# Choosing Classifiers for Decision Fusion

**Kai Goebel**
GE Global Research
K1-5C4A, One Research Circle,
Niskayuna, NY 12309
USA
goebelk@research.ge.com

**Weizhong Yan**
GE Global Research
K1-5B34B, One Research Circle,
Niskayuna, NY 12309
USA
yan@research.ge.com

**Abstract -** *This paper investigates the use of the ρ-correlation as a measure for classifier diversity to aid in the choice of classifiers for a fusion ensemble. Specifically, we define a measure that captures the correlation for n classifiers for binary output as well as for classifier with continuous output. We then suggest the use of the ρ-correlation in classifier selection where classifiers are picked sequentially from a larger pool of classifiers without the need to exhaustively calculate the performance of all possible permutations. We show that this simple method will give close to optimal classification performance. We present examples from real applications for both binary as well as continuous out classifiers.*

**Keywords:** Classifier fusion, decision fusion, correlation, diversity, MCS, information fusion, tracking, classification.

## 1 Introduction

The success of classifier fusion depends on two factors: 1) a pool of diverse individual classifiers to be fused, and 2) the proper combining method. There are two ways to obtain diverse individual classifiers. One approach is to first heuristically pick a number and types of classifiers and then ensure a diverse output, for example by using different data samples in the training phase (bagging and boosting). Another approach is the "overproduce and choose" paradigm [1]. This paper follows the second 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 optimal performance as possible while at the same time avoiding the exhaustive evaluation of all possible classifier combinations.

In classifier fusion, it is desirable to use classifiers that – besides offering reasonable performance – have a mutual low correlation. Obviously, if two classifiers in a three classifier fusion task are completely redundant, many fusion schemes will not only not gain anything but will actually exhibit poorer performance. Some degree of confirmatory information is of course desired, but it is the complementary information that gives the multi-classifier fusion a chance to be successful.

Below, we will first illustrate the background of classifier correlation measures then discuss the proposed correlation analysis in section 3, followed by the selection process and applications to defect diagnostics in section 4 and 5, respectively.

## 2 Background

It has been recognized that the success of a classifier fusion in performance improvement relies on properly choosing individual classifiers to be fused. Selecting classifiers can be performed through exhaustive search with the performance of fusion being the objective function. This is a fairly straightforward method, however, as the number of classifiers increases, it becomes computationally too expensive. In their recent work on methods for designing multiple classifier systems based on the "overproduce and choose" paradigm, Roli et al. [1] described six different approaches for selecting classifiers. Also recently Kuncheva and Jain [2] used GAs to select classifiers and a corresponding feature subset for each classifier at the same time.

Instead of directly using fusion performance as the search objective function, several measures have been proposed for quick quantification of the goodness of a group or pair of classifiers, i.e., how successful the fusion will be when those classifiers are combined. One of the most popular one of these measures is the diversity. Krogh & Vedelsby [3] define diversity as ambiguity that is the variation of the output of ensemble members averaged over unlabeled data. 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.

### 2.1 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 ***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 two-class confusion matrix ***M*** is shown in Fig. 1.

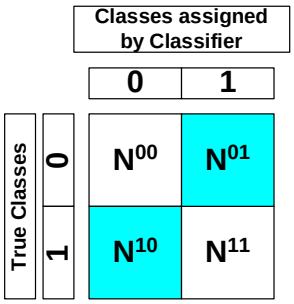

Fig. 1. Typical 2-class confusion matrix.

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:

$$FP = \frac{N^{01}}{N^{00} + N^{01}} \quad (1)$$

$$FN = \frac{N^{10}}{N^{10} + N^{11}} \quad (2)$$

$$TER = \frac{N^{01} + N^{10}}{N^{00} + N^{11} + N^{01} + N^{10}} \quad (3)$$

$$TSR = 1 - TER \quad (4)$$

## 3 Classifier correlation

While it has been well understood that each individual classifier's performance is very important to the performance of a classifier fusion there seems to be less awareness that the dependency between the classifiers to be fused also affects the fusion results. Some studies [7] have shown that the degree of correlation between the classifiers adversely affects the performance of the subsequent classifier fusion. If two classifiers agree everywhere, the fusion of the two classifiers will not achieve any accuracy improvement no matter what fusion method is used. For classifier fusion design, classifier correlation analysis is, therefore, equally important as the classifier performance analysis.

### 3.1 2-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*** as shown in Fig. 2 can be generated for each classifier pair. 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

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

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).

