

# Analysis of a Fusion Method For Combining Marginal Classifiers

Mark D. Happel and Peter Bock

Department of Computer Science, The George Washington University, Washington, DC  
20052, USA  
{mhappel, pbock}@seas.gwu.edu

**Abstract.** The use of multiple features by a classifier often leads to a reduced probability of error, but the design of an optimal Bayesian classifier for multiple features is dependent on the estimation of multidimensional joint probability density functions and therefore requires a design sample size that, in general, increases exponentially with the number of dimensions. The classification method described in this paper makes decisions by combining the decisions made by multiple Bayesian classifiers using an additional classifier that estimates the joint probability densities of the *decision space* rather than the joint probability densities of the *feature space*. A proof is presented for the restricted case of two classes and two features; showing that the method always demonstrates a probability of error that is less than or equal to the probability of error of the marginal classifier with the lowest probability of error.

## 1. Background

Given a set of objects and their corresponding *feature vectors*  $\mathbf{X} = [\chi_1 \chi_2 \dots \chi_d]^T$  in *feature space*  $\Pi$ , one of the fundamental problems of pattern classification is to define a function (a *classifier*)  $\Psi: \Pi \rightarrow \Delta$  that can assign an appropriate *class label*  $\omega_i$  to any given  $\mathbf{X}$  in the feature space. The assignment itself is called a *classification decision*  $\delta \in \Delta$ , and the set of all possible decisions is the *decision space*  $\Delta$ . In a Bayesian classifier, the classification decision is made based on the *a posteriori* probabilities that the input is a member of a given class given the input. For a given input  $\mathbf{X}$ , the *a posteriori* probability for class  $\omega_i$ ,  $p(\omega_i | \mathbf{X})$ , can be calculated using Bayes' rule:

$$p(\omega_i | \mathbf{X}) = \frac{p(\mathbf{X} | \omega_i)P(\omega_i)}{\sum_i p(\mathbf{X} | \omega_i)P(\omega_i)} \quad (1)$$

The Bayesian decision rule selects the class label which corresponds to the maximum *a posteriori* probability. The class-conditional probability density function  $p(\mathbf{X} | \omega_i)$  is often referred to as the *likelihood function* [5], and the likelihood function weighted by the *a priori* probability  $P(\omega_i)$  is referred to here as the *weighted likelihood*. Since the sum of the weighted likelihoods (the denominator in the

equation above) is positive and common to all of the *a posteriori* probabilities, it can be factored out and the comparison made of the weighted likelihoods instead:

$$\delta = \omega_i \text{ such that } p(\mathbf{X} | \omega_i)P(\omega_i) > p(\mathbf{X} | \omega_j)P(\omega_j) \text{ for all } i \neq j \quad (2)$$

If the probability of error attained by a Bayesian classifier is unacceptably high for the requirements of a given problem, two or more features can be used simultaneously to form multivariate joint probability density functions. By using two or more features, the multivariate classifier is often able to achieve a significantly better classification performance than a comparable univariate classifier.

The Bayesian classifier is optimal in the sense that it has the lowest possible probability of error  $\epsilon_\beta$  for a given set of probability density functions [6]. If the classes' density functions are not known, then they must be estimated from sample data. However, the estimation of multivariate density functions in high-dimensional spaces is nontrivial, and may require an unrealistically large design sample size to attain a sufficiently accurate estimate. This "curse of dimensionality" [1] leads to an interesting paradox: as the number of dimensions increases, the theoretical performance of the Bayesian classifier improves but the practical problems involved in implementing such a classifier also increase, resulting in a decline in the actual classification performance beyond a certain threshold dimensionality [6]. Consequently, for situations in which the optimal Bayesian classifier performance is insufficient for  $d$  dimensions, it may not be possible in practice to attain better classification performance using  $d+1$  dimensions, even though the theoretical Bayesian performance should increase.

