On Design and Performance of Metafusers

Nageswara S. V. Rao[a]

[a] Center for Engineering Science Advanced Research
Computer Science and Mathematics Division
Oak Ridge National Laboratory, Oak Ridge, TN 37831
Proceedings of Workshop on Estimation, Tracking and Fusion, 2001.

ABSTRACT

In applications where a fuser for a set of classifiers is to be designed based on finite measurements, it's performance is subject to certain fundamental limitations much like an individual classifier. The fuser may be rendered almost as ineffective as a random guesser by a newer set of problems, even if it performed well on all past classification problems or data sets. Thus it is likely that newer fusers will continue to be developed for the newer problems. The availability of a wide range of fusers makes it difficult to compare them and choose an appropriate one for the application at hand. We propose the concept of a metafuser which combines the available fusers so that its performance is at least as good as the best fuser with a high probability. We show that the isolation property can be a basis for designing such metafusers, which are computed based on measurements. We present a metafuser scheme that can be incrementally maintained to be at least as good as the best classifier as well as best fuser with a specified probability.

1. INTRODUCTION

Over the past decades a bewildering number of classification methods have been developed such as nearest neighbor rules, neural networks, tree methods, support vector machines, and kernel rules [3,2]. These classification methods, simply referred to as classifiers, are quite varied, and their performances are characterized in various ways such as qualitative descriptions, analytical bounds, and empirical errors. Given the data set for a classification problem, one is often overwhelmed by the choice of classifiers and has a tendency to pick "best" classifier from among the available methods. Such choice may be guided by the performance on benchmark problems and/or analytical performance bounds. The No-Free-Lunch theorems establish that on unseen data sets, there is no reason to believe that any one of the methods would be better than others or a random guesser (see Chapter 9 of [3] for a lucid discussion on this topic). More specifically, any classifier could be almost as bad as a random guesser irrespective of its performance on previous data sets including the benchmark problems (Chapter 7 of [2]). This casts a shadow on the benchmarks used to capture the entire class of classification problems, but does not rule out benchmarks for selected subclasses of problems. While it is quite possible that the available classifier methods are adequate for a new data set, no such guarantees can be given. Thus as newer classification problems arise it is likely that newer classifiers will continue to be developed, adding to the wide variety of already existing classifiers.

Concurrent with the recognition of the limitations of general classification methods, a more pragmatic phenomenon has been developing in several communities: remarkable gains have been achieved in the classification performance by fusing or combining various qualitatively disparate classifiers [5]. The efficacy of fusion of classifiers has been demonstration by both analytical and experimental results. It is generally known that a good fuser outperforms the best classifier, and at the same time, a bad fuser choice can result in a performance worse than the worst classifier. Thus it is very important to employ fusion methods that provide concrete performance guarantees

Further author information: (Send correspondence to N. S. V. Rao)
N. S. V. Rao: E-mail: raons@ornl.gov

– in particular, for the fuser to be meaningful it must perform at least as well as the best classifier. If the underlying joint distributions are known the fusion problem has a definite answer: classifiers can be optimally fused by using the available distributed detection methods [19]. Furthermore, if the classifiers are statistically independent, one can employ a linear combination to optimally fuse the outputs of classifiers [1]. In practical classifier systems, however, the design of a fuser is not as direct since independence is seldom satisfied and the underlying distributions are unknown or hard to estimate. One is typically provided a sample, and consequently only an approximation to the optimal fuser is possible. Although the theory of sample-based classifier design has been well developed [2], an analogous theory for fusion of classifiers is developed only to a limited extent. While the optimal fuser can be guaranteed to perform at least as good as the best classifier, an approximation cannot be similarly guaranteed and in fact may not perform well. In an informal sense, the knowledge of distributions provides an infinite dimensional quantity, which in general cannot be provided by a finite sample, thereby imposing fundamental limits on the performance guarantees of a fuser based on the latter. Indeed, the question to fuse or not to fuse is not easily answered in the finite sample case [7]. We provide an answer to this question by utilizing the performance equations of the classifier and fuser, without knowing the distributions. We show that for the fuser to have guarantees at least as good as the best classifier two conditions are needed: (a) the fuser class must satisfy the isolation property which ensures that the optimal fuser performs no worse than the best classifier, and (b) the chosen fuser must be adequately trainable with the given sample.

