Performances of an ATR System via its ROC Manifold

Mark E. Oxley, Steven N. Thorsen and Kenneth W. Bauer, Jr.
Department of Mathematics and Statistics
Department of Operational Sciences
Air Force Institute of Technology
Wright-Patterson AFB, OH, 5433-7765
mark.oxley@afit.edu, steven.thorsen@afit.edu, kenneth.bauer@afit.edu

Christine M. Schubert
Department of Biostatistics
Virginia Commonwealth University
Richmond, VA, 23298-0032
cmschubert@vcu.edu

***Abstract*— A Classification system such as an Automatic Target Recognition (ATR) system with N possible output labels (or decisions) will have N(N-1) possible errors. The Receiver Operating Characteristic (ROC) manifold was created to quantify all of these errors. Truthed data will produce an approximation to a ROC manifold. How well does the approximate ROC manifold approximate the true ROC manifold? Several functionals exist that quantify the approximation ability, but researchers really wish to quantify the performance in the approximate ROC manifold. This paper will review different performance definitions for ROC curves and manifolds, and thus, quantify the fusion of ATR systems. Examples of different performances will be given that are defined on manifolds.**

Keywords: Performance, Evaluation, Classification System, ROC Manifold, functional

I. Introduction

Given a classification system, how does one quantify its performance? What do we mean by the system's performance? Is there a best performance quantifier to use? This paper will discuss many aspects of performances of a system and a family of systems, and as a consequence, will define the performance of the fusion of systems. Examples of different performances will be given in the Examples section.

II. Mathematical Background

This section gives the essential theory and notation in order to discuss the different performances used to evaluate the fusion of ATR systems, and classification systems.

A. Classification Theory

Let $\mathcal{E}$ be a population set of outcomes. These outcomes can be real-life "events"[1]. An event could be a fixed-time event, a space-time event, or a space-time-spectral event, to name a few examples. Let $\mathfrak{E}$ be a σ-algebra of subsets of $\mathcal{E}$, then $(\mathcal{E}, \mathfrak{E})$ is a measurable space [1]. Let P be a probability measure defined on $\mathfrak{E}$, then $(\mathcal{E}, \mathfrak{E}, P)$ is a probability measure space. Let $\mathbf{s}$ be a sensor that senses an event (i.e., an outcome) and produces (raw) datum as its output, i.e., $\mathbf{s} : \mathcal{E} \to \mathcal{D}$, where $\mathcal{D}$ is a (raw) data set. This data set may be too difficult to quantify directly as it may be a collection or series of images or sequences of audio signals, for instance. Thus, a mapping $\mathbf{p}$ defined on $\mathcal{D}$ produces an object called a feature that is a more refined datum, typically a vector of real numbers. The mapping $\mathbf{p}$ then is a processor that takes a (raw) datum and produces a refined datum vector, i.e., $\mathbf{p} : \mathcal{D} \to \mathcal{F}$. Typically, $\mathcal{F}$ is some finite dimensional space but need not be finite nor a linear space. Let $\mathbf{a}$ be a classifier mapping $\mathcal{F}$ into a label set $\mathcal{L}$. That is, $\mathbf{a} : \mathcal{F} \to \mathcal{L}$. Example of a 2-label set is $\mathcal{L} = \{\text{target, non-target}\}$. Our interest for this paper is a label set with N labels, say $\mathcal{L} = \{\ell_1, \ell_2, \ell_3, \ldots, \ell_n\}$. The composition of these mappings yields a classification system $\mathbf{A} \equiv \mathbf{a} \circ \mathbf{p} \circ \mathbf{s}$. The graphical representation of these mappings is given in the following diagram.

$$\mathcal{E} \xrightarrow{\mathbf{s}} \mathcal{D} \xrightarrow{\mathbf{p}} \mathcal{F} \xrightarrow{\mathbf{a}} \mathcal{L}$$

The diagram for the system is written as

$$\mathcal{E} \xrightarrow{\mathbf{A}} \mathcal{L}$$

Since $\mathcal{L} = \{\ell_1, \ell_2, \ell_3, \ldots, \ell_n\}$ is finite then the power set of $\mathcal{L}$, denoted by $\mathfrak{L}$, is the smallest σ-field of subsets of $\mathcal{L}$. Now define the collection of all measurable systems [1] mapping $\mathcal{E}$ into $\mathcal{L}$, by

$$\mathscr{S} = \{\mathbf{A} : \mathcal{E} \to \mathcal{L} \,|\, \mathbf{A} \text{ is measurable } \}.$$