From the preceding discussion, it is apparent that a method for obtaining an improvement in the classification performance for the  $d$ -dimensional Bayesian classifier without requiring the estimation of  $d+1$  dimensional density functions would prove useful. It is intuitively appealing to imagine combining several, lower-dimensional Bayesian classifiers in such a way as to provide a lower error rate than any one of them alone can achieve, and perhaps even to approach the error rate attainable with a higher-dimensional classifier.

Current strategies for obtaining group decisions can be divided into two broad categories: dynamic classifier selection and classifier fusion [12]. Dynamic classifier selection (DCS) strategies attempt to predict or identify, for a given input, the best decision out of the set of decisions made by the individual classifiers. In contrast, classifier fusion algorithms define a function  $\xi: \Delta \rightarrow \Delta$  that can be used to calculate a decision based on the simultaneous decisions of all of the individual classifiers. Classifier fusion methods include majority voting [9], weighted majority voting, averaged Bayesian decisions [13], naive Bayesian classifiers [2, 10], Dempster-Shafer approaches [3, 11], and stacking strategies [4]. Stacking strategies differ from other classifier fusion strategies in that the fusion function  $\xi: \Delta \rightarrow \Delta$  is not defined *a priori* but is instead learned by a "combining classifier" [4]. The combining classifier  $\Psi^*$  receives as input the classification decisions of  $m$  member classifiers  $\Psi_i(\mathbf{X})$  and computes a final classification decision  $\delta^*$ :

$$\delta^* = \Psi^* [\Psi_1(\mathbf{X}), \Psi_2(\mathbf{X}), \dots, \Psi_m(\mathbf{X})] \quad (3)$$

In this paper, a stacking method is proposed as a means of combining marginal decisions into a single, "pseudo-multivariate" decision.

## 2. Proposed Method

The method proposed here is to use the marginal decisions as features, thereby forming a *decision vector*. An additional Bayesian classifier, called a *supervisory classifier*, can then be used to classify the vector of marginal decisions and generate a combined classification decision. The supervisory classifier makes its classification decision based on estimates of the joint probability densities of the decision space rather than the joint probability densities of the feature space.

A block diagram of this architecture is shown in Figure 1. The example shown uses a feature vector  $\mathbf{X} = [\chi_1 \ \chi_2]^T$  that consists of only two features. Like the marginal classifiers  $\Psi_1$  and  $\Psi_2$ , the supervisory classifier  $\Psi_H$  is a Bayesian classifier, allowing the system to be implemented from a common building block. Feature  $\chi_1$  is a random variable with the class-conditional probability density function  $p(\chi_1 | \omega)$ . Likewise, feature  $\chi_2$  is a random variable with the class-conditional probability density function  $p(\chi_2 | \omega)$ . An optimal Bayesian bivariate classifier  $\Psi_\beta(\chi_1, \chi_2)$  would generate a decision  $\delta_\beta$  based on the bivariate class-conditional probability density function  $p(\chi_1, \chi_2 | \omega)$  and the *a priori* probability  $P(\omega)$  using Equation 2. This requires the bivariate density functions to be known beforehand or to be estimated using sample data.

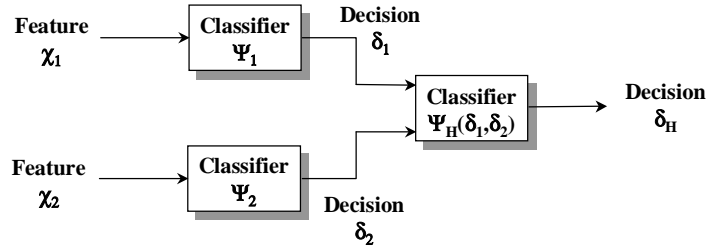


Fig. 1. A Stacking Architecture for Combining Marginal Classification Decisions

However, in this case the intention is to avoid the need to estimate the bivariate density functions. The proposed method accomplishes this by first generating classification decisions  $\delta_1$  and  $\delta_2$  based on the univariate (marginal) density functions. Classifiers  $\Psi_1$  and  $\Psi_2$  can be viewed as performing a nonlinear transformation of features  $\chi_1$  and  $\chi_2$  from the feature space  $\Pi$  to the decision space  $\Delta$ . This mapping will generally be many-to-one, as there are generally far fewer classes defined for a given problem than there are values for the defined features, resulting in considerable compression from feature space to decision space.