We will then show that the fusers themselves are susceptible to the fundamental limitations of the classifiers: there is no single fuser that outperforms the others. Consequently, newer fusers will continue to be designed to address newer data sets much like the classifiers. From a practical user perspective, the fusion concept addressed problem of dealing with a myriad of classifiers by simply fusing them. The unintended consequence of this fusion concept is that now there is a proliferation of fusers such as neural networks, nearest neighbors, projective fusers, and isolation fusers [11]. In this paper we invoke the concept of fusion again but this time on the fusers themselves: by suitably combining the fusers we propose the concept of a *metafuser* such that the resultant system provides guarantees at least as good as the best classifier as well as best fuser with a certain probability. Since metafuser also is a classifier, it too is susceptible to fundamental limitations of a classifier or fuser. But the performance gains due to a metafuser are more pragmatic. Given the data for a new classification problem, the metafuser gives an idea about what is possible with all the available classifiers and fusers. Indeed, the metafuser can potentially perform significantly better than the best classifier as well as best fuser. Based on a finite sample, we provide a test for deciding the employment of a metafuser analogous to the case of a fuser.

We restrict our attention to the classifiers and fusers for which distribution independent performance guarantees can be provided. This formulation is based on Vapnik and Chervonenkis theory [18,17], which has been extensively studied recently in the *probably approximately correct* (PAC) learning paradigm [16,20].

We present the problem formulation in Section 2. The we summarize several known properties of single classifiers in Section 3. We present results on fuser design based on the isolation property in Section 4. The limitations of fusers and design of metafusers are discussed in Section 5.

2. PROBLEM FORMULATION

A classical pattern recognition problem is stated as follows: we are given an independently and identically distributed (iid) *n-sample* $D_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, according to an unknown distribution $P_{X,Y}$, where $X_i \in \Re^d$ and $Y_i \in \{0, 1\}$. The problem is to design a classifier $\phi : \Re^d \mapsto \{0, 1\}$ based on n-sample that ensures a small value for the *probability of misclassification* given by

$$L(\phi) = \int_X I_{\{\phi(X) \neq Y\}} dP_{X,Y},$$

where $I_C(x)$ is the *indicator function* of the set $C \subseteq \Re^d$ such that $I_C(x) = 1$ if $x \in C$ and $I_C(x) = 0$ otherwise. We often suppress the operand x when it is clear from the context.

In the formulation based on Vapnik-Chervonenkis theory [17,2], ϕ is chosen from a class $\mathcal{H}$. Since $P_{X,Y}$ is unknown, exact minimization of $L(.)$ is not possible. Instead, we consider the *empirical error of misclassification* given by

$$\hat{L}(\phi) = \frac{1}{n} \sum_{i=1}^{n} I_{\{\phi(X_i) \neq Y_i\}}.$$

Let $\hat{\phi}_n$ minimize $\hat{L}(.)$ over $\mathcal{H}$. If $\mathcal{H}$ has finite Vapnik-Chervonenkis dimension $V_{\mathcal{H}}$, it is well-known [2] that one can guarantee that

$$P^n_{X,Y}\left[L(\hat{\phi}_n) - \min_{\phi \in H} L(\phi) > \epsilon\right] \leq \delta$$

if n is chosen to be sufficiently large, irrespective of the distribution $P_{X,Y}$. This condition asserts that the misclassification error committed by $\hat{\phi}_n$ is within ϵ of the best possible error, namely $L(\phi^*) = \min_{\phi \in H} L(\phi)$, with a probability of at least $1 - \delta$.

We are given N_C such classifiers corresponding to the classes $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{N_C}$ such that

$$P^n_{X,Y}\left[L(\hat{\phi}_{n;i}) - \min_{\phi \in H_i} L(\phi) > \epsilon\right] \leq \delta_i$$

where $\hat{\phi}_{n;i}$ minimizes $\hat{L}(.)$ over $\mathcal{H}_i$. If the classifiers are statistically independent, it is well-known that the higher the number of classifiers the better is the performance of the fused system [4] based on the majority rule. Such result is not true if independence is not satisfied. Thus, the majority rule is no longer adequate, and the performance can indeed deteriorate if the fusers do not account for the lack of independence.

Our first objective is to decide whether the sample be used to just train the individual classifiers or to train a fuser. We show that the performance equations of the classifier and fuser can be used to answer this question in Theorem 4.2. Then we are given N_F fusers $\hat{f}_1, \hat{f}_2, \ldots \hat{f}_{N_F}$ corresponding to the classes $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_{N_F}$. We answer the question of when a metafuser $\hat{\xi}$ that combines the fusers provides better guarantees than the best classifier as well as best fuser in Theorem 5.1.

