*Proceedings of SPIE Defense and Security Symposium: Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications 2005, pp. 180-191, 2005.*

# Using Correlation-Based Measures to Select Classifiers for Decision Fusion

Kai Goebel[*] & Weizhong Yan[†]
GE Global Research

## ABSTRACT

This paper explores classifier fusion problems where the task is selecting a subset of classifiers from a larger set with the goal to achieve optimal performance. To aid in the selection process we propose the use of several correlation-based diversity measures. We define measures that capture the correlation for n classifiers as opposed to pairs of classifiers only. We then suggest a sequence of steps in selecting classifiers. This method avoids the exhaustive evaluation of all classifier combinations which can become very large for larger sets of classifiers. We then report on observations made after applying that method to a data set from a real-world application. The classifier set chosen achieves close to optimal performance with a drastically reduced set of evaluation steps.

Keywords: Classification; Diversity; Correlation; Accuracy; Decision fusion

## 1. BACKGROUND

In classification problems one attempts to categorize an input vector into at least one of several classes. Where reaching a particular performance goal cannot be accomplished with any particular classifier alone, classifier fusion holds the promise to overcome that ceiling. Since the indiscriminate fusion of all classifiers does in fact not lead to optimal performance (although it may increase the performance above any individual classifier's performance), it is an important task to properly choose the right classifier subset to be fused. While a straightforward answer might seem to exhaustively search the universe of classifiers with fusion performance as the objective function, this turns out to be a potentially rather expensive method. Roli et al.[1] describe the "overproduce and choose" paradigm (in addition to associated selection methods) which assumes one can produce a set of classifiers fit for fusion. Kuncheva and Jain[2] describe the use of GAs to get a handle on the same issue.

Instead of directly using fusion performance as the search objective function, several measures have been proposed for quick quantification of the performance of a group or pair of classifiers. Although some degree of confirmatory information may be desired between classifiers, it is the complementary information that gives classifier fusion a chance to be successful. That is to say, the success of classifier fusion is not guaranteed and depends, among other things, on a successful aggregation method. Classifier fusion will then encounter its own performance ceiling. Krogh & Vedelsby[3] define diversity as ambiguity that is the variation of the output of ensemble members averaged over unlabeled data. Intuitively, diversity is the potential of a set of classifiers to rise above the performance ceiling set by the best classifier. This can only be true to the degree to which classifiers disagree when at least one of them gets the wrong answer. Kuncheva & Whitaker[4] summarize 10 different measures to quantify diversity of a group (>=2) of classifiers. Diversity, as a measure, has been used for selecting ensembles in design of multiple classifier systems[1] and for evaluating and selecting classifiers for a distributed meta-learning system[5]. Diversity has also been used for feature selection for ensembles[6]. Correlation that adversely affects the performance of classifier fusion is another measure. Petrakos et al.[7] proposed a method for classifier correlation analysis for two classifiers. Besides overproducing and then downselecting desired classifiers, one can pick a number (and types) of classifiers while trying to ensure a diverse output. In this paper we explore the former approach. More specifically, the paper is concerned with choosing classifiers from a large pool of classifiers for classifier fusion to achieve classification performance as close to the

[*] goebelk@research.ge.com; phone: (518) 387-4194; fax: 518 387-6104; http://best.me.berkeley.edu/~goebel/kai.html; GE Global Research Center, K1-5C4A, One Research Circle, Niskayuna, NY 12309, USA
[†] yan@research.ge.com; phone: (518) 387-5704; fax: (518) 387-6104; http://www.rpi.edu/~yanw/; GE Global Research Center, K1-5B34B, One Research Circle, Niskayuna, NY 12309, USA

optimal performance as possible while at the same time avoiding the exhaustive evaluation of all possible classifier combinations.

## Classifier performance evaluation

We consider classifier problems where a feature vector $x \in \Re^p$ is to be labeled into one or more of $c$ classes. In order to achieve high overall performance of the classification function, the performance of each individual classifier has to be optimized prior to using it within any fusion schemes. That is, the fusion scheme will be able to improve the overall classification result relative to the performance of the individual classifiers. If several classifiers with only marginal performance are being used, the results cannot necessarily be expected to reach the high performance sought if practical considerations such as computational constraints in an actual implementation are being factored into the selection process (that is, we discount the case where an exceptionally large number of classifiers might accomplish the same performance requirements). On the other hand, if several classifiers are used that work exceptionally well, any further gains will be exceedingly hard to accomplish because opportunity for diversity is diminished. Individual classifier optimization can be performed by selecting appropriate parameters and – where applicable – structure that govern the performance, in addition to the appropriate choice of features.

