Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Classifier Fusion Approaches for Diagnostic Cancer Models

Ioannis N. Dimou, Georgios C. Manikis, Michalis E. Zervakis, Member IEEE

***Abstract*— Classifier ensembles have produced promising results, improving accuracy, confidence and most importantly feature space coverage in many practical applications. The recent trend is to move from heuristic combinations of classifiers to more statistically sound integrated schemes to produce quantifiable results as far as error bounds and overall generalization capability are concerned. In this study, we are evaluating the use of an ensemble of 8 classifiers based on 15 different fusion strategies on two medical problems. We measure the base classifiers correlation using 11 commonly accepted metrics and provide the grounds for choosing an improved hyper-classifier.**

***Index Terms*—classifier fusion, classifier ensembles, diagnostic model, hyper-classifiers,** **SVMs**

I. Introduction

Multi-classifier systems have emerged from experimental evidence that we can provide better results using a collection of relatively poor-performing elementary classifiers, than by utilizing a single fine-tuned one [1].

In this paper, we consider a set of trained classifiers and we are interested in combining their outputs aiming at the highest possible accuracy. The classifiers' outputs are regarded as probabilistic, compensated for prior class probabilities using the training set's class frequencies.

We also assume that our problem consists of *C=2* classes and that both crisp *Yt* and soft *Zt* output labels are available from each base classifier.

Based on these assumptions we can choose to take into account the individual classifiers' confidence estimates and aggregate them in a manner that favors the best performing ones in a specific area of the feature-space. Alternatively we can combine using simpler schemes like majority-voting, min and max that consider all inputs are identical.

Throughout this paper we will refer to the individual classifiers operating on the dataset's features *as level1 (L1)* or *base* classifiers. Fusion schemes and classifiers operating on the outputs of the level1 classifiers will be referred to as *level2 (L2)* classifiers or hyper-classifiers.

The rest of the material is organized as follows. Section II of this paper describes in brief the main approaches in classifier fusion, along with performance and confidence evaluation and selection strategies. In section III we elaborate on the details of the application of different hyper-classifiers to biomedical datasets and the achieved results. Section IV summarizes the experimental findings and conclusions and provides insight to open problems.

II. Classifier Fusion Background

A. Taxonomy of Fusion Methods

In combining classifiers, the key objective is to obtain the most accurate feature mapping by maintaining diversity and simplicity. There are various approaches in related literature as to the available fusion techniques ([1]-[9]). They range from simple majority voting and averaging combinations to Bayesian probabilistic models and hyper-classifiers that work on the feature space composed by the soft outputs of the individual classifiers. There is also significant recent research effort in providing statistical foundations for the existing fusion schemes ([9]). Another trend is to select and use different features for each classifier as in [10].

A major discrimination between the various approaches is based on the type of input data used. Most early classifier fusion approaches used the crisp labels of the individual classifiers. They relied on schemes, such as majority voting, that extract posterior class probability statistics though counting the true and assigned labels per class. Schemes based on this approach include Behavior Knowledge Space method, Naïve Bayes combination and simple or weighted majority voting [5].

Another group of classifier fusion methods utilizes the soft outputs of the level1 classifiers. Having available continuous-valued outputs, i.e. more information per sample, these algorithms are in theory more effective. The continuous support values can represent probabilities and even convey information on the confidence that each specific learner places on its class estimate. Some of the simpler schemes belonging to this group include min, max, average and product combiners. At a more advanced level, one can apply probabilistic product, linear combiners, linear, quadratic and Fisher discriminant functions. State of the art research in this area, however, focuses on using the decision profile (*DP*) to calculate Decision Templates (*DTs*) and Dempster-Shafer membership degrees for each sample. Classical experts like neural networks, logistic classifiers and other linear or nonlinear classifiers can also

This work was supported by Biopattern, IST EU funded project (contract #508803).

