On Design and Performance of Metafusers

Nagi Rao
(Nageswara S.V. Rao)
Center for Engineering Science Advanced Research
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831-6355
email: raons@ornl.gov

May 17, 2001
Workshop on Estimation, Tracking and Fusion
Monterey, CA

This research is sponsored by
Ballistic Missile Defense Organization
Office of Naval Research
U.S. Department of Energy
Office of Science, Engineering Research Program

Presentation Outline

1. **Problem Formulation**
   - 1.1 Single Classifier Performance
   - 1.2 Finite-Sample Limitations
2. **To Fuse or Not To Fuse**
   - 2.1 Set of Classifiers
   - 2.2 Fuser Performance Conditions
3. **Metafusers**
   - 3.1 Fusion of Fusers
   - 3.2 Practical Paradigm
4. **Conclusions**

<u>Classifier</u>

Deterministic Boolean Classifier: $\phi : \Re^d \mapsto \{0, 1\}$

Feature vector: X

Output: Y

Probability of Misclassification:

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

where $I_D(x)$ is the *indicator function* of set $D \subseteq \Re^d$

$$I_D(x) = \begin{cases} 1 & \text{if } x \in C \\ 0 & \text{otherwise.} \end{cases}$$

<u>Objective:</u>

Achieve low values of $L(.)$ based on finite sample: $(X_1, Y_1), \ldots, (X_n, Y_n)$ independently and identically distributed according to <u>unknown</u> distribution $P_{X,Y}$

Bayesian Classifier

Achieves lowest error $L(.)$ of any deterministic classifier
Bayes error $L^* = L(\phi^*)$,

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

Notes:

1. If $P_{X,Y}$ is not known, ϕ^* cannot be computed
2. Regression estimation subsumes classification problem
2. Using finite sample, only an approximation to ϕ^* in possible.

Performance of Classifiers

Classifier: Based on sample with *unknown* distribution

Best is Known in the Limit: $n \to \infty$
Under asymptotic consistency criterion:
— universally consistent methods exist: $L(\phi) \to L(\phi^*)$ as $n \to \infty$
e.g. k-nearest neighbor rule, histogram estimation
—no need to fuse classifiers; just pick one of the above to achieve $L(\phi^*)$

In practice,
we often have a finite sample
— no universal finite sample results

- no best classifier exists for finite sample size
- different classifiers perform well under different conditions

Basic Question:
Can finite sample performance be improved by "fusing" classifiers ?

– Yes, under fairly general conditions and with some effort

Finite Sample Limitations of Classifiers

g_n: classifier based on n-sample
Error probability:

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

Average error probability: $EL_n(g_n) = L(g_n)$.

Devroye (1982):
For $\epsilon > 0$, an integer n and classification rule g_n
there exists a distribution (X, Y) with Bayes risk $L^* = 0$ such that

$$EL_n(g_n) \geq 1/2 - \epsilon$$

Informally, for any classifier, there is a distribution (X, Y)
for which it is almost as bad as a random guesser,
although data is perfectly separable by Bayes classifier.

Implications:

(1) A classifier that performed well on previous data sets can "fail" on next data set.

2) No classifier can perform uniformly better than any other
even if latter is only marginally better than a random guesser.

Result: Quest for better classifiers continues as newer data sets are encountered.

Vapnik-Chervonenkis Theory

General theory with performance guarantees based on finite samples

Hypothesis space: $\mathcal{H} = \{\phi : \Re^d \mapsto \{0, 1\}\}$
Examples: feedforward neural networks, potential functions

Empirical Error of Misclassification:

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

Let $\hat{\phi}$ minimize $\hat{L}(.)$ over $\mathcal{H}$.
— $\hat{\phi}$: empirically best hypothesis

Basic Result: If $\mathcal{H}$ has finite Vapnik-Chervonenkis dimension $V_{\mathcal{H}}$,

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

for sufficiently large n irrespective of distribution $P_{X,Y}$.

—Informally, error of $\hat{\phi}$ is within ϵ of best error $\min_{\phi \in H} L(\phi)$
with probability $1 - \delta$.

Note:
Different $\mathcal{H}$'s result in different best within-class classifiers.