The second step in the proposed method is to combine the decisions  $\delta_1$  and  $\delta_2$  by means of a supervisory classifier,  $\Psi_H$ , which uses decisions  $\delta_1$  and  $\delta_2$  as features. In order to use the Bayesian decision rule, classifier  $\Psi_H$  needs the bivariate class-conditional probability density function  $p(\delta_1, \delta_2 | \omega)$ , which can be estimated using the same design sample used to estimate  $p(\chi_1 | \omega)$  and  $p(\chi_2 | \omega)$ . Since the decision space is a discrete space with relatively few entries (as compared to the number of feature values in feature space),  $p(\delta_1, \delta_2 | \omega)$  can be estimated using a much smaller sample size than can  $p(\chi_1, \chi_2 | \omega)$ . Note that the supervisory classifier can "override" the

decisions of the marginal classifiers, choosing a different class from any of the marginal decisions. This *override property* is a key factor in the superiority of the proposed method to voting methods. Further information on the operation of the method and the utility of the override property may be found in [7, 8].

### 3. Properties of the Proposed Method

To better understand the behavior of the proposed method, research to date has focused on the simplest case: a problem requiring the classification of patterns into one of two classes (Grey or 'G' and Black or 'B') based on the values of two features. The proposed method divides the feature space into multiple *partitions* based on the locations of the decision surfaces determined by the marginal classifiers. It then generates classification decisions for each partition by choosing the maximum class-conditional probability (i.e., the class with the largest volume under its weighted likelihood within the partition). The marginal classification decisions, partitions, and partition classification decisions are shown in Figure 2.

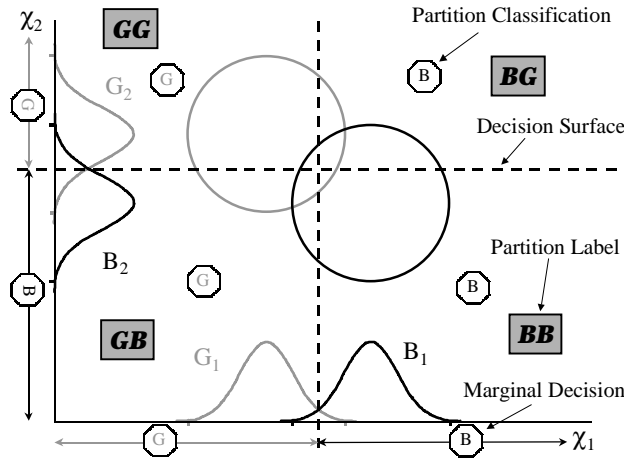
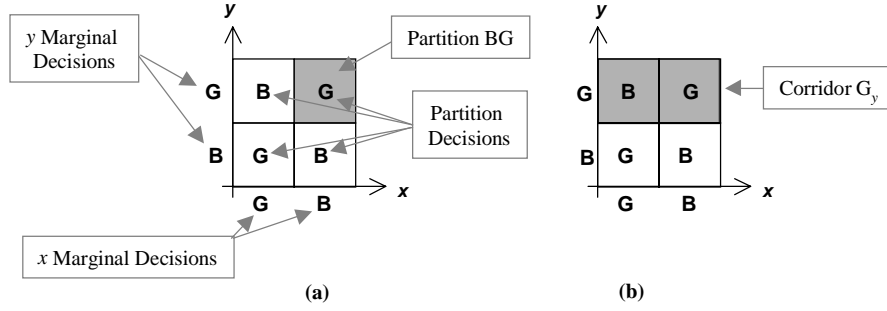


Fig. 2. Classifier Partitions and Classification Decisions

One of the goals of the preliminary research effort was to determine the properties of the proposed method without requiring assumptions regarding the parametric forms of the complete or marginal likelihoods to be made. Consequently, it was found to be advantageous to graphically represent the partitions and the associated marginal and partition classifications, without explicitly representing the likelihoods themselves. A *partition plot* for the example in Figure 2 is shown in Figure 3.

