 36, no. 1-2, pp. 33–58, 1999.

[12] U. Knauer and U. Seiffert, "Cascaded Reduction and Growing of Results Set for Combining Object Detectors," in *Multiple Classifier Systems*, ser. LNCS, Z.-H. Zhou, F. Roli, and J. Kittler, Eds. Springer-Verlag Berlin Heidelberg, 2013, vol. 7872, pp. 121–133.

[13] W. Khreich, E. Granger, A. Miri, and R. Sabourin, "Iterative Boolean combination of classifiers in the ROC space: An application to anomaly detection with HMMs," *Pattern Recognition*, vol. 43, no. 8, pp. 2732–2752, 2010.

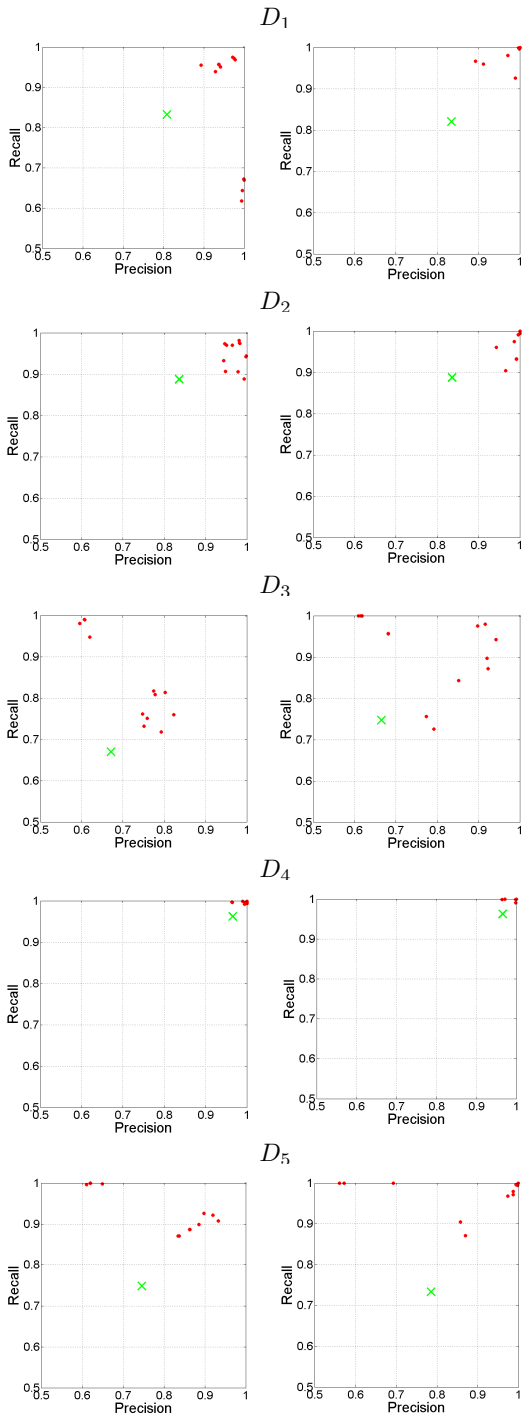

Fig. 1. Precision Recall plot for ensemble decisions (red dots) compared to the best individual classifier (green cross). The left column shows L_2-based classifier ensembles. The right column includes γ-divergence based classifiers.