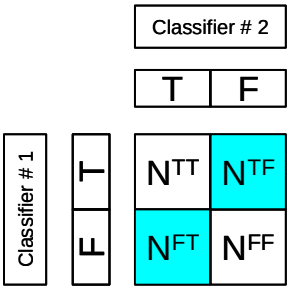


Fig. 2. Correlation Analysis Matrix

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

Table 1: Results from experiment for 2 classifiers.

| *Answer classifier 1* | *Answer 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

### 3.2 n-Classifier correlation analysis

We proposed an extension of the 2 class correlation coefficient to n different classifiers [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 $\rho$–correlation, the larger the redundancy. In particular, the $\rho$–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. The $\rho$–correlation coefficient gets larger as the number of wrong answers are the same for many answers. 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 $\rho$–correlation coefficient is then

$$\rho_n = \frac{nN^f}{\sum_{i=1}^{2^n-2} N_i^c + nN^f} \qquad (6)$$

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 more conveniently be rewritten as

$$\rho_n = \frac{nN^f}{N - N^f - N^t + nN^f} \qquad (7)$$

Consider a 3-classifier example which is the same as previous 2-classifier example except that a third classifier was added that gets the answer wrong in 50% of the cases. The calculation of $\rho_n$ yields: $\rho_n$ =0.21

Although the newly added classifier has poor performance, its addition reduces the overall redundancy of the classifier assembly.

It is interesting to note that the $\rho$–correlation does not record redundancy with any particular classifier (for n>2) but with a set of classifier only. For illustrative purpose, consider the simplistic cases shown in Table 2 and Table 3 [8]:

Table 2: 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 $\rho$–correlation is $\rho_n$ = 0.5

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

| *Answer classifier 1* | *Answer classifier 2* | *Answer classifier 3* |
|---|---|---|
| T | F | T |
| F | T | T |
| F | T | F |
| T | T | T |
| F | F | F |

The $\rho$–correlation is $\rho_n$ = 0.5

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.

It has to be noted that the calculation of the $\rho$–correlation factor can be performed on multi-class scenarios as well because the factor is only concerned with the correctness of the outcome.

### 3.3 Classifiers with continuous output

For classifiers that give continuous output such as confidences, partial class membership etc. we propose a slightly different operator $\rho_{n_c}$ [8]. Let $N_c^f$ be the sum of all false classifier outputs that are greater than decision threshold $t_c$; $o_i^f$ be the fused output per case i, $o_i^f = \frac{\sum_{j=1}^{n_{classifiers}} o_{j,i}}{n_{classifiers}}$ and $o_{j,i} = \begin{cases} o_{j,i}^{raw} & \text{if } o_{j,i}^{raw} > t_c \\ 1 - o_{j,i}^{raw} & \text{otherwise} \end{cases}$; $t_c$ is the decision threshold for class membership; $N_c^f$ is the sum of cases that are false per threshold t, $N_c^f = \sum_{i=1}^{N_f} o_i^f$ where $N_f$ is the number of false cases while $N_c^t = \sum_{j=1}^{N_t} o_j^f$ is the sum of true fused cases where $N_t$ is the number of true cases. Then the $\rho$–correlation $\rho_{n_c}$ is

$$\rho_{n_c} = \frac{nN_c^f}{N - N_c^f - N_c^t + nN_c^f} \qquad (8)$$

Note that $\rho_{n_c} \neq 0$ for completely redundant information if the output is not completely symmetric (which is typically the case).

Table 4: Classifier output for 2 classifiers in continuous format

| *True state* | *Output classifier 1* | *Output classifier 2* | *Normalized adjusted cumulative output $o_i$* |
|---|---|---|---|
| 1 | 0.7 | 0.6 | 0.65 |
| 1 | 0.6 | 0.3 | 0.7 |
| 1 | 0.3 | 0.6 | 0.65 |
| 0 | 0.2 | 0.7 | 0.75 |
| 0 | 0.6 | 0.8 | 0.7 |
| 0 | 0.9 | 0.8 | 0.85 |
| 0 | 0.1 | 0.6 | 0.75 |
| 1 | 0.2 | 0.6 | 0.75 |
| 0 | 0.4 | 0.3 | 0.65 |
| 0 | 0.4 | 0.4 | 0.6 |
| 1 | 0.8 | 0.6 | 0.7 |
| 0 | 0.2 | 0.1 | 0.85 |
| 0 | 0.3 | 0.7 | 0.7 |
| 1 | 0.7 | 0.9 | 0.8 |
| 0 | 0.4 | 0.4 | 0.6 |
| 0 | 0.7 | 0.4 | 0.65 |

Consider now 2 classifiers with continuous output bounded by [0,1] as shown in Table 4. Table 4 also shows the normalized adjusted output o.