Let Θ be a set of parameters that might be a multi-dimensional vector of parameters. For each $\theta \in \Theta$ let $\mathbf{a}_\theta$ be an classifier mapping $\mathcal{F}$ into the label set $\mathcal{L}$. That is, $\mathbf{a}_\theta : \mathcal{F} \to \mathcal{L}$ for each $\theta \in \Theta$. The composition of these mappings yields a classification system $\mathbf{A}_\theta \equiv \mathbf{a}_\theta \circ \mathbf{p} \circ \mathbf{s}$. We define the family of the classification systems, or for brevity, the classification system family (CSF), to be $\mathbb{A} \equiv \{\mathbf{A}_\theta : \theta \in \Theta\}$. Thus, $\mathbb{A}$ is a subset of $\mathscr{S}$. We define the collection of families of classification systems to be

$$\mathscr{F} = \{\mathbb{A} \subset \mathscr{S} : \mathbb{A} \text{ is nonempty}\}.$$

B. Two Classification Systems

Consider the case when two sensors, $\mathbf{s}_1$ and $\mathbf{s}_2$, observe events occurring in the same population set $\mathcal{E}$. Assume they produce data in the data sets $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Further, assume each sensor has its own processor, $\mathbf{p}_1$ and $\mathbf{p}_2$, which

[1]In probability theory an event is a set of outcomes. Here we use it in the informal sense.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **JUL 2008**	2. REPORT TYPE	3. DATES COVERED **00-00-2008 to 00-00-2008**
4. TITLE AND SUBTITLE **Performances of an ATR System via its ROC Manifold**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Institute of Technology,Department of Mathematics and Statistics,Department of Operational Sciences,Wright Patterson AFB,OH,45433**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**		
13. SUPPLEMENTARY NOTES **11th International Conference on Information Fusion, June 30 ? July 3, 2008, Cologne, Germany.**		
14. ABSTRACT **see report**		
15. SUBJECT TERMS		

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**6**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

maps datum in $\mathcal{D}_1$ to features in $\mathcal{F}_1$ and $\mathcal{D}_2$ to features in $\mathcal{F}_2$, respectively. In particular, assume $\mathbf{p}_1 : \mathcal{D}_1 \to \mathcal{F}_1$ and $\mathbf{p}_2 : \mathcal{D}_2 \to \mathcal{F}_2$. Suppose there is a family of classifiers for $\mathbf{p}_1$ and $\mathbf{s}_1$ given by $\{\mathbf{a}_\theta : \theta \in \Theta\}$ and another family of classifiers $\{\mathbf{b}_\phi : \phi \in \Phi\}$ for $\mathbf{p}_2$ and $\mathbf{s}_2$, outputting labels in the label set $\mathcal{L}$. Thus, $\mathbf{a}_\theta : \mathcal{F}_1 \to \mathcal{L}$ for each $\theta \in \Theta$ and $\mathbf{b}_\phi : \mathcal{F}_2 \to \mathcal{L}$ for each $\phi \in \Phi$. The composition of these mappings yield classification systems represented by the diagram.

$$\begin{tikzcd}
 & \mathcal{D}_1 \arrow[r, "\mathbf{p}_1"] & \mathcal{F}_1 \arrow[r, "\mathbf{a}_\theta"] & \mathcal{L} \\
\mathcal{E} \arrow[ur, "\mathbf{s}_1"] \arrow[dr, "\mathbf{s}_2"'] & & & \\
 & \mathcal{D}_2 \arrow[r, "\mathbf{p}_2"] & \mathcal{F}_2 \arrow[r, "\mathbf{b}_\phi"] & \mathcal{L}
\end{tikzcd}$$

Now define the system $\mathbf{A}_\theta \equiv \mathbf{a}_\theta \circ \mathbf{p}_1 \circ \mathbf{s}_1$ for each $\theta \in \Theta$ and $\mathbf{B}_\phi \equiv \mathbf{b}_\phi \circ \mathbf{p}_2 \circ \mathbf{s}_2$ for each $\phi \in \Phi$, and denote the two classification system families $\mathbb{A} \equiv \{\mathbf{A}_\theta : \theta \in \Theta\}$ and $\mathbb{B} \equiv \{\mathbf{B}_\phi : \phi \in \Phi\}$.