Our objective is to "fuse" the classifiers and fusers so that the fused system performs at least as well as the best individual classifier and fuser based on the sample only. If the joint error distributions of the classifiers are known, then the fusion problem can be solved using the existing maximum likelihood estimation methods [19]. The main challenge of the present formulation is due to the lack of knowledge of error distributions. Problems of this kind are of relatively recent interest with most works dealing with computing a close-to-optimal fusion rule within a class [12] or sample-based implementation of fusion rules derived for known distributions case [6]. Very few results exist for the present problem that guarantee that the fused system is at least as good as the best classifier or best combination of classifiers (with some exceptions [13,14]).

3. SINGLE CLASSIFIER: LIMITATIONS AND PERFORMANCE BOUNDS

The lowest possible error achievable by any deterministic classifier is given by the *Bayes error* $L^* = L(\phi^*)$, where $\phi^* : \Re^d \mapsto \{0, 1\}$ is defined as

$$\phi^*(x) = \begin{cases} 1 & \text{if } P_{X,Y}[Y = 1|X = x] \geq P_{X,Y}[Y = 0|X = x] \\ 0 & \text{otherwise} \end{cases}$$

In cases, where the distribution $P_{X,Y}$ is not known, ϕ^* cannot be computed. In practice, one is presented with a finite and typically small sample. Then, only an approximation to $L(\phi^*)$ can be computed in general, but there

are fundamental limitations on the performance achievable by any classifier based on finite samples. We now state a theorem from [2] which closely matches our formulation. We note that there are a number of similar results, popularly grouped under the term No-Free-Lunch theorems, and a discussion on them can be found in [3]. Let g_n be a classifier computed based on n-sample such that its error probability is given by

$$L_n(g_n) = P\{Y \neq g_n(X)|D_n\}.$$

The average error probability of g_n is $EL_n(g_n) = L(g_n)$.

THEOREM 3.1. *[2] Let $\epsilon > 0$ be an arbitrary small number. For an integer n and classification rule g_n, there exists a distribution (X, Y) with Bayes risk $L^* = 0$ such that $EL_n \geq 1/2 - \epsilon$.*

This theorem states that for any classifier, there is a distribution (X, Y) for which it is almost as bad as a random guesser, although the data is perfectly separable by the Bayes classifier. The distribution of X used in the proof is discrete and uniform, and the distribution of Y is the main contributor to the error. The distribution depends on n and in this sense there is no single distribution that renders all classifiers useless. Note that the situation is dramatically different for infinite sample sizes, since for a number of estimators such as k-nearest neighbor and cubic histogram, we have $EL_n \to L^*$ as $n \to \infty$.

There are several practical implications of this theorem for finite sample sizes. Different data sets that arise in various applications correspond to different distributions $P_{X,Y}$.

(1) A classifier that performed well on all the data sets that it was tested so far can fail on the next data set.

2) No classifier can be guaranteed to perform uniformly better any other classifier, even if the latter turns out to be only marginally better than a random guesser.

As a result, the quest for classifiers better than the available ones will continue as newer data sets are encountered.

On the positive side, it is quite possible that a newer classification problem is solved well by the existing classification methods. Also, different classification methods will offer different strengths, and the distributions that render one classifier useless may not do the same to a different classifier. Now the question is how one would make use of the wealth of classifiers already in existence to design one for a new data set.

While no finite sample guarantees can be given in EL_n, one can provide weaker guarantees under certain additional conditions. In particular, the performance of $\hat{\phi}_n$ that minimizes $\hat{L}(.)$ can be characterized by the properties of $\mathcal{H}$ and the sample size. We now present a method based on the Vapnik-Chervonenkis dimension. Let $\mathcal{A}$ be a collection of measurable sets of R^d. For $(z_1, z_2, \ldots, z_n) \in \{\Re^d\}^n$, let $\mathcal{N}_{\mathcal{A}}(z_1, z_2, \ldots, z_n)$ denote the number of different sets in

$$\{\{z_1, z_2, \ldots, z_n\} \cap A : A \in \mathcal{A}\}.$$

The nth *shatter coefficient* of $\mathcal{A}$ is

$$s(\mathcal{A}, n) = \max_{(z_1, z_2, \ldots, z_n) \in \{\Re^d\}^n} \mathcal{N}_{\mathcal{A}}(z_1, z_2, \ldots, z_n).$$

Then, the *Vapnik-Chervonenkis* (VC) dimension of $\mathcal{A}$, denoted by $V_{\mathcal{A}}$, is the largest integer $k \geq 1$ such that $s(\mathcal{A}, k) = 2^k$. The following important identity [17,20] relates the shatter coefficient to VC dimension:

$$s(\mathcal{A}, n) = \begin{cases} 2^n & \text{if } n \leq V_{\mathcal{A}} \\ 2\frac{n^{V_{\mathcal{A}}}}{V_{\mathcal{A}}!} & \text{if } n > V_{\mathcal{A}} \end{cases}$$