The calculation of $\rho_{n_c}$ yields: $\rho_{n_c}$ =0.2441

where

$$N_c^f = 1.55$$
$$N_c^t = 4.85 .$$

## 4 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 or until the $\rho$–correlation increases.

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. With that acknowledgment, we continue with the stated assumption

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 $\rho_n$=0.67. The joint $\rho$–correlation of the three classifiers is $\rho_n$ =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 $\rho_n$ =0.5; i.e., the $\rho$–correlation decreased.

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 |

## 5 Application to classifiers with continuous output

We show here an application to classifiers with 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 [9]. The classifiers were chosen to be all feedforward neural nets. The use of different input features to different networks and varied network configuration try to ensure a reasonable diversity. The accuracy was calculated subject to the Neyman-Pearson criterion where the true positive rate was set fixed at TPR=98%. In real-world classification problems, the classifier performance is often constrained by a given true positive or false positive rate. The goal of any further classifier design is then to reduce the false positive rate while maintaining the desired true positive rate. The accuracy is then directly proportional to the false positive rate. 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.

After finding the best classifier, ("classifier 1"), the $\rho$-correlation is calculated for the pairs of classifier 1 with the remaining classifiers. The lowest $\rho$–correlation combination is pair 1-7. Choosing the classifier with higher accuracy as the base classifier, i.e., classifier 1, the $\rho$–correlations with classifier 1 are as displayed in Table 7. Table 7 also shows the joint accuracy.

Table 7: Joint $\rho$–correlation and joint accuracy of classifier 1 with classifier n

| *classifier n* | *$\rho$–correlation* | *joint accuracy* |
|---|---|---|
| 7 | 0.0088 | 0.6850 |
| 4 | 0.0106 | 0.6937 |
| 8 | 0.0120 | 0.7110 |
| 10 | 0.0132 | 0.6869 |
| 3 | 0.0138 | 0.6894 |
| 9 | 0.0140 | 0.6686 |
| 6 | 0.0142 | 0.6676 |
| 2 | 0.0148 | 0.6871 |
| 5 | 0.0149 | 0.6736 |

The joint accuracy of classifier 1 with classifier n is plotted against the classifier used with classifier 1 as shown in Fig. 3. This figure also shows the base accuracy. What is interesting to note is that the joint accuracy has almost no statistical correlation *c* with the single accuracy. (*c*=0.01). That is, choosing classifiers based on their performance may not lead to the best joint performance. At the same time, the $\rho$–correlation (while still low) has a somewhat higher correlation (*c*=-0.44) with joint accuracy.

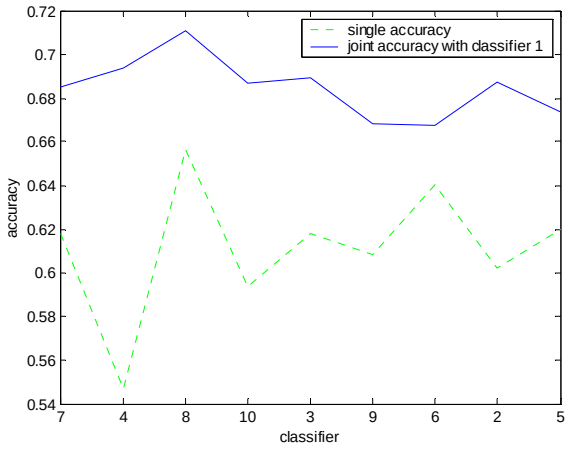


Fig. 3. Single accuracy versus pair-wise joint accuracy.

Table 8 shows the result of adding successively the classifier with remaining highest joint $\rho$–correlation to the existing set. It can be noted that the overall accuracy increases on average. It must also be noted that there are some downward steps where the overall accuracy drops. Fig. 4 illustrates that trend. There seems to be also a saturation effect that is consistent with the experience that there is an optimal number of classifiers to achieve maximum performance. Indeed, the best joint performance is reached with 6 classifiers (acc=0.7405).

Table 8: Mean of classifier 1 with remaining classifiers

| *Classifiers n* | *Joint accuracy* |
|---|---|
| 1, 7 | 0.6850 |
| 1, 7, 4 | 0.7091 |
| 1, 7, 4, 8 | 0.7099 |
| 1, 7, 4, 8, 10 | 0.7070 |
| 1, 7, 4, 8, 10, 3 | 0.7405 |
| 1, 7, 4, 8, 10, 3, 9 | 0.7278 |
| 1, 7, 4, 8, 10, 3, 9, 6 | 0.7375 |
| 1, 7, 4, 8, 10, 3, 9, 6, 2 | 0.7342 |
| 1, 7, 4, 8, 10, 3, 9, 6, 2, 5 | 0.7344 |