After design a confusion matrix $\boldsymbol{M}$ can be generated for each classifier using labeled training data[8]. The confusion matrix lists the true classes $c$ versus the estimated classes $\hat{c}$. Because all classes are enumerated, it is possible to obtain information not only about the correctly classified states ($N^{00}$ and $N^{11}$), but also about the false positives ($N^{01}$) and false negatives ($N^{10}$). The top-left entry of the confusion matrix is dedicated to the normal case $N^{00}$. The first row – except the first entry – contains the $N^{01}$. The off-diagonal elements – except the first row – contains the $N^{10}$. Sometimes a further distinction is made between false negatives and false classifieds where the false classifieds are defined to be the off-diagonal elements of the confusion matrix except the first row and the first column. A typical n-class confusion matrix $\boldsymbol{M}$ is shown in Fig. 1 a) for n=4. This confusion matrix collapses to the binary case where n=2 as shown in Fig 1. b)

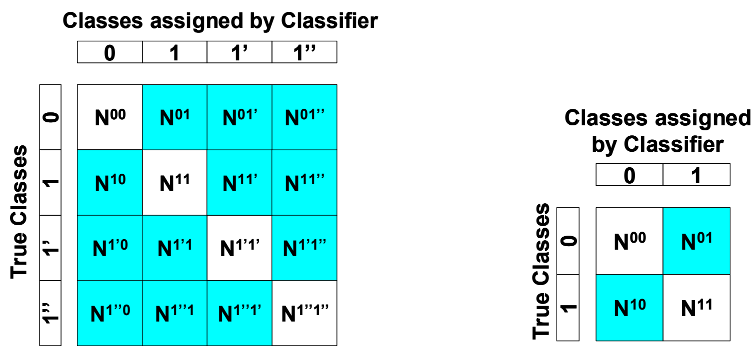


**Figure 1 – a) n-class confusion matrix; b) 2-class confusion matrix**

Continuing with the binary case, from the confusion matrix of each classifier, the false positive (FP) error, the false negative (FN) error, the total error rate (TER), and the total success rate (TSR) can be calculated for the classifier. These error rates are defined as in Equations 1 – 4. The total error rate (TER) or the total success rate (TSR) is typically used as a simple measure for overall performance of a classifier:

$$FP = \frac{N^{01}}{N^{00} + N^{01}} \tag{1}$$

$$FN = \frac{N^{10}}{N^{10} + N^{11}} \tag{2}$$

$$TER = \frac{N^{01} + N^{10}}{N^{00} + N^{11} + N^{01} + N^{10}} \tag{3}$$

$$TSR = 1 - TER \tag{4}$$

A performance requirement often times encountered in practice is the Neyman-Pearson criterion where a particular performance goal is stated as a particular minimum true positive rate or true negative rate. The goal is then to optimize the complementary performance measure (i.e., FN or FP, respectively) and, since the terms $N^{01}$ or $N^{10}$ become constants, the TER or TSR become proportional to the FN or FP.

## CLASSIFIER CORRELATION

Fusion performance depends on a number of different factors. These include the fusion method, the performance of the individual classifiers and the correlation between the different classifiers. Research is ongoing to determine the interplay of these components. In this paper, we are concerned mostly with the latter of the three components. Studies have shown[7] that an increased degree of classifier correlation may adversely affects the performance of the subsequent classifier fusion. Clearly, if two classifiers agree everywhere, fusing these two classifiers will not achieve any accuracy improvement no matter what fusion method is used. In classifier fusion design, classifier correlation analysis is, therefore, equally important as the classifier performance analysis.

### Classifier correlation analysis

Petrakos et al.[7] describe a classifier correlation analysis for two classifiers. Based on the classifier outputs on the labeled training data, a 2x2 matrix ***N*** can be generated for each classifier pair as shown in Fig. 2. The off-diagonal numbers directly indicate the correlation degree of the two classifiers. The smaller the two off-diagonal numbers are, the higher the correlation between the two classifiers will be. The proportion of specific agreement which we call here the correlation, $\rho_2$, is defined in[7] as

$$\rho_2 = \frac{2 \times N^{FF}}{N^{TF} + N^{FT} + 2 \times N^{FF}} \tag{5}$$

where $N^{TT}$ implies that both classifiers classified correctly, $N^{FF}$ means both classifiers classified incorrectly, $N^{TF}$ represents the case of the 1st classifier classified correctly and 2nd classifier classified incorrectly, and $N^{FT}$ stands for the 2nd classifier classified correctly and 1st classifier classified incorrectly as further shown in Fig. 2. In order for classifier fusion to be effective in performance improvement, the correlation, $\rho_2$, has to be small (low correlation).

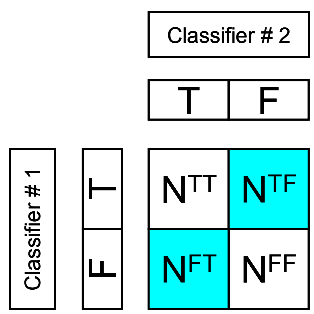


**Figure 2 – Correlation Analysis Matrix**

Note that one fundamental assumption of this approach is the availability of classifier labels.

Consider the output of 2 classifiers as enumerated in Table 1.