The two classification systems developed above map outcomes from the population set into different data, feature, and label sets, which are then used to fuse the classification systems together. There are, however, other ways to label the outcomes from the event set. In this discussion, classification systems can map outcomes into either the same or different data sets or the same or different feature sets. The sets which must remain the same for the mathematical development contained herein are the event set $\mathcal{E}$ and the two-class label set $\mathcal{L}$. Therefore, the classification systems must be acting from the same event set, map into either the same or different data and feature sets and eventually map into the same label set. That is,

$$\begin{tikzcd}
 & \mathcal{L} \\
\mathcal{E} \arrow[ur, "\mathbf{A}_\theta"] \arrow[dr, "\mathbf{B}_\phi"'] & \\
 & \mathcal{L}
\end{tikzcd}$$

C. Fusion Rules

There are two types of fusion for classification systems. The first type allows for the families of classification systems which are to be fused to have exactly the same label set. We mean exactly the same, and not isomorphic, so that if the label set is, in fact, $\mathcal{L} = \{\text{target, non-target}\}$ for each family, then this means that the actual definition of a target label is identical for each. This allows for each family to partition the population set in the same way. This type of information fusion we call *within fusion;* the other type is called *across fusion* [2]. The diagram for label fusion for two systems is

$$\begin{tikzcd}
 & \mathcal{L} \arrow[dd, dotted, "\mathfrak{R}"] \\
\mathcal{E} \arrow[ur, "\mathbf{A}_\theta"] \arrow[dr, "\mathbf{B}_\phi"'] & \\
 & \mathcal{L}
\end{tikzcd}$$

no matter which type of fusion it is. Given two CSFs $\mathbb{A}$ and $\mathbb{B}$ and a fusion rule $\mathfrak{R}$, then a new family $\mathbb{C}$ is produced and defined by

$$\mathbb{C} = \mathfrak{R}(\mathbb{A}, \mathbb{B}) \equiv \{\mathfrak{R}(\mathbf{A}_\theta, \mathbf{B}_\phi) : \theta \in \Theta, \phi \in \Phi\}.$$

D. Receiver Operating Characteristic (ROC) Curves

For a 2-class label set $\mathcal{L} = \{t, n\}$, (t denotes target and n denote nontarget) the errors are false positive (type I error, α) and false negative (type II error, β). Let $P_{FP}(\mathbf{A}_\theta)$ denote the probability that the classification system $\mathbf{A}_\theta$ labels an event as a target label, t, given that the event is really a non-target event. Let $P_{FN}(\mathbf{A}_\theta)$ denote the probability of false negative classification by the system $\mathbf{A}_\theta$, then $P_{FN}(\mathbf{A}_\theta)$ is the probability that the classification system $\mathbf{A}_\theta$ labels an event as a non-target label, n, given that the outcome is really a target event. The ROC curve is the graph of the ROC function.

Definition 1: (ROC function, ROC curve) Let $\mathbb{A} = \{\mathbf{A}_\theta : \theta \in \Theta\}$ be a family of classification systems defined on the probability space $(\mathcal{E}, \mathfrak{E}, P)$ mapping to the label set $\mathcal{L} = \{t, n\}$ with parameter set Θ. For each $p \in [0, 1]$, define the set

$$\Theta_p \equiv \{\theta \in \Theta : P_{FP}(\mathbf{A}_\theta) \leq p\}.$$

For $p \in [0, 1]$, if Θ_p is nonempty then define

$$f_{\mathbb{A}}(p) = \max\{P_{TP}(\mathbf{A}_\theta) : \theta \in \Theta_p\}. \tag{1}$$

If Θ_p is empty then $f_{\mathbb{A}}(p)$ is not defined. The function $f_{\mathbb{A}}$ is called the *ROC function*. The graph of $f_{\mathbb{A}}$ is called the *ROC curve*.

Since every classification system family will have a ROC curve (determined by the parameter set), then there is a mapping $\digamma$ that take a CSF $\mathbb{A}$ and produces its ROC curve $f_{\mathbb{A}}$. That is, $\digamma(\mathbb{A}) = f_{\mathbb{A}}$.