I. N. Dimou, G. C. Manikis and M. E. Zervakis are with the Technical University of Crete, 73100 Chania, Crete, GREECE (e-mail: jdimou@systems.tuc.gr).

be used. The later approach however requires reshaping (unfolding) the DP to form a new output feature space. The resulting L2 feature space suffers from very concentrated distribution of the outputs around 0 and 1, which invalidates algorithms like linear discriminant functions.

B. Classifier Selection

There exists extensive literature on the subject of classifier selection ([6],[11],[12],[13]). Most of the approaches, however, focus on choosing the best performing single learning machine. In the context of classifier fusion, the weight is shifted to selecting a number of classifiers that perform optimally as a group. This implies notions of feature space coverage, diversity and combined confidence. In fact, the importance of the later aspects often shadows the actual individual performance as a decision factor. The above process should not be confused with classifier selection schemes that identify one local expert for an area of the feature space and assign every sample in that area to that specific classifier only.

One of the first metrics considered for checking classifier overlap is the Correlation Coefficient. Correlation is a well known statistical measure that can be calculated for pairs of classifiers *(i,j)* using crisp or soft labeling as an outcome. For two binary classifier outputs, it can be defined as:

$$C2_{i,j} = \frac{N^{TT} \cdot N^{FF} - N^{FT} \cdot N^{TF}}{\sqrt{N^{T*} \cdot N^{F*} \cdot N^{*T} \cdot N^{*F}}} \quad (1)$$

where
N^{TT}: # of common true cases for both classifiers
N^{TF}: # of cl.1 true cases that cl.2 marked as false
N^{FT}: # of cl.2 true cases that cl.1 marked as false
N^{FF}: # of common false cases for both classifiers

For a set of M classifiers, the averaged correlation coefficient of all pairs of classifiers is taken:

$$\overline{C2} = \frac{2}{M(M-1)} \sum_{\substack{i,j=1,\ldots,M \\ i \neq j}} C2_{ij} \quad (2)$$

Correlation coefficient with a high value corresponds to low diversity between the classifiers.

Another important measure that showed strong relation among learners is the Q-statistic. It is used for assessing the level and sign of dependency between a pair of classifiers with crisp output labeling.

$$Q2_{ij} = \frac{N^{TT} \cdot N^{FF} - N^{FT} \cdot N^{TF}}{N^{TT} \cdot N^{FF} + N^{FT} \cdot N^{TF}} \quad (3)$$

Classifiers that tend to recognize the same objects correctly will have positive values of Q. In addition, Q varies between -1 and 1, where -1 means full negative dependence and +1 full positive dependence. The higher the Q statistic value, the less diversity there is between the classifiers.

Contrary to the Q-statistic, the disagreement measure is equal to the probability that two classifiers disagree on their decisions. Such a measure is informative about the degree of correlation between the classifiers' outcome, assigning a high value to high correlation. If we are only interested in the fail cases, the Double-Fault measure gives the probability of a pair of classifiers both having wrong decision. The information about the simultaneous errors that are committed is believed to be more useful than the knowledge of when both classifiers are correct. Pairs of classifiers that provide a high value of Double Fault are highly correlated.

The Kappa-statistic was used in this sturdy as a mixture of several other correlation indices, which not only provides a measure of the degree of agreement, but it also gives information about the degree of agreement beyond chance:

$$k = \frac{2 \cdot \left(N^{TT} \cdot N^{FF} - N^{FT} \cdot N^{TF}\right)}{\left(N^{TT} + N^{TF}\right) \cdot \left(N^{FF} + N^{TF}\right) + \left(N^{TT} + N^{FT}\right) \cdot \left(N^{FF} + N^{FT}\right)} \quad (4)$$

For a set of M classifiers, the averaged kappa-statistic of all pairs of classifiers is taken. A value of kappa below 0.40 is considered to represent poor agreement beyond chance, values between 0.40 and 0.75 indicate fair agreement, and values beyond 0.75 indicate excellent agreement.

All the above-considered measures are pair-wise and reflect the relationship of each classifier couple. Usually they can be extended through some form of aggregation to L-wise measures, which can depict the diversity of the entire pool of learning algorithms.