By examining all of the possible partition arrangements, it is possible to determine upper and lower bounds on the probability of error  $\epsilon_{\text{H}}$  of the proposed method as applied to a 2 class and 2 feature case without making assumptions as to the form or parameters of the likelihoods. Several lemmas that are helpful for establishing these bounds are proved below. It is also important to note that the magnitude of the hypervolume under a portion of the complete density function is equal to the

magnitude of the associated hypervolume under the corresponding marginal density function [8]. The first lemma will be used to discard some of the possible candidate classification arrangements due to the inconsistency between the partition classifications and the corresponding marginal classification.



**Fig. 3.** Partition plots for the analysis. Note that marginal decisions are shown along the respective axes, while the partition decisions are shown within each partition. Therefore, in (a) the shaded partition BG (i.e.,  $\delta_x = B$  and  $\delta_y = G$ ) has been classified as G. All of the partitions that correspond to a given marginal decision form a *corridor*, shown as a shaded region in (b).

**Lemma 1:** Suppose that all of the partitions  $\theta_k$  in a given corridor  $c$  of the decision space are assigned the same class label  $\omega_i$ . Then the class label assigned to the corresponding decision region by the corresponding marginal classifier will be  $\omega_i$ .

**Proof:** The assignment of label  $\omega_i$  to all of the partitions  $\theta_k$  in the corridor implies that, within each of these partitions, the volume under the  $p(X|\omega_i)P(\omega_i)$  surface is greater than the volume under any other classes' surface:

$$\int_{\theta_k} p(X|\omega_i)P(\omega_i) dX > \int_{\theta_k} p(X|\omega_j)P(\omega_j) dX; \quad 1 \geq j \geq N, i \neq j. \quad (4)$$

Likewise, it follows that the corridor's  $\omega_i$  volume, composed of the sum of the  $\omega_i$  volumes within each partition in the corridor, will also be the largest class-conditional volume within the corridor:

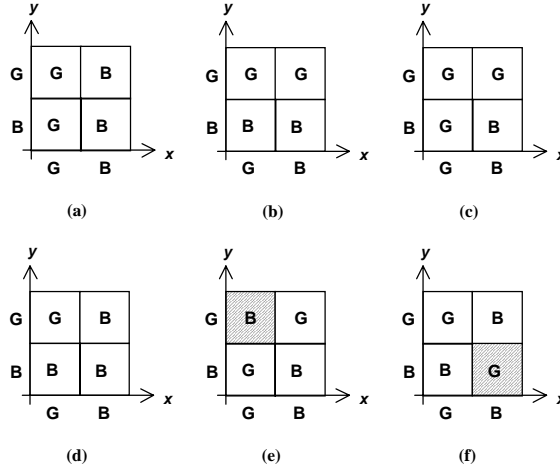
$$\int_c p(X|\omega_i)P(\omega_i) dX > \int_c p(X|\omega_j)P(\omega_j) dX; \quad 1 \geq j \geq N, i \neq j. \quad (5)$$

As noted above, the volume under the complete weighted likelihood within a corridor is equal to the volume (or area, if one-dimensional) under the corresponding marginal weighted likelihood. Consequently,

$$\int p(X|\omega_i)P(\omega_i) dX > \int p(X|\omega_j)P(\omega_j) dX; \quad 1 \geq j \geq N, i \neq j. \quad (6)$$

and the classification decision, in accordance with the Bayesian decision rule will be  $\omega_i$ . *QED.*

As a result of Lemma 1, ten of the candidate classification arrangements can be discarded, since they contain corridors with partitions that share classification labels that are not the same as the label of the corresponding marginal classification region. The six remaining candidate classification arrangements, shown in Figure 4, are all valid arrangements.