Then we have the following result

$$P^n_{X,Y}\left[\sup_{\phi \in \mathcal{A}} |\hat{L}(\phi) - L(\phi)| > \epsilon\right] \leq 8s(\mathcal{A}, n)e^{-n\epsilon^2/32}.$$

which in turn implies that

$$P^n_{X,Y}\left[L(\hat{\phi}_n) - \min_{\phi\in H} L(\phi)| > \epsilon\right] \leq 8s(\mathcal{H},n)e^{-n\epsilon^2/128}.$$

Thus, given a sample of size

$$n = \frac{128}{\epsilon^2}\left(\ln s(\mathcal{H},n) + \ln(8/\delta)\right)$$

we have

$$P^n_{X,Y}[L(\hat{\phi}_n) - \min_{\phi\in\mathcal{H}} L(\phi) > \epsilon] < \delta,$$

irrespective of the distribution $P_{X,Y}$. Here ϵ is called the *precision* and δ which is a function of ϵ, n and $\mathcal{H}$ is called the *confidence*. When needed we explicitly show the dependence of δ on ϵ, n and H by representing it as a function $\delta_{\hat{\phi}_n}(\epsilon, n, \mathcal{H})$. We also suppress some of the operands of $\delta_{\hat{\phi}_n}(.)$ when the choice is clear from the context.

We consider a generic form for the confidence of a classifier $\hat{g}$ chosen from the family $\mathcal{G}$ based on n-sample given by

$$\delta_{\hat{g}}(\epsilon, n, \mathcal{G}) = \Delta_{\hat{g}}(n)e^{-a\epsilon^2 n}.$$

In the above case we have $\Delta_{\hat{g}}(n) = 8s(\mathcal{H},n)$ and $a = 1/128$. This is typical of other formulations such as finite hypothesis classes, linear combinations, vector spaces, and neural networks [2] as shown below.

H	$\Delta_{\hat{g}}$	a
finite	$\lvert H\rvert$	$\frac{1}{128}$
linear combinations with d terms	$8n^d + 1$	$\frac{1}{128}$
vector space with dimensionality d_v	$8n^{d_v} + 1$	$\frac{1}{128}$
threshold neural networks with k hidden nodes and d inputs	$8(ne)^{kd+2k+1}$	$\frac{1}{128}$

4. ISOLATION PROPERTY: TO FUSE OR NOT TO FUSE

We first consider a family of fuser functions $\mathcal{F} : \{f : \{0,1\}^N \mapsto \{0,1\}\}$ such that the fused output is given by $f[\hat{\phi}_1(X), \hat{\phi}_2(X), \ldots, \hat{\phi}_{N_C}(X)]$, denoted by $f(Z)$, where $Z = (\hat{\phi}_1(X), \hat{\phi}_2(X), \ldots, \hat{\phi}_{N_C}(X))$. The error probability of the fused system is given by

$$L_F(f) = \int I_{\{f(Z)\neq Y\}} dP_{X,Y}.$$

Note that Z is a deterministic function of X given the sample. For computational convenience, we utilize the following alternative formula

$$L_F(f) = \int [f(Z) - Y]^2 dP_{X,Y}.$$

Note that $|\mathcal{F}| \leq 2^{2^{N_C}}$ since $\mathcal{F}$ consists of at most all Boolean functions on N_C variables. Consider the function class

$$\mathcal{G} = \{f(\phi_1(X), \phi_2(X), \ldots, \phi_{N_C}(X)) : \phi_1 \in \mathcal{H}_1, \phi_2 \in \mathcal{H}_2, \ldots, \phi_N \in \mathcal{H}_{N_C}\}.$$

Here $f(\phi_1(.), \phi_2(.), \ldots, \phi_{N_C}(.))$ specifies a subset of $\Re^d$, and hence $\mathcal{G}$ specifies a family of sets of $\Re^d$.

We now consider that we are given an n-sample. Then there are two ways of utilizing the sample: (a) train the classifiers using the entire sample, or (b) use part of the sample to train the classifiers and the other part to train the fuser. The choice depends on the performance equations of the classifiers and the fuser.

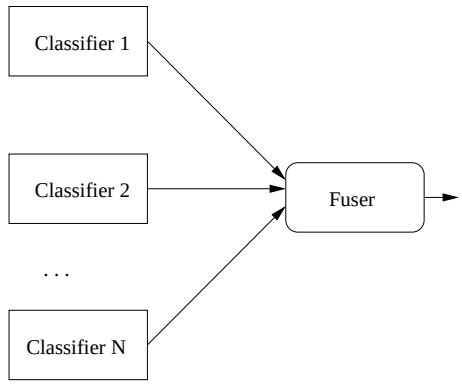

Figure 1. Single fuser.