E. ROC Manifolds

Assume the label set $\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_n\}$ where $n > 2$, and the classification system $\mathbf{A} : \mathcal{E} \to \mathcal{L}$ is designed to map the outcomes in the event set $\mathcal{E}_i \subset \mathcal{E}$ to ℓ_i for each $i = 1, ..., n$. Define the probability of true positive classification for a given label ℓ_i of the classification system $\mathbf{A}$ by the conditional probability

$$P_{i|i}(\mathbf{A}) \equiv \Pr\{\mathbf{A}(e) = \ell_i \mid e \in \mathcal{E}_i\} = \frac{\Pr\left(\mathbf{A}^\natural(\{\ell_i\}) \cap \mathcal{E}_i\right)}{\Pr\left(\mathcal{E}_i\right)}.$$

The probability that system $\mathbf{A}$ classifies an outcome as label ℓ_i when the outcome is truly classified as label ℓ_j, is

$$P_{i|j}(\mathbf{A}) = \Pr\{\mathbf{A}(e) = \ell_i \mid e \in \mathcal{E}_j\} = \frac{\Pr\left(\mathbf{A}^{\natural}(\{\ell_i\}) \cap \mathcal{E}_j\right)}{\Pr\left(\mathcal{E}_j\right)}. \tag{2}$$

We use the notation $P_{i|j}(\mathbf{A})$ to convene the fact that $P_{i|j}$ is a real-valued function with the system $\mathbf{A}$ as its input. The conjunctive equations of the system are

$$\sum_{i=1}^{n} P_{i|j}\left(\mathbf{A}\right) = 1 \text{ for each } j = 1, 2, \ldots, n \tag{3}$$

and are true for every system $\mathbf{A} : \mathcal{E} \to \mathcal{L}$ [3], [4]. Only the $i|i$ terms are correct classifications, the other $n-1$ terms are the errors of system $\mathbf{A}$ and, consequently, from equations (3) we have

$$\sum_{i=1, i\neq j}^{n} P_{i|j}(\mathbf{A}) = 1 - P_{j|j}(\mathbf{A}) \quad \text{for each } j = 1, 2, \ldots, n. \tag{4}$$

For system $\mathbf{A}$ define the $n \times n$ matrix $\mathtt{P}(\mathbf{A})$ to be the matrix whose i, j entry is the value $P_{i|j}(\mathbf{A})$ for every $i, j \in \{1, ..., n\}$, that is,

$$\mathtt{P}(\mathbf{A})_{i,j} = P_{i|j}(\mathbf{A}) = \frac{\Pr(\mathbf{A}^{\natural}(\{\ell_i\}) \cap \mathcal{E}_j)}{\Pr(\mathcal{E}_j)}.$$

Notice that the diagonal entries of the matrix $\mathtt{P}(\mathbf{A})$ are the correct classifications and the off-diagonal entries are the errors associated with misclassification. By property (3) the transposed matrix $\mathtt{P}(\mathbf{A})^{\mathsf{T}}$ is a stochastic matrix. Also, all the entries of this matrix have values lying in the interval $[0, 1]$. Let $\mathcal{M}_n$ denote the set of $n \times n$ matrices whose entries lie in $[0, 1]$, that is,

$$\mathcal{M}_n = \{\mathtt{M} = (\mathtt{M}_{i,j}) : \mathtt{M}_{i,j} \in [0, 1] \text{ for every } i, j \in \{1, 2, , .., n\}\}$$

then $\mathtt{P}(\mathbf{A}) \in \mathcal{M}_n$. Define the matrix $\mathtt{J} \in \mathcal{M}_n$ by

$$\mathtt{J} = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ 1 & 1 & 0 & & 1 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 0 \end{bmatrix}.$$

Matrix $\mathtt{J}$ will be used to remove the correct classifications and keep only the errors of the system. Specifically, let $\widetilde{\mathtt{P}}(\mathbf{A})$ denote the $n \times n$ matrix given by the Hadamard product with $\mathtt{J}$

$$\widetilde{\mathtt{P}}(\mathbf{A}) = \mathtt{J} \odot \mathtt{P}(\mathbf{A}).$$

Let $\mathcal{Z}_n$ denote the set of matrices in $\mathcal{M}_n$ with zero diagonal entries and off-diagonal entries are real numbers between 0 and 1, that is,

$$\mathcal{Z}_n = \{\mathtt{M} \in \mathcal{M}_n : \mathtt{M}_{i,i} = 0 \text{ for all } i = 1, 2, ..., n\}.$$

Now we define the error set of the classification system family.