**Table 1: Results from experiment for 2 classifiers**

| Output classifier 1 | Output classifier 2 |
|---|---|
| T | T |
| T | F |
| F | T |
| T | F |
| F | F |

| | |
|---|---|
| F | F |
| T | F |
| F | T |
| T | T |
| T | T |
| T | T |
| T | T |
| T | F |
| T | T |
| T | T |
| F | T |

The calculation of $\rho_2$ yields $\rho_2$ =0.36. Had classifier 2 been completely redundant to classifier 1, the correlation would have been $\rho_2$ =1

**n-Classifier Correlation Analysis**

We proposed an extension of the 2-class correlation coefficient to n different classifiers $\rho_{n1}$[8]. The notion that redundancy is described by the individual true and false answers of the classifiers is retained from the 2-class correlation analysis. The larger the $\rho_{n1}$–correlation, the larger the redundancy. In particular, the $\rho_{n1}$–correlation goes to zero if the individual incorrect answers are disjoint for all answers. That implies that there is always at least one correct answer from some classifier for any case available. It also implies that there is at least one incorrect answer but the point in this analysis is to identify the potential of correct classification that hinges on the correct answer. The $\rho_{n1}$–correlation coefficient gets larger as the number of wrong answers are the same for many answers. $\rho_{n1}$ is one when all classifiers agree. Let $N^f$ be the number of experiments where all classifiers give a wrong answer, $N_i^c$ be the number of experiments with combinations of correct and incorrect answers; $c$ is the combination of correct and incorrect answers (for 2 classifiers: $c \in \{wr, rw\}$; for 3 classifiers: $c \in \{wwr,\ wrw,\ rww,\ wrr,\ rwr,\ rrw\}$, etc.); $n$ is the number of classifiers. The $\rho_{n1}$–correlation coefficient is then

$$\rho_{n_1} = \frac{nN^f}{\sum_{i=1}^{2^n-2} N_i^c + nN^f} \tag{6}$$

If $N$ is the number of experiments and $N^r$ is the number of experiments for which all classifiers had a right answer, equation 6 can be rewritten[8] as

$$\rho_{n_1} = \frac{nN^f}{N - N^f - N^t + nN^f} \tag{7}$$

Although we achieved good results with this measure[9], $\rho_{n1}$ may appear overly complicated. Alternatively, a somewhat simpler measure $\rho_{n_2}$ is introduced as

$$\rho_{n_2} = \frac{N^f}{N - N^t} \tag{8}$$

Measure $\rho_{n2}$ drops the term $(n-1)N^f$ from the denominator and applies a different scalar to the overall equation. Both, measure $\rho_{n_1}$ and $\rho_{n_2}$, use the joint incorrect class output as the central term. Assuming 2 classifiers with a 50% correct classification rate, and varying the number of cases in agreement, both measures behave exactly the same as illustrated in Figure 3 where measures 1 and 2 are completely overlapping. Indeed, if there is no joint incorrect output, both measures goes to zero, irrespective of the joint correct classifier output.

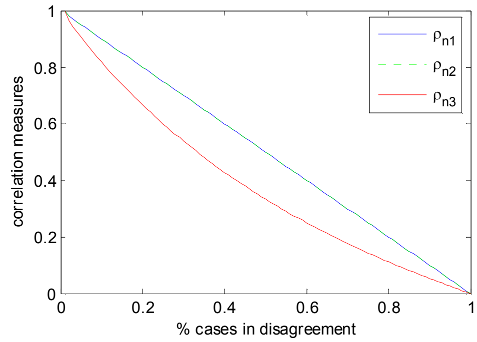


**Figure 3 – Correlation measure as a function of case disagreement**

This may actually not be a desired property. Consider a pathological case where the first classifier again has a 50% correct classification rate (as in the example earlier) but now the second classifier is ordered such that it first exhausts all jointly false cases while the total number of cases in disagreement was still rising. Then measures $\rho_{n1}$ and $\rho_{n2}$ would drop to zero before all cases are evaluated since the numerator is zero. To address this concern, we explore a second alternative $\rho_{n3}$ that is defined by

$$\rho_{n_3} = \frac{N^t + N^f}{N} \tag{9}$$

Instead of the ratio of jointly false calls over the remainder of all calls minus the correct calls (as in equation 8) $\rho_{n3}$ uses the ratio of the sum of all agreement cases (irrespective of good or bad calls) over all cases.

All measures have the property of zero correlation where there is no agreement. They also attain the value one where complete agreement exists (cf Figure 3). It has to be noted that the calculation of the $\rho$–correlation factors can be performed on multi-class scenarios as well because the factor is only concerned with the correctness of the outcome. However, without loss of generality, we will continue with the 2-class problem. Using the measures on the classifier output example of Table 1, the results becomes $\rho_{n1}$=0.3636; $\rho_{n2}$=0.5625; $\rho_{n3}$=0.2222. Now consider a 3-classifier example which is the same as the previous 2-classifier example except that a third classifier was added that gets the answer wrong in 50% of the cases as shown in Table 2.

**Table 2: 3-classifier output example**

| Output classifier 1 | Output classifier 2 | Output classifier 3 |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 0 | 1 |
| 0 | 1 | 0 |
| 1 | 0 | 1 |
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |
| 1 | 1 | 0 |
| 0 | 1 | 1 |

The calculation of $\rho_n$ yields: $\rho_{n1}$ = 0. 2000; $\rho_{n2}$ = 0. 2500; $\rho_{n3}$ = 0. 0769. Note that, although the newly added classifier has poor performance, its addition reduces the overall correlation of the classifier assembly. To reiterate, this may not guarantee a fusion output that improves overall performance, but it enables the desired fusion output.