Consider that each classifier is trained with n-sample, and for simplicity $\hat{\phi}_{n;i}$ and $\delta_{\hat{\phi}}(\epsilon, n, \mathcal{H})$ are represented by $\hat{\phi}_i$ and $\delta_{\hat{\phi}}(n)$ respectively. If one of the classifier is to be chosen, the lowest achievable error is given by $\min_{i=1}^{N_C} L(\phi_i^*)$. Since the classifiers can be correlated in an arbitrary manner, the empirically best classifier $\hat{\phi}_{\min} = \arg\min_i \hat{L}(\hat{\phi}_i)$ yields the following guarantee

$$P_{X,Y}^n \left[L(\hat{\phi}_{\min}) - \min_{i=1}^{N_C} L(\phi_i^*) > \epsilon \right] < \delta_{\hat{\phi}_1}(n) + \delta_{\hat{\phi}_2}(n) + \ldots + \delta_{\hat{\phi}_{N_C}}(n),$$

based on n-sample.

Consider that we split n-sample into two disjoint n_1-subsample and n_2-subsample, such that $n_1 + n_2 = n$. Fuser is obtained in two steps once the classifiers have been trained with n_1-sample: (a) a training set (Z_1, Y_1), (Z_2, Y_2), $\ldots$, (Z_{n_2}, Y_{n_2}), where $Z_i = (\hat{\phi}_1(X_i), \hat{\phi}_2(X_i), \ldots, \hat{\phi}_N(X_i))$, is derived from the classifiers and n_2-sample, and (b) the fuser is derived by minimizing empirical error over $\mathcal{F}$ based on this training set. Let f^* minimize $L_F(.)$ over $\mathcal{F}$. Note that f^* cannot be exactly computed since $P_{X,Y}$ is unknown. Instead, we minimize the empirical error given by

$$\hat{L}_F(f) = \frac{1}{n} \sum_{i=1}^{n} [f(Z_i) - Y_i]^2.$$

Let $\hat{f}$ minimize $\hat{L}_F(.)$ over $\mathcal{F}$. The fuser, thus, provides a *better guarantee* if $\delta_F(n_2) < \delta_{\hat{\phi}_1}(n) + \delta_{\hat{\phi}_2}(n) + \ldots + \delta_{\hat{\phi}_{N_2}}(n)$ where

$$P_{X,Y}^n \left[L_F(\hat{f}) - \min_{i=1}^{N} L(\phi_i^*) > \epsilon \right] < \delta_F(n_2).$$

We now discuss a method to guarantee this condition which requires that the fuser class satisfies two properties. The first property is as follows:

Definition 1. The fuser class $\mathcal{F}$ satisfies the *isolation property* [15,10] if it contains the following N_C functions: for all $i = 1, 2, \ldots, N_C$ we have $f_i(z_1, z_2, \ldots, z_N) = z_i$.

Several well-known classes satisfy the isolation property [10]. This property is trivially satisfied if $\mathcal{F}$ consists of all Boolean functions of N variables. The linear combinations and projective fusers [9] satisfy this property. Although it is sufficient to include N functions in $\mathcal{F}$ to satisfy this property, in general a richer class performs better in practice. We first present a result when $\mathcal{F}$ is finite.

THEOREM 4.1. *Consider that $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_{N_C}$ are the classifiers chosen from classes $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{N_C}$, respectively. Let the finite fuser class $\mathcal{F}$ satisfy the isolation property. Then fuser $\hat{f}$ trained on n_2-subsample of n-sample, provides better guarantee than the best classifier under the condition*

$$|\mathcal{F}| \leq \frac{1}{2} \sum_{i=1}^{N} \delta_i(n) e^{a\epsilon^2 n_2}.$$

Proof: We first have

$$\begin{aligned} L_F(f^*) &= \min_{f \in \mathcal{F}} \int [f(Z) - Y]^2 dP_{X,Y} \leq \min_{i=1}^{N_C} \int [f_i(Z) - Y]^2 dP_{X,Y} \\ &\leq \min_{i=1}^{N_C} \int [\phi_i^*(X) - Y]^2 dP_{X,Y} = \min_{i=1}^{N_C} \int I_{\{\phi_i^*(X) \neq Y\}} dP_{X,Y} = \min_{i=1}^{N_C} L(\phi_i^*) \end{aligned}$$

where the third step is a direct consequence of the isolation property. Consequently the event $\{L_F(\hat{f}) - \min_{i=1}^{N_C} L(\phi_i^*) > \epsilon\}$ implies the event $\{L_F(\hat{f}) - L_F(f^*) > \epsilon\}$. Thus we have