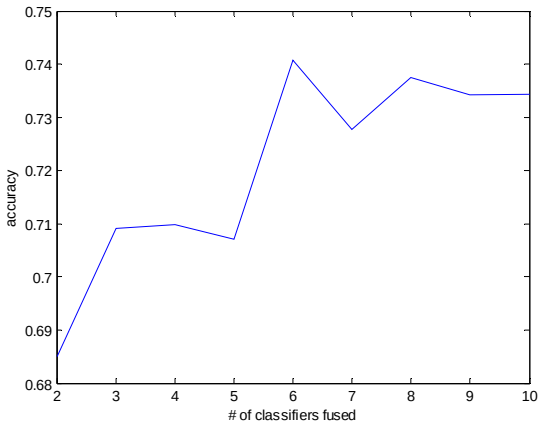


Fig. 4. Fusion performance versus number of fused classifiers

After establishing the curve as shown in Fig. 4, it is straightforward to read off the set of classifiers that leads to maximum performance. Some "noise" in the curve suggests that there are other factors that are not captured in this measure.

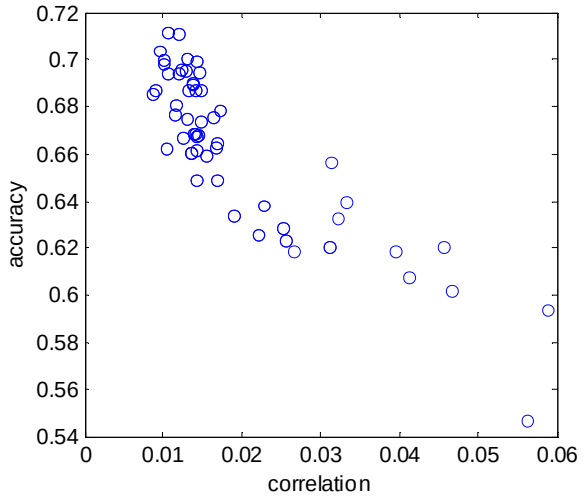


Fig. 5. Accuracy versus $\rho$-correlation.

Fig. 5 shows the accuracy against $\rho$-correlation which suggests a good correspondence between the two measures. The statistical correlation is $c$=-0.84. In comparison, the statistical correlation between the ambiguity measure V [3] and the pair-wise accuracy is only $c$=0.65

## 6 Conclusions and summary

This paper showed how the $\rho$-correlation can be used for classifier selection. The $\rho$-correlation measures the correlation between n classifiers for crisp as well as for soft class assignment. The measure takes into account not only whether an output was right or wrong but also the closeness to the decision boundary. In other words, it takes into account how right or how wrong an answer was.

Using the measure successively on the base set with the remaining classifiers, a relatively fast way is found to find a set of classifiers that leads to high fusion performance. Some noise in the output points to factors that have not been fully captured by the correlation measure. In real-life applications, the class-specific performance is only one measure of interest. 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. Other uses of the correlation could also include its use within a fusion algorithm.

## References

[1] F. Roli, G. Giacinto, and G. Vernazza, Methods for designing multiple classifier systems, *Proc. of MCS 2001*, pp78-87, 2001.

[2] L. I. Kuncheva and L. C. ain, Designing classifier fusion systems by genetic algorithms, *IEEE Trans. on Evol. Comp.*, Vol. 4, No. 4, pp327-36, 2000.

[3] A. Krogh and J. Vedelsby, 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] L. I. Kuncheva, and C. J. Whitaker, Ten measures of diversity in classifier ensembles: limits for two classifiers, *IEEE Workshop on Int. Sensor Processing*, Birmingham, February 2001

[5] A. L. Prodromidis, S. Stolfo, and P. K. Chan, Pruning classifiers in a distributed meta-learning system, *Proc. 1st Nat. Conf. New Inf. Tech. (NIT'98)*, Athens, Greece, 1998.

[6] D.W. Optiz, Feature selection for ensembles, *Proc. 16th Intl. Conf. AI*, pp. 379-384, 1999.

[7] M. Petrakos, I. Kannelopoulos, J. Benediktsson, and M. Pesaresi, The effect of correlation on the accuracy of the combined classifier in decision level fusion, *Proc. IEEE 2000 Intl. Geo-science and Remote Sensing Symp.*, Vol. 6, 2000.

[8] K. Goebel, W. Yan, and W. Cheetham, A method to calculate classifier correlation for decision fusion, *Proc. IDC 2002*, Adelaide, 11-13 February, 2002., pp. 135-140, 2002.

[9] W. Yan and K. Goebel, Designing Classifier Ensembles with Constrained Performance Requirements, *Proc SPIE Defense & Security Symposium, Multisensor Multisource Information Fusion: Architectures, Algorithms, and Applications 2004*, pp78-87, 2004.