**Fig. 4.** Legal Arrangements of Partition Classifications. The shaded partitions in arrangements (e) and (f) represent decision overrides.

**Lemma 2:** Suppose that, for every corridor associated with a given marginal classifier  $x$ , the partitions within a given corridor are assigned the same class label. Then the probability of error  $\epsilon_H$  for the proposed method is equal to the probability of error  $\epsilon_x$  of the associated marginal classifier.

**Proof:** Refer to Figure 4(a). Let  $R_\omega$  represent the region classified as class  $\omega$  by the marginal classifier, and let  $C_\omega$  represent the associated corridor. The probability of error for a given classifier is equal to the sum of the probabilities of the less likely classes (i.e., all of the classes except the most likely one). Consequently, the probability of error for the  $x$  marginal classifier in Figure 4(a) is given by:

$$\epsilon_x = \int_{R_G} p(x | \omega_B) P(\omega_B) dx + \int_{R_B} p(x | \omega_G) P(\omega_G) dx \quad (7)$$

and the probability of error for the proposed method in Figure 8 is given by:

$$\epsilon_H = \iint_{C_G} p(x, y | \omega_B) P(\omega_B) dx dy + \iint_{C_B} p(x, y | \omega_G) P(\omega_G) dx dy \quad (8)$$

The volume under the complete weighted likelihood for each class within a given corridor is equal to the volume (or area, in this case) under the marginal weighted likelihood for the same class.

$$\int_{R_{\omega_j}} p(x | \omega_i) P(\omega_i) dx = \iint_{C_{\omega_j}} p(x, y | \omega_i) P(\omega_i) dx dy; \quad i \neq j \quad (9)$$

Consequently, it is apparent by substituting terms that  $\epsilon_x = \epsilon_H$ . In addition, it can be shown [8] that the probability of error  $\epsilon_x$  of the associated marginal classifier will be less than or equal to the probability of error  $\epsilon_y$  of the other marginal classifier. *QED*.

The next lemma concerns two of the remaining four arrangements. In these arrangements, three of the four partitions share a common class label, as shown in the examples in Figures 4(c) and 4(d). It can be shown that these arrangements result in a probability of error  $\epsilon_H$  that is equal to or less than the smaller probability of error of the two marginal classifiers.

**Lemma 3:** Suppose that the proposed method is used to discriminate between two-dimensional patterns belonging to two classes with arbitrary joint probability density functions, and that three of the four possible partitions are assigned the same class label. Then the probability of error for the proposed method  $\epsilon_H$  is less than the smaller probability of error of the associated marginal classifiers.

**Proof:** Refer to Figure 4(d). Let  $\epsilon_p$ ,  $1 \leq i \leq 4$ , represent the contribution of marginal region  $R_i$  to the marginal classifier's probability of error (*i.e.*, the volume under the weighted likelihood of the unchosen class in that region). Let  $I_p$  represent the volume under the weighted likelihood curve for class  $I$  within partition  $P$  (*i.e.*,  $B_{GB}$  would refer to the volume under the weighted likelihood for class B in partition GB). By Lemma 2, the marginal probabilities of error  $\epsilon_x$  and  $\epsilon_y$  can be related to the volumes within the partitions and hence to the probability of error of the proposed method:

$$\epsilon_x = (B_{GB} + B_{GG}) + (G_{BB} + G_{BG}) \quad (10)$$

$$\epsilon_y = (B_{GG} + B_{BG}) + (G_{GB} + G_{BB}) \quad (11)$$

$$\epsilon_H = B_{GG} + G_{GB} + G_{BG} + G_{BB} \quad (12)$$

The proposed classifier, which selects the largest volume within a partition, classified partition GB as B and partition BG as B. This implies that

$$G_{GB} \leq B_{GB} \quad \text{and} \quad G_{BG} \leq B_{BG} \quad (13)$$

Assume that  $\epsilon_x \leq \epsilon_H$ . This implies that

$$B_{GB} + B_{GG} + G_{BB} + G_{BG} \leq B_{GG} + G_{GB} + G_{BG} + G_{BB} \quad (14)$$