$$P_{X,Y}^n \left[L_F(\hat{f}) - \min_{i=1}^{N_C} L(\phi_i^*) > \epsilon \right] \leq P_{X,Y}^n \left[L_F(\hat{f}) - L_F(f^*) > \epsilon \right] \leq 2|\mathcal{F}| e^{-a\epsilon^2 n_2}$$

where the last step is due to the finiteness of $|\mathcal{F}|$ [17]. □

A minimal realization of this theorem can be based on $\mathcal{F} = \{f_1, f_2, \ldots, f_{N_C}\}$ as per the isolation property defined above. In this case $L(f^*) = \min_{i=1}^{N_C} L(\phi_i^*)$, but if $\mathcal{F}$ is chosen to be much larger, then $L(f^*)$ could be significantly smaller. We wish to emphasize that this fusion method can be easily applied without identifying the best classifier, while still ensuring its performance in the fused system.

The above result is not very useful if $\mathcal{F}$ is not finite, for example as in the case of linear fusers. Also, while it is possible to select $\hat{f}$ by examining all members of $\mathcal{F}$, such method is not computation is not practical. Instead, $\mathcal{F}$ is often embedded in a continuously indexed space where gradient-style algorithms can be employed. In particular, for linear combinations, the computation problem is polynomial-time solvable [8]. Also in several cases the classifiers output is a number from the interval $[0, 1]$. To account for both the cases we consider that $\mathcal{F} : \{f : [0, 1]^{N_C} \mapsto [0, 1]\}$. In this case the fuser $\hat{f} \in \mathcal{F}$ is characterized similar to a classifier by $\delta_{\hat{f}}(n) = \Delta_{\hat{f}}(n) e^{-a\epsilon^2 n}$. Then the following result follows along the lines of Theorem 4.1.

THEOREM 4.2. *Consider that $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_{N_C}$ are the classifiers chosen from classes $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{N_C}$, respectively. If the fuser class $\mathcal{F}$ satisfies the isolation property, then fuser $\hat{f}$ trained on n_2-subsample of n-sample, provides better guarantee than the best classifier under the condition*

$$\Delta_{\hat{f}}(n_2) \leq \sum_{i=1}^{N_C} \Delta_{\hat{\phi}_i}(n) e^{-a\epsilon^2 n_1}.$$

Proof: The first part of the proof is identical to that in Theorem 4.1. For the second part we have

$$P_{X,Y}^n \left[L_F(\hat{f}) - \min_{i=1}^{N_C} L(\phi_i^*) > \epsilon \right] \leq P_{X,Y}^n \left[L_F(\hat{f}) - L_F(f^*) > \epsilon \right] \leq \Delta_{\hat{f}}(n_2) e^{-a\epsilon^2 n_2}$$

which yields the required inequality. □

To answer to question to fuse or not, the condition of the above theorem can be converted into that on the sample size given by

$$n_1 < \frac{1}{a\epsilon^2} \ln \left[\frac{\sum_{i=1}^{N_C} \Delta_{\hat{\phi}_i}(n)}{\Delta_{\hat{f}}(n - n_1)} \right].$$

Note that depending on the performance guarantees of the fuser and classifiers the right hand side might turn out to be larger than n, in which case no fuser is utilized – we will simply pick the classifier with the lowest training error. If this condition is satisfied for some $n_1 < n$ then the fuser is employed. Note that this decision can be made based on sample size and the performance bounds of the classifiers and fusers.

Intuitively speaking, the isolation property guarantees that the error of the best fuser from $\mathcal{F}$ is no worse than that of best classifier. The bound on the sample size specifies that a sufficiently good approximation to the best fuser can be computed based on the sample.

5. METAFUSERS

We first discuss limitations of fusers by the application of Theorem 3.1. A fuser is also a classifier and hence is subject to the limitations discussed in Section 3: there is no single best fuser and newer data sets likely require new fusion methods. This is evidenced by a myriad of fusers developed including nearest neighbor rules, feedforward sigmoid networks, vector space methods, isolation fusers and projective fusers [11].

By extending the fusion concept one level higher, we develop a metafuser which fuses the outputs of the fusers as shown in Figure 2. Consider that n-sample is split into n_1-subsample, n_2-subsample and n_3-subsample such that $n_1 + n_2 + n_3 = n$. Consider that we are given N_C classifiers and N_F fusers. A repeated application of the result from Theorem 4.2 yields the following result.