Single Classifier: Performance Bounds

Sample Size Estimate: For sufficiently large sample size n we have

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

irrespective of $P_{X,Y}$

Generic form for δ showing explicit dependence:
from classifier family $\mathcal{G}$ based on n-sample

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

Examples:

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	$8(ne)^{kd+2k+1}$	$\frac{1}{128}$

Performance of Set of Classifiers

Classifiers:

N_C hypothesis classes: $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{N_C}$
Best within class classifiers:

$$\phi_i^* = \min_{\phi_i \in \mathcal{H}_i} L(\phi_i)$$

$$\hat{\phi}_i = \min_{\phi_i \in \mathcal{H}_i} \hat{L}(\phi_i)$$

If single classifier is chosen, lowest achievable error is

$$\min_{i=1}^{N_C} L(\phi_i^*)$$

But, such selection is not possible by finite sample.

Performance Equation:

Given classifiers: $\hat{\phi}_1$, $\hat{\phi}_2$, ..., $\hat{\phi}_{N_C}$:
Classifiers can be correlated in an arbitrary manner:
For empirically best classifier:

$$\hat{\phi}_{\min} = \arg\min_i \hat{L}(\hat{\phi}_i)$$

we have

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

Fuser Computation

Family of fuser functions:
$\mathcal{F} : \{f : [0,1]^{N_C} \mapsto \{0,1\}\}$

Fuser output: $f(Z) = f[\hat{\phi}_1(X), \hat{\phi}_2(X), \ldots, \hat{\phi}_{N_C}(X)]$

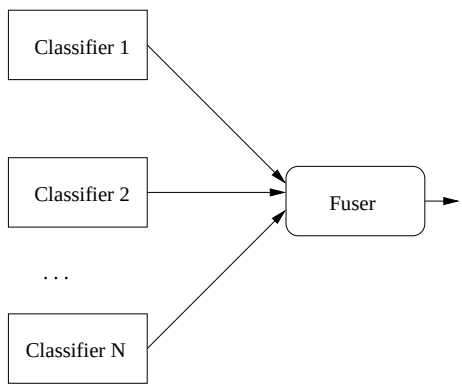


Figure 1: Single fuser.

Fuser Computation:
Split n-sample into two disjoint n_1- and n_2-subsamples;
Two steps procedure:
Step1: Trained the classifiers with n_1-sample:
Step2:
(a) generate a training set (Z_1, Y_1), (Z_2, Y_2), ..., (Z_{n_2}, Y_{n_2})
where $Z_i = (\hat{\phi}_1(X_i), \hat{\phi}_2(X_i), \ldots, \hat{\phi}_N(X_i))$, using n_2-sample, and
(b) minimize empirical error over $\mathcal{F}$ based on this training set.

<u>Relative Performance of Fuser</u>

Error probability of fuser:

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

<u>Fuser $\hat{f}$ provides a better guarantee:</u>
under the condition

$$\delta_F(n_2)\delta_{\hat{\phi}_1}(n) + \delta_{\hat{\phi}_1}(n) + \ldots + \delta_{\hat{\phi}_{N_C}}(n).$$

where

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

Note: Fuser class is finite, i.e. $|\mathcal{F}| \leq 2^{2^N}$

Isolation Fuser

Optimal fuser: f^* minimizes $L_F(.)$ over $\mathcal{F}$
Optimal empirical fuser: $\hat{f}$ minimizes $\hat{L}_F(.)$ over $\mathcal{F}$.

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

Isolation Property:
$\mathcal{F}$ contains the following N functions:
for all $i = 1, 2, \ldots, N$ we have $f_i(z_1, z_2, \ldots, z_N) = z_i$.

Theorem 1. Finite fuser class:
Classifiers $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_{N_C}$ chosen from $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{N_C}$.
Finite fuser class $\mathcal{F}$ has 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 of Theorem 1

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 $|\mathcal{F}|$

Infinite Fuser Classes

Fuser Class: $\mathcal{F} : \{f : [0,1]^{N_C} \mapsto [0,1]\}$
Fuser $\hat{f} \in \mathcal{F}$ is characterized by $\delta_{\hat{f}}(n) = \Delta_{\hat{f}}(n) e^{-a\epsilon^2 n}$.