Apart from these, there is also p-correlation assesses the agreement on misclassification and is defined as:

$$p_2 = \frac{2 \cdot N^{FF}}{N^{TF} + N^{FT} + 2 \cdot N^{FF}} \quad (5)$$

Low values of p-correlation indicate a high possibility for the classification fusion approach to be effective in performance improvement. Largely relating to information theory models, the metric of ensemble entropy shows the level of disagreement among the outputs from a set of classifiers. Entropy varies between 0 and 1, where 0 indicates that all classifier outputs are identical, and 1 indicates the highest possible diversity.

Kohavi-Wolpert variance measures the average variance from binomial distributions of the outputs for each classifier. A large value indicates major and probably useful diversity among the used models.

Having evaluated a set of 20 classifiers using the above metrics, we concluded that for the more difficult breast cancer dataset (ds3) the trained models showed high diversity (Q:[0.22,0.46], p-corr: [0.22,0.37]). The other two datasets produced highly correlated classifiers (Q:[0.63,0.92], p-corr: [0.55,0.81] for ds2). This pair of metrics reflects the trend of the diversity values reported by the other metrics described in section II.B. In these cases, we were forced to reduce the classifiers' pool to a subset of 8 classifiers for satisfactory ensemble performance.

C. Issues of Performance Evaluation

Evaluation of the results of a multi-classifier system can

be done in a number of ways depending on the problem. For this particular application we chose to use a cross-validated accuracy metric, which tends to be the standard in related work. We provide the mean test-set accuracy for each single classifier after performing 100 randomized runs.

For the hyper-classifiers, we used the accuracy of the overall system as performance measure. It should be stressed here that a different pool of more efficient experts could upgrade L1 accuracies, but at the same time might leave a narrower margin of improvement for the hyper-classifiers. All hyper-classifiers used the same *DP*, which consists of the outputs of the optimized level1 experts. Wherever hard labels were needed, a second hard-*DP* matrix containing these labels was extracted. The AUC performance metric could only be defined for L2 classifiers that work on DP feature space and have a selectable threshold. Therefore, for consistency we used only the cross-validated accuracy as a measure of performance for all hyper-classifiers.

III. Application to Biomedical Datasets

We present an experimental evaluation of the above state-of-the-art classifier fusion algorithms in a practical problem. More specifically, we applied the 8 individual classifiers that were chosen from the entire pool, to one artificial and two biomedical datasets. The aim was to provide accurate diagnosis of different types of cancer based on the available predictors.

Dataset 1 (ds1) consists of a set 8 features of an artificially created banana-shaped dataset along with an indicator of two-class membership. A total of 1000 samples were produced and split in a stratified manner in one training set consist in of 700 samples and one test-set with the remaining 300 samples. Each case set has a 10% prior distribution of positives.

The preprocessing for dataset1 includes standardization of the datasets' variables to the [0,1] range. Individual classifier optimization was not extensive, as this work focuses on L2 comparison and not L1 performance fine-tuning. Stratified 100-fold cross validation was applied to the base classifiers to improve performance estimates and ensure better generalization capability. For each run the dataset is split to training (70%) set and test (30%) set with proportional class probabilities (~10% positives).

The second and third datasets are obtained from the UCI Repository of Machine Learning Datasets [15]. Data found in these two sets describes a large number of cases of brain and breast cancer patients from Germany and Yugoslavia.

Both datasets 2 and 3 have normalized predictors in the [0,1] range and are available in 100x stratified randomizations. The dimensionality and features of the three used datasets are shown in Table I.

TABLE I
THE BENCHMARK DATASETS

	b-shaped (ds1)	brain (ds2)	breast (ds3)
# features	8	20	9
# train patterns	700	700	200
# test patterns	300	300	77
% of positives	10%	30%	29%

Several classifiers with different parameters were applied to each problem and a selection process was carried out based on the metrics described in section II.B. As a result 8 classifiers were chosen to be utilized in the hyper-classifier models. The classifiers used include an LS-SVM with rbf kernel, an LS-SVM with linear kernel, a linear and a quadratic distance classifier, a Naive Bayes classifier, a Probabilistic Neural Net, a Radial basis neural network mapping and a Fisher discriminant function.