THEOREM 5.1. *Consider that the classifiers $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_{N_C}$ are fused by each of the fusers $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_{N_F}$, which are in turn fused by a metafuser $\hat{\xi}$. If the metafuser class $\mathcal{M}$ satisfies the isolation property and one of the fuser class $\mathcal{F}_k$ satisfies the isolation property, then the metafuser $\hat{\xi}$ provides better guarantee than the best classifier as well as best fuser under the condition*

$$\Delta_{\hat{f}}(n_2) \leq \sum_{i=1}^{N_C} \Delta_{\hat{\phi}_i}(n) e^{-a\epsilon^2 n_1/4}$$

for some fuser $\hat{f}_i$, and

$$\Delta_{\hat{\xi}}(n_3) \leq \sum_{i=1}^{N_F} \Delta_{\hat{f}_i}(n_2 + n_3) e^{-a\epsilon^2 n_2/4}.$$

Proof: The first condition is same as in Theorem 4.2 except the precision is $\epsilon/2$ instead of ϵ. Let ξ^* minimize $L_F(.)$ over $\mathcal{M}$. Then we have,

$$P^n_{X,Y}\left[L_F(\hat{f}_k) - L_F(f^*_k) > \epsilon/2\right] \leq \Delta_{\hat{f}_i}(n_2) e^{-a\epsilon^2 n_2/4}$$

$$P^n_{X,Y}\left[L_F(\hat{\xi}) - L_F(\xi^*) > \epsilon/2\right] \leq \Delta_{\hat{\xi}}(n_3) e^{-a\epsilon^2 n_3/4}$$

By the isolation property of $\mathcal{M}$ and $\mathcal{F}_k$, we have

$$L_F(\xi^*) \leq \min_{j=1}^{N_F} L_F(f^*_j) \leq L_F(f^*_k) \leq \min_{i=1}^{N_C} L(\phi^*_i).$$

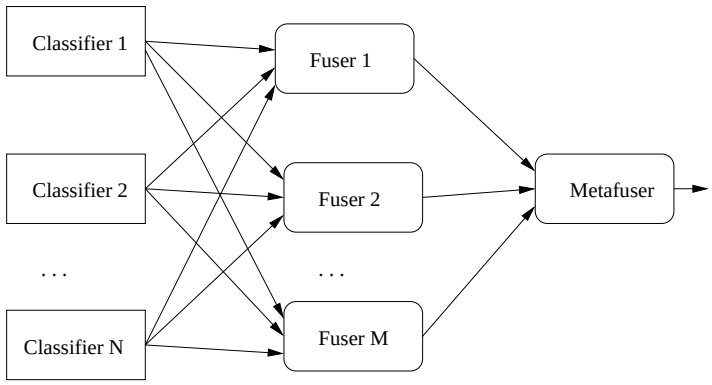

Figure 2. Concept of metafuser.

Using the first inequality and the arguments of Theorem 4.2, $\hat{\xi}$ has better guarantee than $\hat{f}_i$ with precision $\epsilon/2$ and hence with precision ϵ. Since

$$\begin{aligned} |L_F(\hat{\xi}) - \min_{i=1}^{N_C} L(\phi_i^*)| &\leq |L_F(\hat{\xi}) - \min_{j=1}^{N_F} L(f_j^*)| + |\min_{j=1}^{N_F} L(f_j^*) - \min_{i=1}^{N_C} L(\phi_i^*)| \\ &\leq |L_F(\hat{\xi}) - \min_{j=1}^{N_F} L(f_j^*)| + |L(f_k^*) - \min_{i=1}^{N_C} L(\phi_i^*)| \end{aligned}$$

we have

$$P_{X,Y}^n \left[L_F(\hat{\xi}) - \min_{j=1}^{N_F} L(f_j^*) > \epsilon/2 \right] \leq P_{X,Y}^n \left[L_F(\hat{\xi}) - \min_{i=1}^{N_C} L(\phi_i^*) > \epsilon \right] \leq \Delta_{\hat{\xi}}(n_3) e^{-a\epsilon^2 n_3},$$

which yields the second inequality, and ensures that the metafuser‘s guarantees is at least as good as that of best classifier. □

The two inequalities in this theorem can be used to obtain the sample sizes needed to train the classifiers, fusers, and metafuser. A practical paradigm is to employ all available classifiers and fusers as shown in Figure 2 such that at least one of the fusers and the metafuser both satisfy the isolation property. The metafuser could be based on a simple linear combination so that it can be computed very efficiently [10]. When a new classifier is developed it can be incorporated into this scheme by simply retaining the isolation fuser and the metafuser; the resultant system is at least as good as the currently available classifiers in terms of guarantees. If a new fuser is available, it can be simply integrated into the scheme by retraining the metafuser. Thus the performance guarantee of the metafuser is at least as good as that of best classifier as well as fuser with a probability. This status can be incrementally maintained along with the development of new classifiers and fusers.