Definition 2: (Error set) Given a classification system family $\mathbb{A}$ define its error set $E_{\mathbb{A}}$ to be

$$E_{\mathbb{A}} = \{\widetilde{\mathtt{P}}(\mathbf{A}) : \mathbf{A} \in \mathbb{A}\}.$$

Observe that $E_{\mathbb{A}} \subset \mathcal{Z}_n$ since the diagonal entries are always zero. This set is comprised of many points (matrices), however, we seek those "closest" to the origin (zero matrix) as we define in the following ROC function.

Definition 3: (ROC function) Given a classification system family $\mathbb{A}$ define its ROC function $\Upsilon_{\mathbb{A}}$ to be, for every $\widetilde{\mathbf{P}} \in \mathcal{Z}_n$

$$\begin{aligned} &\Upsilon_{\mathbb{A}}(\widetilde{\mathbf{P}}) \\ = &\begin{cases} \text{smallest } \alpha > 0 & \text{when} \quad \alpha\widetilde{\mathbf{P}} \in E_{\mathbb{A}} \\ \infty & \text{when} \quad \alpha\widetilde{\mathbf{P}} \notin E_{\mathbb{A}} \text{ for all } \alpha \geq 0 \end{cases} \\ = &\min\left\{\alpha \in [0, \infty] : \alpha\widetilde{\mathbf{P}} \in E_{\mathbb{A}}\right\}. \end{aligned}$$

Therefore, $\Upsilon_{\mathbb{A}} : \mathcal{Z}_n \to [0, \infty]$.

The idea of this definition comes from Minkowski's functional (see [5] for its use in optimization.) An equivalent definition of $\Upsilon_{\mathbb{A}}$, useful for computations, is given in the following theorem.

Theorem 1: Given a classification system family $\mathbb{A}$ with n labels, for every $\widetilde{\mathbf{P}} \in \mathcal{Z}_n$, with $\widetilde{\mathbf{P}} \neq \widetilde{\mathbf{0}}$

$$\begin{aligned} &\Upsilon_{\mathbb{A}}(\widetilde{\mathbf{P}}) = \\ &\frac{1}{\left\|\widetilde{\mathbf{P}}\right\|} \min\left\{\left\|\widetilde{\mathbf{Q}}\right\| : \widetilde{\mathbf{Q}} \in E_{\mathbb{A}} \text{ and } \left\langle \widetilde{\mathbf{Q}}, \widetilde{\mathbf{P}} \right\rangle = \left\|\widetilde{\mathbf{Q}}\right\| \left\|\widetilde{\mathbf{P}}\right\|\right\} \end{aligned}$$

and $\Upsilon_{\mathbb{A}}(\widetilde{\mathbf{P}}) = \infty$ otherwise.

The ROC function allows us to define the frontier of the error set, which we call the ROC frontier. The ROC frontier will be the ROC manifold if it satisfies the manifold criterion.

Definition 4: (Manifold) [6] A *manifold* is a topological space that is locally Euclidean, that is, around every point in the space, there is a neighborhood that is topologically the same as the open unit ball in $\mathbb{R}^m$ for some positive integer m.

Definition 5: (ROC frontier, ROC manifold) Given a classification system family $\mathbb{A}$ with n labels, define its ROC frontier $M_{\mathbb{A}}$ to be the set

$$M_{\mathbb{A}} = \{\widetilde{\mathbf{P}} \in \mathcal{Z}_n : \Upsilon_{\mathbb{A}}(\widetilde{\mathbf{P}}) = 1\}.$$

If this set is a manifold, then $M_{\mathbb{A}}$ is called the *ROC manifold* for $\mathbb{A}$.

For $\mathbb{A} = \{\mathbf{A}_\theta : \theta \in \Theta\}$ we assume the parameter set, Θ, is homeomorphic to $\mathbb{R}^m$ for some positive integer m with the usual Euclidean topology. Consequently, the error set, $E_{\mathbb{A}}$, is also homeomorphic to some finite dimensional space, and thus, $M_{\mathbb{A}}$ will be a manifold.