Eliminating common terms yields

$$B_{GB} \leq G_{GB} \quad (15)$$

which contradicts equation 13. Therefore,  $\epsilon_x > \epsilon_H$ . Similarly, assume that  $\epsilon_y \leq \epsilon_H$ . This implies that

$$B_{GG} + B_{BG} + G_{GB} + G_{BB} \leq B_{GG} + G_{GB} + G_{BG} + G_{BB} \quad (16)$$

Eliminating common terms yields

$$B_{BG} \leq G_{BG} \quad (17)$$

which also contradicts Equation 13. Therefore,  $\epsilon_y > \epsilon_H$ . Since  $\epsilon_H < \epsilon_x$  and  $\epsilon_H < \epsilon_y$ ,  $\epsilon_H$  is less than the smaller of  $\epsilon_x$  and  $\epsilon_y$ . A similar argument can be used to show that similar contradictions also result when partition BB is the sole partition classified as B. *QED.*

The final lemma concerns the two remaining arrangements, Figures 4(e) and 4(f), which share the property of having a partition that has been classified differently than either of the corresponding marginal regions. As discussed previously, this amounts to having the supervisory classifier "override" all of the decisions made by the marginal classifiers, choosing a class that was not selected by either of the marginal classifiers.

**Lemma 4:** Suppose that the proposed method is used to discriminate between two-dimensional patterns belonging to two classes with arbitrary joint probability density functions, and that a partition is assigned a class label that is different than the label assigned to either of the two corresponding marginal classification regions. Then the probability of error for the proposed method  $\epsilon_H$  is less than the smaller probability of error of the associated marginal classifiers.

**Proof:** This proof will be performed in a manner similar to that of Lemma 3. Assume that the partition and marginal region classifications are given by Figure 4(f). Let  $\epsilon_i$ ,  $1 \leq i \leq 4$ , represent the contribution of marginal region  $R_i$  to the marginal classifier's probability of error (i.e., the volume under the weighted likelihood of the unchosen class in that region). Let  $I_p$  represent the volume under the weighted likelihood curve for class  $I$  within partition  $P$ . The marginal probabilities of error  $\epsilon_x$  and  $\epsilon_y$  can be related to the volumes within the partitions and hence to the probability of error of the proposed classifier:

$$\epsilon_x = (B_{GB} + B_{GG}) + (G_{BB} + G_{BG}) \quad (18)$$

$$\epsilon_y = (B_{GG} + B_{BG}) + (G_{GB} + G_{BB}) \quad (19)$$

$$\epsilon_H = B_{GG} + G_{GB} + G_{BG} + B_{BB} \quad (20)$$

The proposed classifier, which selects the largest volume within a partition, classified partition GB as B, partition BB as G, and partition BG as B. This implies that

$$G_{GB} \leq B_{GB}, G_{BG} \leq B_{BG}, \text{ and } B_{BB} \leq G_{BB} \quad (21)$$

Assume that  $\epsilon_x \leq \epsilon_H$ . This implies that

$$B_{GB} + B_{GG} + G_{BB} + G_{BG} \leq B_{GG} + G_{GB} + G_{BG} + B_{BB} \quad (22)$$

Eliminating common terms yields

$$B_{GB} + G_{BB} \leq G_{GB} + B_{BB} \quad (23)$$

Since the volumes are all positive quantities, this contradicts Equation 21. Therefore,  $\epsilon_x > \epsilon_H$ . Similarly, assume that  $\epsilon_y \leq \epsilon_H$ . This implies that

$$B_{GG} + B_{BG} + G_{GB} + G_{BB} \leq B_{GG} + G_{GB} + G_{BG} + B_{BB} \quad (24)$$

Eliminating common terms yields

$$B_{BG} + G_{BB} \leq G_{BG} + B_{BB} \quad (25)$$

Since the volumes are all positive quantities, this contradicts Equation 21. Therefore,  $\epsilon_y > \epsilon_H$ . Since  $\epsilon_H < \epsilon_x$  and  $\epsilon_H < \epsilon_y$ ,  $\epsilon_H$  is less than the smaller of  $\epsilon_x$  and  $\epsilon_y$ . A similar argument can be used to show that similar contradictions also result when partition GG is classified as B. *QED.*

By using the lemmas that were proven in the previous section, it is possible to construct a proof for Theorem 1.