The performance of the eight base classifiers in terms of accuracy is shown in Table II.

TABLE II
COMPARISON OF L1 CLASSIFIERS' PREFORMANCE

L1 classifier	Accuracy		
	ds1	ds2	ds3
Linear LS-SVM	0.912	0.701	0.714
RBF-kernel LS-SVM	0.933	0.703	0.714
LDC	0.796	0.627	0.713
QDC	0.823	0.662	0.695
Naïve Bayes Classifier	0.818	0.727	0.734
Prob. Neural Net	0.914	0.832	0.713
Backprop Neural Net	0.834	0.747	0.688
Fisher discriminant	0.784	0.633	0.721

As seen in the above table, datasets ds2 and ds3 pose greater difficulty but gained more from the application of the hyper-classifier system than ds1. In fact, improvement margins are slimmer for the later. This difference can be explained in part due to the reduced dimensionality and less available features in these problems. The confidence intervals of each method are calculated by sampling a Bernoulli distribution with parameter p equal to the respective accuracy, as illustrated in Fig.1.

Using the soft and crisp labels obtained by the set of L1 classifiers, the *DP* of each sample in the dataset was constructed.

The classifier fusion strategy includes the 8 methods analyzed in section II.A plus 5 trainable classifiers. The majority-voting, BKS and "Naive"-Bayes fusion algorithms were implemented using the crisp L1 labels. The DPs' soft labels were used to directly implement min, max, average, and product fusion schemes. Additional training was needed to apply the more elaborate probabilistic-product, Dempster-Shafer and Decision Templates hyper classifiers. After reshaping the L1 outputs into a new L2 feature space, we were also able to employ LDC, QDC, a Probabilistic Neural Net and 2 SVMs as hyper-classifiers.

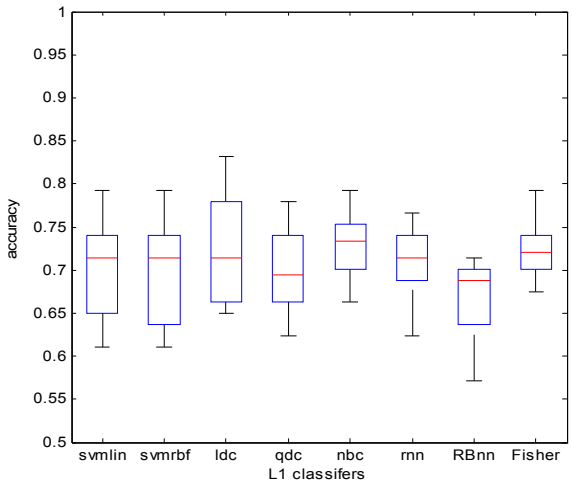

Fig. 1. L1 classifiers' confidence intervals for ds3

The comparative results of the L2 classifiers are shown in Table III. The performance of the simpler non-trainable combiners like min, max, average and product appears high in comparison to the more elaborate schemes. However, repeated runs of the entire evaluation process on all datasets have shown that DTs and DS are in general more robust.

Another observation is that LDC and QDC fail to produce satisfactory results as hyper-classifiers, most likely because the L2 feature set is not normally distributed. At the same time, the Support Vector Machines and Neural Network trained on the L2 feature set give promising results. The 2 SVMs show nearly identical performance suggesting that the feature space has at least partially linear boundaries. Apart from this, their ability to map the Decision Profile to a better overall accuracy outperforms even some of the information-theoretic hyper-classifiers (BKS, DS).