Let $\mathscr{R}$ denote the collection of ROC curves ($n = 2$) or manifolds ($n > 2$). That is,

$$\begin{aligned} \mathscr{R} &= \{f_{\mathbb{A}} : \mathbb{A} \text{ is a CSF defined on } \mathcal{E}\} \quad \text{for} \quad n = 2 \\ &= \{M_{\mathbb{A}} : \mathbb{A} \text{ is a CSF defined on } \mathcal{E}\} \quad \text{for} \quad n > 2 \end{aligned}$$

III. Performances

We choose a real-valued functional ρ that takes a system $\mathbf{A}$ as its input and yields a positive real number as its output. We call $\rho(\mathbf{A})$ the *performance* of $\mathbf{A}$, and ρ is called a *performance functional*. Without loss of generality, we assume that a larger value of $\rho(\mathbf{A})$ is better performance. Consequently, given two systems $\mathbf{A}$ and $\mathbf{B}$, if

$$\rho(\mathbf{A}) \leq \rho(\mathbf{B})$$

then we say $\mathbf{B}$ is better than $\mathbf{A}$ with respect to ρ. This will induce a partial ordering $\overset{\rho}{\preceq}$ on systems in $\mathscr{S}$, and hence, we write

$$\mathbf{A} \overset{\rho}{\preceq} \mathbf{B}.$$

The system performance functional ρ ($\mathscr{S}$-functional for brevity) induces a family performance functional ϱ ($\mathscr{F}$-functional for brevity) on a classification system family $\mathbb{A}$ by the following definition

$$\varrho(\mathbb{A}) = \max_{\mathbf{A}\in\mathbb{A}} \rho(\mathbf{A}) = \max_{\theta\in\Theta} \rho(\mathbf{A}_\theta).$$

Problem 1: Given a performance $\mathscr{F}$-functional ϱ and given a set of label fusion rules $\mathsf{LABrules}$ we seek the best fusion rule $\mathfrak{R}^* \in \mathsf{LABrules}$ such that the performance

$$\varrho\left(\mathfrak{R}^*(\mathbb{A},\mathbb{B})\right) \geq \varrho\left(\mathfrak{R}(\mathbb{A},\mathbb{B})\right)$$

for all choices $\mathfrak{R} \in \mathsf{LABrules}$. That is,

$$\varrho\left(\mathfrak{R}^*(\mathbb{A},\mathbb{B})\right) = \max_{\mathfrak{R}\in\mathsf{LABrules}} \varrho(\mathfrak{R}(\mathbb{A},\mathbb{B})).$$

The optimal classification system family will be $\mathbb{C}^* = \mathfrak{R}^*(\mathbb{A},\mathbb{B})$ and the optimal fusion rule $\mathfrak{R}^*$ indicates how the two families will be fused. But, have we done fusion here? It depends on the performance. If

$$\varrho\left(\mathfrak{R}^*(\mathbb{A},\mathbb{B})\right) \geq \max\{\varrho(\mathbb{A}), \varrho(\mathbb{B})\}$$

then $\mathbb{C}^* = \mathfrak{R}^*(\mathbb{A},\mathbb{B})$ is THE optimal classification system family (with respect to ϱ.)

Definition 6: (Dual Set) Define the collection $\mathscr{S}^\star$ of real-valued functionals ρ defined on systems mapping outcomes from the measurable space $(\mathcal{E},\mathfrak{E})$ into the label set $\mathcal{L}$ to be

$$\mathscr{S}^\star = \{\rho : \mathscr{S} \to \mathbb{R}\}.$$

We call $\mathscr{S}^\star$ the *dual*[2] set of $\mathscr{S}$. We are interested in nonnegative functionals so we restrict this set further.

Definition 7: (Nonnegative Dual Set). Define the collection $\mathscr{S}^{\star+}$ to be the nonnegative, real-valued functionals ρ defined on systems mapping outcomes from the measurable space $(\mathcal{E},\mathfrak{E})$ into the label set $\mathcal{L}$ to be

$$\mathscr{S}^{\star+} = \{\rho : \mathscr{S} \to \mathbb{R}^+\}$$

where $\mathbb{R}^+ = \{r \in \mathbb{R} : r \geq 0\}$.

[2]We take care in using the word "dual set" here not to be confused with the "dual space" as founded in functional analysis. A dual space is a linear space of linear functionals over the field in use. We do not assume our functionals are linear since we do assume the sets have algebraic structure that make them linear spaces. If the label set was, in fact, a subfield of $\mathbb{R}$ then it would the same.