It is interesting to note that none of the $\rho$–correlation measures record correlation with any particular classifier (for n>2) but with a set of classifier only. For illustrative purpose, consider the simplistic cases shown in Table 3 and Table 4[8]:

**Table 3: Output for 3 classifiers**

| Output classifier 1 | Output classifier 2 | Output classifier 3 |
|---|---|---|
| T | F | F |
| F | T | F |
| F | T | T |
| T | T | T |
| F | F | F |

The $\rho$–correlations are $\rho_{n1}$ = 0.5, $\rho_{n2}$ = 0.4, $\rho_{n3}$ = 0.25,

**Table 4: Output for 3 classifiers with different output for 3rd classifier**

| Output classifier 1 | Output classifier 2 | Output classifier 3 |
|---|---|---|
| T | F | T |
| F | T | T |
| F | T | F |
| T | T | T |
| F | F | F |

The $\rho$–correlations are $\rho_{n1}$ = 0.5, $\rho_{n2}$ = 0.4, $\rho_{n3}$ = 0.25. Obviously the third classifier is different in the two example cases above. However, the degree of correlation is the same because it does not matter whether it is correlated to the first or to the second classifier. Rather it is only relevant that it correlated to the combination of the first two classifiers. Although all measures scale between zero and one, they do this at different rates.

## CLASSIFIER SELECTION

The $\rho$–correlation coefficient can be used for different purposes such as classifier selection, classifier simulation, and within the fusion algorithm itself. We discuss here only the issue of classifier selection and refer for some initial thoughts on classifier simulation and fusion estimate refinement to Goebel et al. [8].

As mentioned, classifier selection should be carried out such that the least redundancy is maintained. First, one needs to select an appropriate performance measure which is typically comprised of the (possibly weighted) false positives, false negatives, and false classified. For a 2-class classifier, the TER as introduced in Eq. 4 can be used. Then, assuming a suitable set of classifiers is available, the best performing classifier is chosen. Next, the classifier with lowest joint correlation will be added. Note that this does not imply that the two best performing classifiers are fused. This process is repeated until the desired number of classifiers has been reached. When there is no classifier limit, the lowest $\rho$-correlation is an indication of best classifier performance. Note that in practice “noise” effects (possibly due to partially overlapping non-linear decision boundaries of the classifiers) are observed that mask a unique identification of the best set. To find the best set, it has been proven advisable to apply a “filter” which can be accomplished by continuing past the apparent first increase to see whether the pattern holds and then choosing the best set.

This method assumes that there is some inherent advantage in using the best classifier in the fusion scheme. This seems to make sense intuitively although it is acknowledged that, theoretically, the performance of several non-optimal classifiers may in some cases outperform a set of classifiers that includes the best classifier. We continue without further proof of that assertion.

Consider now the following example where classifier 3 is completely redundant to the second classifier as enumerated in Table 5.

**Table 5: 3rd classifier added**

| Output classifier 1 | Output classifier 2 | Output classifier 3 |
|---|---|---|
| T | T | T |
| T | F | F |
| F | T | T |
| F | F | F |
| F | F | F |

The $\rho$–correlation of classifier 1 and classifier 2 is $\rho_{n1}$=0.67. The joint $\rho$–correlation of the three classifiers is $\rho_{n1}$ =0.75, i.e., the $\rho$–correlation increased. This gives us a quantified measure for rejecting the third classifier. Had the third classifier instead been as shown in Table 6, the index would have been $\rho_{n1}$ =0.5; i.e., the $\rho$–correlation decreased.

In summary, the selection methods proposes to start with the best classifier as the base classifier, then assess the pairwise correlation with the remaining classifiers, choose the classifier with the lowest correlation, then assess the correlation of that set with the remaining classifiers, and so on. We argue that this will lead to a successful set of classifiers that is better than using a brute-force approach that requires the performance assessment of all permutations. This assertion will be examined using real-world data in the application example below.

**Table 6: Different 3rd classifier**

| Output classifier 1 | Output classifier 2 | Output classifier 3 |
|---|---|---|
| T | T | T |
| T | F | F |
| F | T | T |
| F | F | T |
| F | F | F |

## APPLICATION TO CLASSIFIERS WITH CONTINUOUS OUTPUT

We consider a suite of classifiers that produces continuous output. Specifically, 10 classifiers were designed to tackle defect detection for inspection data. A host of several hundred features was available from which smaller sets were selected for the individual classifiers using a genetic algorithm driven selection process[10]. The classifiers were chosen to be all feedforward neural nets. The use of different input features to different networks and varied network configuration attempts to ensure a reasonable diversity (although using a set of classifiers each employing a different technique might be provide even more diversity) . The accuracy was calculated subject to the Neyman-Pearson criterion where the true positive rate was set fixed at TPR=98%. In this specific application, the classification task is a 2-class problem where one of the classes can be broken down into several sub-classes. The fusion was performed using a simple averaging scheme.

The question we are trying to answer is which classifiers should be fused? Let classifier 1 be the classifier with the best performance. Without fusion, we would use just this classifier and be done. With fusion, we would like to add more classifiers to the scheme but do not know in advance which set of classifiers to use. In the example discussed here, there are 1023 different combinations but only one gives highest performance. Indeed there is a set of classifier combinations that does reasonably well but there is also a set of combinations that stands out (see Figure 11). Remarkably, the best performing combinations do not use all classifiers. Rather, the fusion performance drops when a saturation point is surpassed (for the given fusion method and the set of classifiers at hand). Figure 4 shows all possible combinations.