TABLE III
COMPARISON OF L2 CLASSIFIERS' PERFORMANCE

L1 Output	L2 Classifier	Accuracy		
		ds1	ds2	ds3
hard	Majority Vote	0.880	0.796	0.710
	Naive Bayes	0.836	0.632	0.580
	BKS	0.960	0.802	0.750
soft	Minimum	0.781	0.803	0.635
	Maximum	0.884	0.863	0.860
	Average	0.960	0.897	0.882
	Product	0.897	0.881	0.871
	Probabilistic Prod.	0.863	0.864	0.682
	L2 LDC	0.682	0.536	0.618
	L2 QDC	0.687	0.632	0.644
	L2 Naïve Bayes	0.711	0.855	0.786
	L2 Prob. Neural Net	0.821	0.855	0.786
	L2 linear LS-SVM	0.878	0.702	0.648
	L2 rbf LS-SVM	0.885	0.732	0.803
	Decision Templates	0.891	0.850	0.698
	Dempster-Schafer	0.903	0.841	0.762
	mean	0.845	0.784	0.732

Finally, a noteworthy improvement is observed in the ds3 "difficult" dataset where the L1 classifiers' diversity was high and the corresponding single-model accuracies moderate.

IV. Conclusions

This paper demonstrates how a variety of different classifier fusion techniques can be used to augment the diagnostic performance of individual models in the context of practical biomedical applications. As seen in the previous analysis, the performance of hyper-classifiers varies largely depending on the nature of the individual L1 outputs and the coverage of the feature space. Having homogenized inputs and properly selected components, a classifier ensemble can outperform the best single expert and provide results that are more robust. The integration and optimization of these ensembles into a single system is still an open problem.

References

[1] L. I. Kuncheva, J. C. Bezdek, R. P. W. Duin, "Decision templates for multiple classifier fusion: An experimental comparison", Pattern Recognition, 34, (2), 2001, 299-314.
[2] H. Altincay, "On naive Bayesian fusion of dependent classifiers.", Pattern Recognition Letters, vol. 26, pp. 2463-2473, 2005.
[3] K. Tumer, J. Ghosh, "Classifier combining: Analytical results and implications", AAAI 96 - Workshop in Induction of Multiple Learning Models, 1995.
[4] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24, No. 2, February 2002.
[5] D. Ruta, B.Gabrys, "an overview of classifier fusion methods", Computing and Information Systems, 7 (2000) p.1-10.
[6] K. Woods, W. Ph. Kegelmeyrer, K. Bowyer, "Combination of multiple classifiers using local accuracy estimates", IEEE transactions on pattern analysis and machine intelligence, vol. 19, no. 4, april 1997.
[7] G. Fumera, F. Roli, "Linear combiners for classifier fusion: some theoretical and experimental results", Proc. Int. Workshop on Multiple Classifier Systems (LNCS 2709), Springer, Guildford, Surrey, 2003, pp. 74--83.
[8] L. I. Kuncheva, L. C. Jain, "Designing classifier fusion systems by genetic algorithms.", IEEE Transactions on Evolutionary Computing, USA, Vol. 4, No. 4, November 2000, pp. 327-336.
[9] J. Kittler, M. Hatef, R. P. W. Duin, J. Matas, "On combining classifiers", IEEE transactions on pattern analysis and machine intelligence, vol. 20, no. 3, March 1998
[10] K. Chen, L. Wang, H. Chi, "Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification", International Journal of Pattern Recognition and Artificial Intelligence, 11(3), 1997, pp. 417-445.
[11] D. Ruta, B. Gabrys, "Classifier selection for majority voting", Information Fusion 6: 63-81, 2005.
[12] K. Goebel, W. Yan, "Choosing classifiers for decision fusion", Proceedings of the Seventh International Conference on Information Fusion, vol. I, pp. 563-568, 2004.
[13] L. I. Kuncheva, "Switching between selection and fusion in combining classifiers: an experiment.", IEEE trans. on systems, man, and cybernetics—part b: cybernetics, vol. 32, no. 2, April 2002.
[14] C.A. Shipp, L.I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers.", Information Fusion, 3 (2), 2002, 135-148
[15] D.J. Newman, S. Hettich, C.L. Blake, C.J Merz, "UCI Repository of machine learning databases", [www.ics.uci.edu/~mlearn/] (1998)