Definition 8: Define the collection $\mathscr{F}^{\star+}$ of nonnegative, real-valued functionals ϱ defined on classification system families mapping from the measurable spaces $(\mathcal{E},\mathfrak{E})$ into $\mathcal{L}$ to be

$$\mathscr{F}^{\star+} = \{\varrho : \mathscr{F} \to \mathbb{R}^+\}.$$

Now, suppose the performance $\varrho(\mathbb{A})$ is determined via the ROC curve $f_\mathbb{A}$ (or the ROC manifold), that is, assume there is a ROC functional φ that maps a ROC curve (or manifold) to a number (see [3]), then

$$\varrho(\mathbb{A}) = \varphi(f_\mathbb{A}). \tag{5}$$

Define the mapping $\digamma$ that takes a CSF $\mathbb{A}$ and outputs its ROC curve $f_\mathbb{A}$. Therefore, equation (5) can be written as

$$\varrho(\mathbb{A}) = \varphi(\digamma(\mathbb{A})).$$

Definition 9: (ROC functional) Define the collection $\mathscr{R}^{\star+}$ of nonnegative, real-valued functionals φ defined on ROC curves/manifolds to be

$$\mathscr{R}^{\star+} = \{\varphi : \mathscr{R} \to \mathbb{R}^+\}.$$

That is, given a ROC manifold $M_\mathbb{A}$ and a ROC functional φ then $\varphi(M_\mathbb{A})$ is a nonnegative real number. Since we wish this to be true for all CSFs, then we have

$$\varrho = \varphi \circ \digamma.$$

This equation tells us that given a ROC functional φ one can make the performance functional ϱ just by the composition with $\digamma$. Therefore, there is an induced mapping $\digamma^\bigstar$, called the *conjugate mapping* (that is conjugate to $\digamma$) defined by

$$\varrho = \digamma^\bigstar(\varphi) \equiv \varphi \circ \digamma.$$

Thus, $\digamma^\bigstar : \mathscr{R}^{\star+} \to \mathscr{F}^{\star+}$.

Now we can restate problem 1 as

Problem 2: Given a ROC functional φ, and given a set of label fusion rules $\mathsf{LABrules}$ we seek the best fusion rule $\mathfrak{R}^* \in \mathsf{LABrules}$ such that the performance

$$\varphi\left(\digamma\left(\mathfrak{R}^*(\mathbb{A},\mathbb{B})\right)\right) \geq \varphi\left(\digamma\left(\mathfrak{R}(\mathbb{A},\mathbb{B})\right)\right)$$

for all choices $\mathfrak{R} \in \mathsf{LABRules}$. That is,

$$\varphi\left(\digamma\left(\mathfrak{R}^*(\mathbb{A},\mathbb{B})\right)\right) = \max_{\mathfrak{R}\in\mathsf{LABRules}} \varphi\left(\digamma\left(\mathfrak{R}(\mathbb{A},\mathbb{B})\right)\right) \tag{6}$$

This may not look like we are gaining anything new until we look at the term $\digamma(\mathfrak{R}(\mathbb{A},\mathbb{B})) = f_{\mathfrak{R}(\mathbb{A},\mathbb{B})}$ (or $M_{\mathfrak{R}(\mathbb{A},\mathbb{B})}$). This is the ROC manifold of the fused system.

IV. Results

This section contains the main result that concerns determining a ROC functional φ given the performance functional ϱ.

Theorem 2: Given φ an $\mathscr{R}$-functional there exists an unique ϱ $\mathscr{F}$-functional given by

$$\varrho = \varphi \circ \digamma.$$

This defines the *conjugate mapping* $\digamma^\bigstar$ defined by

$$\digamma^\bigstar(\varphi) = \varphi \circ \digamma \qquad \text{for all} \quad \varphi \in \mathscr{R}^\star.$$

The domain of $\digamma^{\bigstar}$ is all of $\mathscr{R}^{\star}$, and $\digamma^{\bigstar}:\mathscr{R}^{\star}\rightarrow\mathscr{F}^{\star}$.

Proof: The proof of this theorem is straightforward. ■

Theorem 3: Given $\varphi\in\mathscr{R}^{\star+}$ there exists an unique $\varrho\in\mathscr{F}^{*+}$ given by the conjugate mapping

$$\digamma^{\bigstar}(\varphi)=\varphi\circ\digamma.$$

The domain of $\digamma^{\bigstar}$ is all of $\mathscr{R}^{\star+}$, and $\digamma^{\bigstar}:\mathscr{R}^{\star+}\rightarrow\mathscr{F}^{\star+}$.