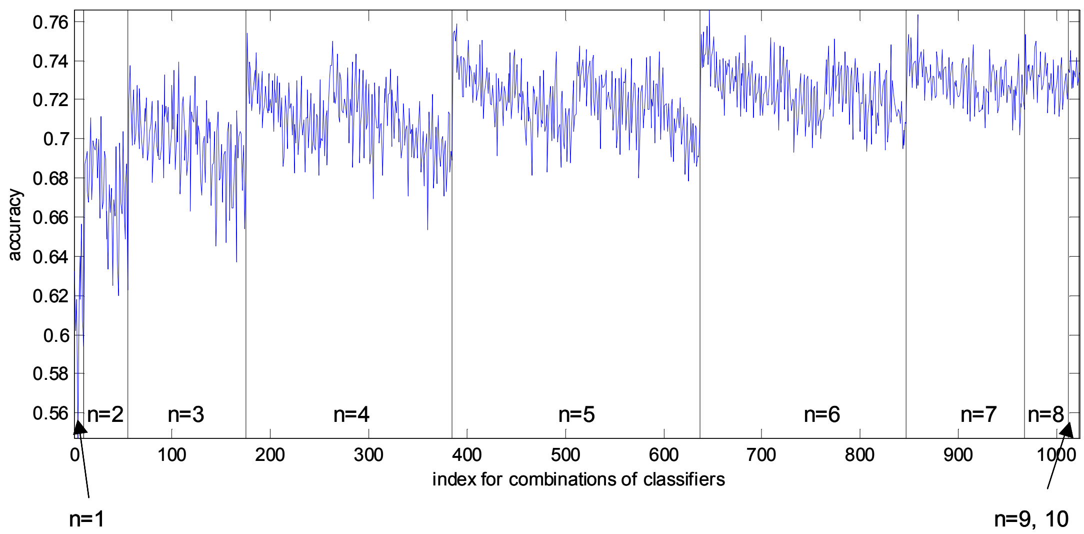


**Figure 4 – Accuracy for all classifier combinations**

The accuracy varies from 62% TPR to 76% TPR. From left to right, first all individual classifier performance is shown, next all sets to 2 classifiers (n=2), next the permutation of 3 classifiers (n=3), and so forth. As a general trend, it appears that it is beneficial to use more classifiers in the fusion scheme. However, within a set of n classifiers to be fused, there are (what one might consider considerable) differences in accuracy. Moreover, the increase in accuracy is not linear. Rather, incremental improvement becomes smaller as the number of classifiers to be fused increases. It eventually seems to approach saturation although the increase might continue for more classifiers than the maximum number considered here. One can also observe within a set of classifiers a performance drop which may have to do with the a priori ordering of the classifiers. This would support the notion that it is beneficial to fuse well performing classifiers.

Another important take-away is that the maximum performance is with a configuration of 6 classifiers, followed by a particular combination of 7 classifiers. While it is not clear how robust these results are, it is nonetheless interesting to note that the best performance is not with all 10 classifiers. Indeed, there are 169 combinations of sub-10 classifier sets that produce better performance than the 10 classifier set.

In Goebel and Yan[9], we explored that notion that it is reasonable to use the best classifier as the base classifier and assess $\rho$-correlation for pairs of the best classifier with the remaining classifiers. In contrast to that study, (which used a slightly different definition of the correlation) we focus here on the $\rho_{n1} - \rho_{n3}$ measures. Similar in both studies is the hypothesis that the correlation is a reasonable measure for fusion performance. Fig. 5 shows the pairwise accuracy for the classifiers under consideration

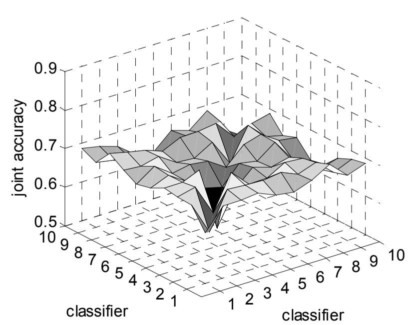


**Figure 5 – Fusion accuracy for classifier pairs**

Fig. 6 shows the pairwise correlation of the classifiers for the $\rho$-measures.

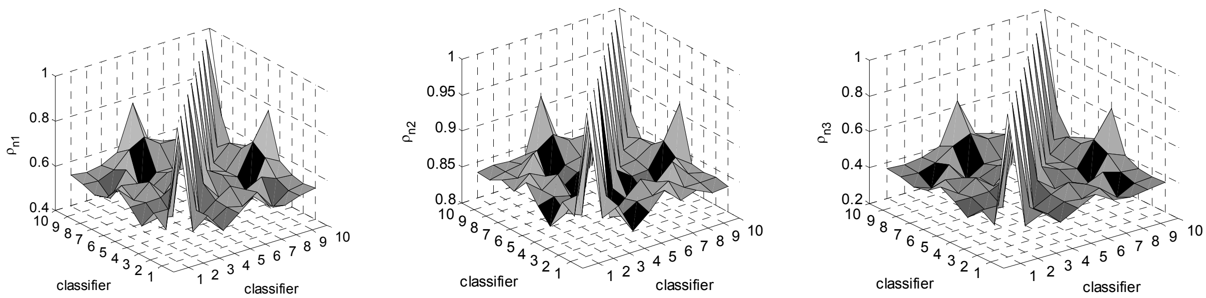


**Figure 6 – Correlation $\rho_{n1} - \rho_{n3}$ of classifier pairs**

To assist in the visual comparison, the correlation is plotted in Fig. 7 in reverse scale, i.e., 1- $\rho$