6. CONCLUSIONS

In classification problems where a fuser is to be designed based on finite measurements, we showed that it’s performance is subject to fundamental limitations: it may be almost as ineffective as a random guesser for a newer set of problems. This result is true even if the fuser performed uniformly well on all past data sets. Thus one could expect that newer fusers will continue to be developed for the newer problems much like the classifiers themselves. Thus a practitioner must contend with a wide range of fusers. We proposed a metafuser to combine the available fusers to probabilistically ensure performance is at least as good as the best fuser. These metafusers themselves are computed based on measurements using the performance equations of the individual classifiers and fusers. We provided

tests to analytically determine if a fuser and metafuser can provide better performance guarantees compared to best classifier. We also derived finite sample performance bounds for the metafusers.

A major limitation of this work is the utilization of performance bounds to make decisions about deploying the fusers and metafuser. In practice the bounds could be rather loose and perhaps do not provide the same degree of approximation. Indeed, it is quite possible that a classifier with a tighter bound but with worse performance may be preferred over the one with a looser bound. But, in practice the bounds are about the only performance equations known in general for the classifiers and fusers. It is temping to derive results analogous to Theorem 5.1 based on actual performance equations rather than bounds. Such results are of limited practical use since the actual performance equations are mostly unknown. We split the sample to train the fusers and metafusers. It is possible to use the entire sample to train the classifiers and fusers as in [10]. For this method the conditions derived in the present paper are not valid. The derivation of the required performance equations for the fusers is technically more involved, which is a topic of future investigation.

ACKNOWLEDGMENTS

This research is sponsored by the Engineering Research Program of the Office of Basic Energy Sciences, U.S. Department of Energy, under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC, Office of Naval Research under order No. N00014-96-F-0415, and Ballistic Missile Defense Organization under MIPR No. 0100568954.

REFERENCES

1. C. K. Chow. Statistical independence and threshold functions. *IEEE Trans. Electronic Computers*, EC-16:66–68, 1965.
2. L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.
3. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, Inc., 2001. Second Edition.
4. B. Grofman and G. Owen, editors. *Information Pooling and Group Decision Making*. Jai Press Inc., Greenwich, Connecticut, 1986.
5. J. Kittler and F. Roli, editors. *Multiple Classifier Systems*, volume 1857. Springer-Verlag, Berlin, 2000.
6. N. S. V. Rao. Distributed decision fusion using empirical estimation. *IEEE Transactions on Aerospace and Electronic Systems*, 33(4):1106–1114, 1996.
7. N. S. V. Rao. To fuse or not to fuse: Fuser versus best classifier. In *SPIE Conference on Sensor Fusion: Architectures, Algorithms, and Applications II*, pages 25–34. 1998.
8. N. S. V. Rao. Multiple sensor fusion under unknown distributions. *Journal of Franklin Institute*, 336(2):285–299, 1999.
9. N. S. V. Rao. Projective method for generic sensor fusion problem. In *IEEE/SICE/RSJ International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 1–6. 1999.
10. N. S. V. Rao. Finite sample performance guarantees of fusers for function estimators. *Information Fusion*, 1(1):35–44, 2000.
11. N. S. V. Rao. Multisensor fusion under unknown distributions: Finite sample performance guarantees. In A. K. Hyder, editor, *Multisensor Fusion*. Kluwer Academic Pub., 2002.
12. N. S. V. Rao and S. S. Iyengar. Distributed decision fusion under unknown distributions. *Optical Engineering*, 35(3):617–624, 1996.
13. N. S. V. Rao and E. M. Oblow. Majority and location-based fusers for PAC concept learners. *IEEE Trans. on Syst., Man and Cybernetics*, 24(5):713–727, 1994.
14. N. S. V. Rao and E. M. Oblow. N-learners problem: System of PAC learners. In *Computational Learning Theory and Natural Learning Systems, Vol IV: Making Learning Practical*, pages 189–210. MIT Press, 1997.
15. N. S. V. Rao, E. M. Oblow, C. W. Glover, and G. E. Liepins. N-learners problem: Fusion of concepts. *IEEE Transactions on Systems, Man and Cybernetics*, 24(2):319–327, 1994.
16. L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
17. V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.
18. V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.
19. P. K. Varshney. *Distributed Detection and Data Fusion*. Springer-Verlag, 1997.
20. M. Vidyasagar. *A theory of Learning and Generalization*. Springer-Verlag, New York, 1997.