Theorem 2. Infinite fuser class:
Classifiers: $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_{N_C}$ chosen from $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{N_C}$
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

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

To fuse or not to fuse:

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

Informally speaking,

Fuser class must be:
(i) capable of "picking" best classifier, and
(ii) small enough to be done efficiently

Finite-Sample Limitations of Fusers

A fuser is also a classifier - subject to the fundamental limitations

– there is no single best fuser

– newer data sets might require new fusers

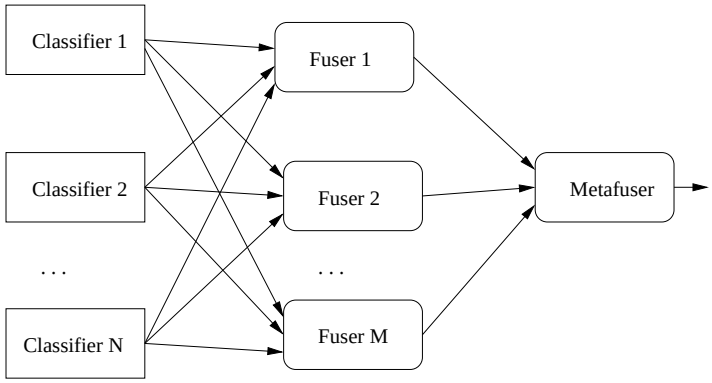
myriad of available fusers: nearest neighbor, sigmoid networks, vector space methods, isolation fusers and projective fusers

What have we done ?

Started out to deal with multiple classifiers,
but ended up creating more fusers !

– Fusion engineer's job is more difficult now

Metafuser



Extension of fusion concept one level higher: metafuser fuses fusers

Computation of Metafuser:

Split n-sample into n_1-, n_2- and n_3-subsamples

1. Train N_C classifiers with n_1-subsample
2. Train N_F fusers with n_2-subsample
3. Train metafuser using n_3-subsample

Performance of Metafuser

Theorem 3.:
classifiers: $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_{N_C}$
fused by each of fusers: $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_{N_F}$,
fused by a metafuser: $\hat{\xi}$
Metafuser class $\mathcal{M}$ has isolation property and
one fuser class $\mathcal{F}_k$ has isolation property.
The metafuser $\hat{\xi}$ has better guarantee than
best classifier as well as best fuser under the conditions

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

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

To fuse or not to fuse:
Two inequalities yield sample sizes to train classifiers, fusers, and metafuser.

Proof

The first condition is same as in Theorem 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

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

Using the first inequality and the arguments of Theorem 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

$$P^n_{X,Y}\left[L_F(\hat{\xi}) - \min_{j=1}^{N_F} L(f^*_j) > \epsilon/2\right] \leq P^n_{X,Y}\left[L_F(\hat{\xi}) - \min_{i=1}^{N_C} L(\phi^*_i) > \epsilon\right] \leq \Delta_{\hat{\xi}}(n_3)e^{-a\epsilon^2 n_3},$$

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

Adaptive paradigm

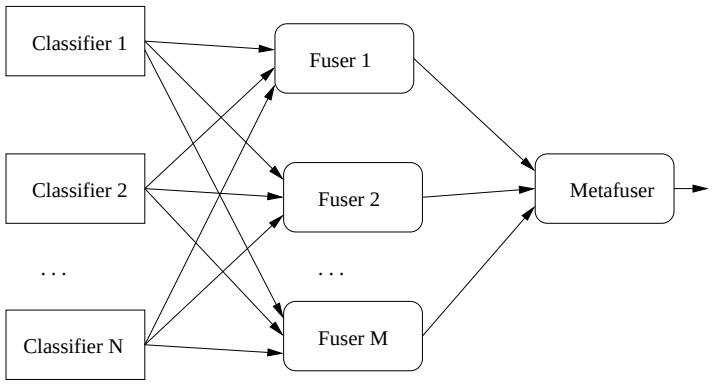
Scheme:
Employ all available classifiers and fusers such that
at least one fuser and metafuser have isolation property.

Simple Metafuser: linear combination
– it can be computed very efficiently

When a new classifier becomes available:
Plug it in, and retrain isolation fuser and metafuser;
– resultant system is at least as good as the current best classifier

If a new fuser becomes available:
Pulg it in, and retrain the metafuser
– system is at least as good as that of best classifier as well as fuser

The guarantee can be incrementally maintained as new classifiers and fusers become available



Conclusions

Summary:

1. Fuser performance is subject to fundamental limitations:
– can be as ineffective as a random guesser for newer problems.
– practitioner must contend with a number of fusers.

2. Proposed metafusers to combine fusers to ensure performance as good as best fuser.
– derived design and performance equations

Perspective:

A limitation: utilization of performance bounds
– bounds could be loose and non-uniform
In practice, bounds are about only performance equations known.

Research Topics:

1. Another Method: use entire sample to train classifiers and fusers
   – we split sample to train the fusers and metafusers
   Derivations are technically more involved
   What are the boundaries of performance between two methods ?
2. Generalization of isolation property