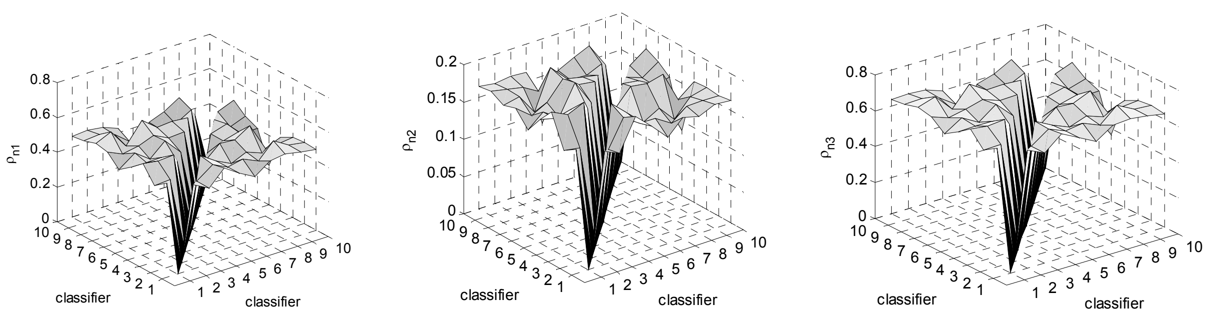


**Figure 7 – Correlation with reverse scale 1-$\rho$**

**Table 7: $\rho$ measure and statistical correlation to accuracy for pairwise classifiers**

| Measure | Statistical Correlation |
|---|---|
| $\rho_{n1}$ | 0.74 |
| $\rho_{n2}$ | 0.66 |
| $\rho_{n3}$ | 0.72 |

The statistical correlation between the measures and the pairwise accuracy is reasonably high for all $\rho$ measures. It is summarized in Table 7. In comparison, the statistical correlation between the ambiguity measure $V^3$ and the pair-wise accuracy is slightly lower, $c$=0.65.

Extending the fusion to sets of 3 classifiers, the combination 2, 5, 8 gives the highest accuracy. The statistical correlation is summarized in Table 8.

**Table 8: $\rho$ measure and statistical correlation to accuracy for combinations of 3 classifiers**

| Measure | Statistical Correlation |
|---|---|
| $\rho_{n1}$ | 0.80 |
| $\rho_{n2}$ | 0.69 |
| $\rho_{n3}$ | 0.74 |

Figure 8 – 10 show the correlations $\rho_{n1}$, $\rho_{n2}$, and $\rho_{n3}$ for the combinations. To visualize the similarity with the fusion accuracy, the correlations are displayed using 1-$\rho_{nx}$.