**Theorem 1:** Suppose that two features  $x$  and  $y$  are used to discriminate between patterns belonging to two classes M and B for which the class-conditional bivariate probability density functions are known. Then  $\epsilon_\beta \leq \epsilon_H \leq \min(\epsilon_x, \epsilon_y)$ .

**Proof:** For a two class and two feature scenario, there are 16 candidate arrangements of partition classifications. In accordance with Lemma 1, ten of those arrangements are inconsistent with the properties of any arbitrary likelihood function and can therefore never occur. This leaves the six arrangements shown in Figure 4. In accordance with Lemma 2 through 4, these arrangements yield the following probability of error  $\epsilon_H$ :

$$\epsilon_H \leq \min(\epsilon_x, \epsilon_y) \quad (26)$$

The Bayesian classifier is optimal in the sense that it has the lowest possible probability of error  $\epsilon_\beta$  for a given set of probability density functions. Therefore,

$$\epsilon_\beta \leq \epsilon_H \leq \min(\epsilon_x, \epsilon_y) \quad (27)$$

*QED.*

## 4. Conclusions

The use of multiple features by a classifier often leads to a reduced probability of error, but the design of an optimal Bayesian classifier for multiple features requires a design sample size that, in general, increases exponentially with the number of dimensions. This project explores a method of combining the classification decisions of multiple classifiers, each utilizing a different subset of the set of features, into a single decision. The current research has focused on the restricted problem of classifying two classes given two features. It has been proven that, for this restricted

problem, the method always demonstrates a probability of error that is greater than or equal to the probability of error of the optimal joint Bayesian classifier and less than or equal to the probability of error of the marginal classifier with the lowest probability of error.

## References

1. Bellman, R., 1961. *Adaptive Control Processes: A Guided Tour*. New Jersey: Princeton University Press.
2. Bloch, I., January 1996. "Information Combination Operators for Data Fusion: A Comparative Review With Classification" in *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 26, no. 1, pp. 52-67.
3. Buade, D., and P. Girardi, 1997. "A Target Identification Comparison of Bayesian and Dempster-Shafer Multisensor Fusion" in *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, Vol. 27, No. 5, pp. 569-577.
4. Dietterich, T., 1997. "Machine Learning Research: Four Current Directions" in *AI Magazine*, Winter 1997, pp. 97-136.
5. Duda, R., and P. Hart, 1973. *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons, Inc.
6. Fukunaga, K., 1990. *Introduction to Statistical Pattern Recognition* (2<sup>nd</sup> ed.). Boston: Academic Press, Inc.
7. Happel, M., and P. Bock, 2000. "Overriding the Experts: A Stacking Method For Combining Marginal Classifiers" in Proceedings of the 13th International FLAIRS Conference. Menlo Park, CA: AAAI Press, forthcoming.
8. Happel, M., 1999. *A Fusion Method for Combining Marginal Classification Decisions using an Override-Capable Classifier* (unpublished dissertation proposal). Washington, DC: The George Washington University.
9. Lam, L., and C. Suen, 1997. "Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance" in *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, Vol. 27, No. 5, pp. 553-568.
10. Mitchell, T., 1997. *Machine Learning*. Boston: McGraw-Hill.
11. Shafer, G., 1976. *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton University Press.
12. Woods, K., W. Kegelmeyer Jr., and K. Bowyer, 1997. "Combination of Multiple Classifiers using Local Accuracy Estimates" in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 405-410.
13. Xu, L., A. Krzyżak, and C. Suen, 1992. "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition" in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 3, pp. 418-435.