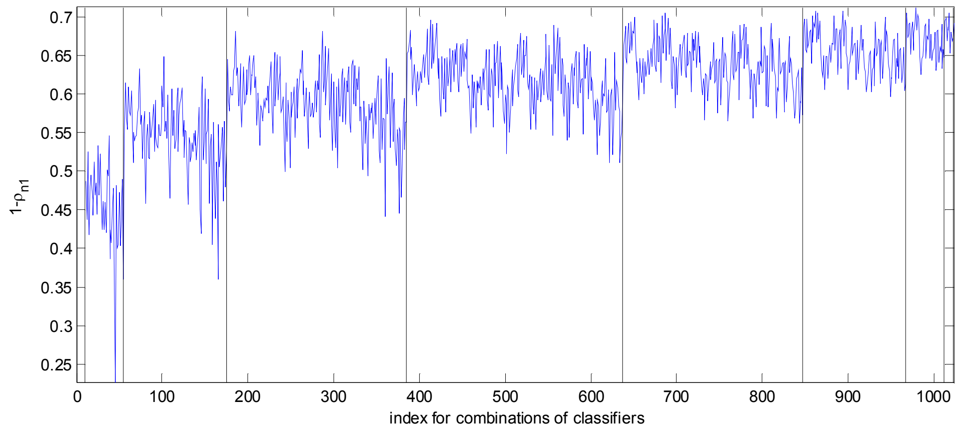


**Figure 8 – Correlation $\rho_{n1}$ for all classifier combinations.**

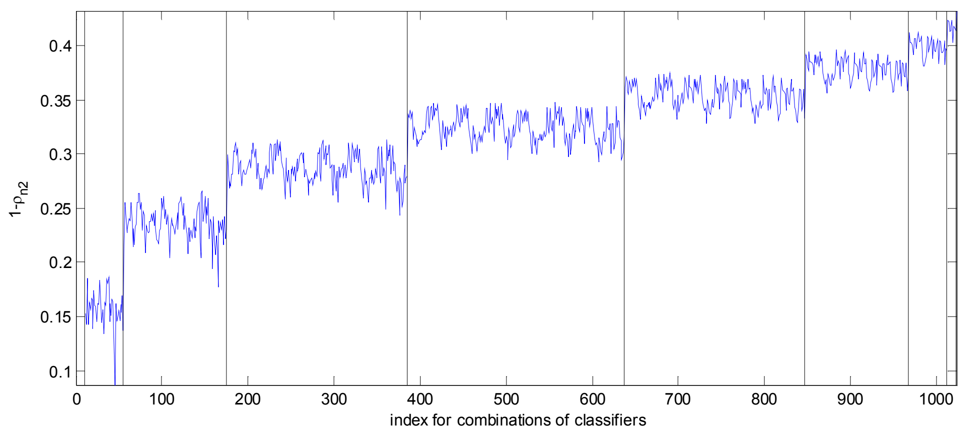


**Figure 9 – Correlation $\rho_{n2}$ for all classifier combinations.**

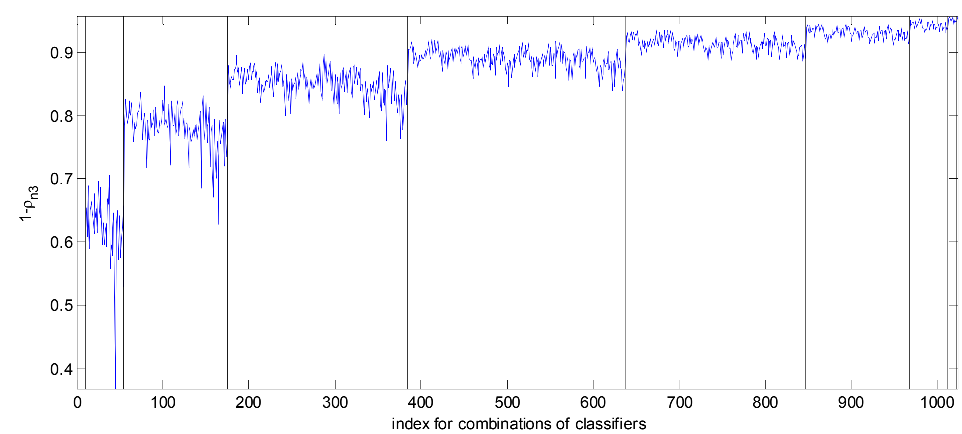


**Figure 10 – Correlation $\rho_{n3}$ for all classifier combinations.**

The statistical covariance is 0.0050, 0.0420, 0.0088, for $\rho_{n1}$, $\rho_{n2}$, and $\rho_{n3}$, respectively.

Coming back to the question how one would select the best set of classifiers, one could, of course, simply evaluate the performance of all possible classifier combinations and select the best performer. Where this is not possible, the correlation analysis may be helpful in identifying the best set of classifiers. Table 9 shows the classifier performance based on the selection rule “next best” where the next best individual classifier was added to the fusion suite. Here, too, we observe that the performance fluctuates somewhat, making identification of the best classifier set a 2-pass process.

**Table 9: Classifier Selection based on “Next Best”Performance only**

| **Classifiers** | 8 | 8, 6 | 8, 6, 1 | 8, 6, 1, 5 | 8, 6, 1, 5, 3 | 8, 6, 1, 5, 3, 7 | 8, 6, 1, 5, 3, 7, 9 | 8, 6, 1, 5, 3, 7, 9, 2 | 8, 6, 1, 5, 3, 7, 9, 2, 10 | 8, 6, 1, 5, 3, 7, 9, 2, 10, 4 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Performance** | 0.6563 | 0.6979 | 0.7093 | 0.7230 | 0.7172 | 0.7350 | 0.7207 | 0.7444 | 0.7317 | 0.7344 |

Table 10 shows the performance of classifier sets whose selection was aided by the correlation measures. The resulting best fusion performance is higher than using the “next best” strategy. The best combination for this strategy turns out to be the second best overall achievable performance (which may well be a coincidence). This also means that the overall highest performance was not found with this method. However, as illustrated in Figure 11, the found set (red circle at far end of curve) extends fairly high up to the end of the distribution compared to the performance achieved with the “next best: strategy (green circle).

**Table 10: Classifier Selection based on Correlation**

| **Classifiers** | 8 | 8, 2 | 8, 2, 1 | 8, 2, 1, 4 | 8, 2, 1, 4, 3 | 8, 2, 1, 4, 3, 6 | 8, 2, 1, 4, 3, 6, 7 | 8, 2, 1, 4, 3, 6, 7, 5 | 8, 2, 1, 4, 3, 6, 7, 5, 9 | 8, 2, 1, 4, 3, 6, 7, 5, 9, 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Performance($\rho_{nx}$)** | 0.6563 | 0.6997 | 0.7251 | 0.7438 | 0.7585 | 0.7573 | 0.7633 | 0.7531 | 0.7336 | 0.7344 |

Note that we use $\rho$-correlation not used directly to make the final decision on the classifier set. Rather, the correlation is used to select the next classifier to be included in the set. The final decision is based on fusion performance which has to be computed for only 10 sets of classifiers as opposed to for more than a 1000 sets.

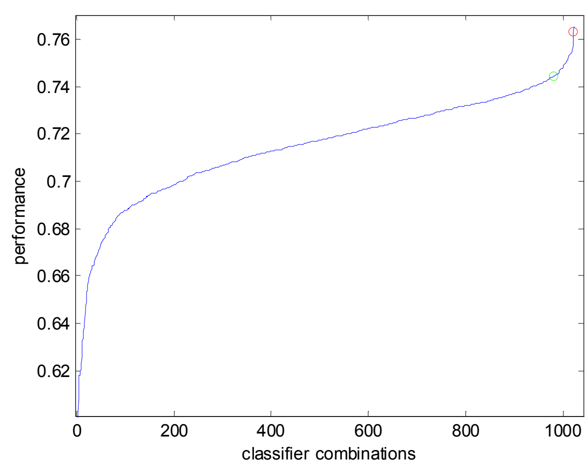


**Figure 11 – Ordered fusion performance for all combinations; red circle: performance accomplished with set found by correlation method; green circle: performance accomplished with set found by “next best” method.**

Figure 12 illustrates that the performance of the classifier sets chosen by the correlation method is consistently better than those using the “next best” method.

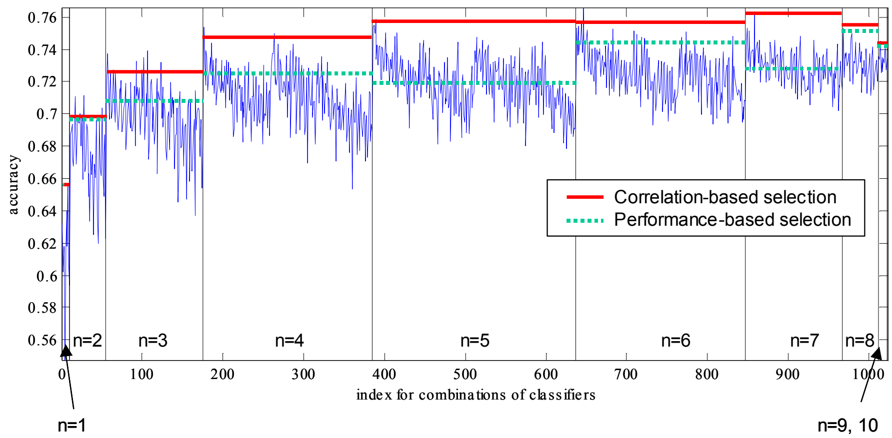


**Figure 12 – Comparison of performance found within different classifier sets for correlation-based method and "next best" method**

## CONCLUSIONS AND SUMMARY

This paper examined classifier selection from a given set of classifiers. Using all classifiers can lead to suboptimal performance. Where the exhaustive evaluation of classifiers cannot be easily accomplished, diversity measures can aid in finding close to optimal classifier sets. In particular, a group of $\rho$-correlation measures was introduced that show reasonable correspondence to fusion performance. Using the measure successively on the base set with the remaining classifiers, a relatively fast way is found to arrive at a final set of classifiers. Data from a real-world example were used to confirm the merit of the measures. Some noise in the output points to factors that have not been fully captured by the correlation measure. Because often times not all classes are equally important, a cost measure should be considered that takes a more comprehensive view of the classification task. It should be noted that performance may not be the only criterion by which classification is measured. Other issues such as solution robustness have been deliberately been excluded and should be subject to future research.

## REFERENCES


1. Roli, F., Giacinto, G., and Vernazza, G. "Methods for Designing Multiple Classifier Systems", *Proceedings of MCS 2001*, pp78-87, 2001.
2. Kuncheva L. I., and Jain, L. C. "Designing Classifier Fusion Systems by Genetic Algorithms", *IEEE Transactions on Evolutionary Computation*, Vol. 4, No. 4, pp327-36, 2000.
3. Krogh A. and Vedelsby, J. "Neural Network Ensembles, Cross Validation, and Active Learning", In G. Tesauro, D. Touretzky, and T. Leen, (editors), *Advances in Neural Information Processing Systems*, Vol. 7, pp231-238, Cambridge, MA, MIT Press, 1995.
4. Kuncheva, L. I., and Whitaker, C. J. "Ten Measures of Diversity in Classifier Ensembles: Limits for Two Classifiers", *IEEE Workshop on Intelligent Sensor Processing*, Birmingham, February 2001
5. Prodromidis, A. L., Stolfo, S., and Chan, P. K., "Pruning Classifiers in a Distributed Meta-Learning System", *Proceedings of 1st National Conference on New Information Technology (NIT'98)*, Athens, Greece, 1998.
6. Optiz, D. W., "Feature Selection for Ensembles", *Proceedings of 16th Intl. Conference of AI*, pp. 379-384, 1999.
7. Petrakos, M., Kannelopoulos, I., Benediktsson, J., and Pesaresi, M. "The Effect of Correlation on the Accuracy of the Combined Classifier in Decision Level Fusion", *Proceedings of IEEE 2000 International Geo-Science and Remote Sensing Symposium*, Vol. 6, 2000.
8. Goebel, K., Yan, W., and Cheetham, W. "A Method to Calculate Classifier Correlation for Decision Fusion", *Proceedings of IDC 2002*, Adelaide, 11-13 February, 2002., pp. 135-140, 2002.
9. Goebel K., and Yan, W. "Choosing Classifiers for Decision Fusion", *Proceedings of the Seventh International Conference on Information Fusion*, vol. I, pp. 563-568, 2004.
10. Yan, W., and Goebel, K. "Designing Classifier Ensembles with Constrained Performance Requirements", *Proceedings of SPIE, Multisensor Multisource Information Fusion: Arch., Algs., and Appl. 2004*, pp78-87, 2004.