Proof: Let $\varphi\in\mathscr{R}^{\star+}$ then $\varphi(M_{\mathbb{A}})\geq 0$ for every ROC manifold, $M_{\mathbb{A}}\in\mathscr{R}$. Then

$$\left[\digamma^{\bigstar}(\varphi)\right](\mathbb{A})=\left[\varphi\circ\digamma\right](\mathbb{A})=\varphi\left(\digamma\left(\mathbb{A}\right)\right)=\varphi\left(M_{\mathbb{A}}\right)\geq 0$$

for every $\mathbb{A}\in\mathscr{F}$. Therefore,

$$\digamma^{\bigstar}(\varphi)\in\mathscr{F}^{\star+}.$$

■

V. Examples

This section contains well-known examples of performance quantifiers of classifications system [7]. Let $\mathbf{A}$ denote a classification system and $\mathbb{A}$ denote a family of classification systems.

A. case $n=2$

Assume the label $\mathcal{L}=\{t,n\}$ for this subsection.

Example 1: **True Positive** (TP) also called the hit rate, recall, and sensitivity

$$\rho_{TP}(\mathbf{A})=P_{TP}(\mathbf{A})$$

Example 2: **True Negative** (TN) also called the correct rejection.

$$\rho_{TN}(\mathbf{A})=P_{TN}(\mathbf{A})$$

Example 3: **False Positive** (FP) also called false alarm, and Type I error

$$\rho_{FP}(\mathbf{A})=P_{FP}(\mathbf{A}).$$

Example 4: **False Negative** (FN) also called the Type II error

$$\rho_{FN}(\mathbf{A})=P_{FN}(\mathbf{A}).$$

Example 5: **Accuracy** (ACC)

$$\rho_{ACC}(\mathbf{A})=P_{TP}(\mathbf{A})+P_{TN}(\mathbf{A})$$

Example 6: **Specificity** (SPC)

$$\rho_{SPC}(\mathbf{A})=1-P_{FP}(\mathbf{A})$$

Example 7: **Positive Predictive Value** (PPV) also called precision

$$\rho_{PPV}(\mathbf{A})=\frac{P_{TP}(\mathbf{A})}{P_{TP}(\mathbf{A})+P_{FP}(\mathbf{A})}$$

Example 8: **Negative Predictive Value** (NPV)

$$\rho_{NPV}(\mathbf{A})=\frac{P_{TN}(\mathbf{A})}{P_{TN}(\mathbf{A})+P_{FN}(\mathbf{A})}$$

Example 9: **False Discovery Rate** (FDR)

$$\rho_{\ FDR}(\mathbf{A})=\frac{P_{FP}(\mathbf{A})}{P_{FP}(\mathbf{A})+P_{TP}(\mathbf{A})}$$

Example 10: **Matthews Correlation Coefficient** (MCC) [8], [9] is used in machine learning as a means to quantify the 2-class classification system $\mathbf{A}$. For brevity, let $t_p=P_{TP}(\mathbf{A})$, $t_n=P_{TN}(\mathbf{A})$, $f_p=P_{FP}(\mathbf{A})$ and $f_n=P_{FN}(\mathbf{A})$ then

$$MCC(\mathbf{A})=\frac{t_pt_n-f_pf_n}{\sqrt{\left(t_p+f_p\right)\left(t_p+f_n\right)\left(t_n+f_p\right)\left(t_n+f_n\right)}}.$$

If any of the four sums in the denominator is zero, the denominator can be arbitrarily set to one; this results in a Matthews Correlation Coefficient of zero, which can be shown to be the correct limiting value. It takes into account true and false positives and negatives and is generally regarded as a balanced quantifier which can be used even if the classes are of very different sizes. It returns a value between -1 and +1. A coefficient of +1 represents a perfect performance, 0 an average random performance and -1 the worst possible performance. That is, it quantifies the performance of the classification system $\mathbf{A}$ Since MCC can be negative, we add 1 to get a nonnegative performance functional

$$\rho_{MCC}(\mathbf{A})=MCC(\mathbf{A})+1.$$

By the disjunction equations we see that

$$MCC(\mathbf{A})=\frac{1-(f_p+f_n)}{\sqrt{1-(f_p-f_n)^2}}$$

so that

$$\rho_{MCC}(\mathbf{A})=\frac{1-(f_p+f_n)+\sqrt{1-(f_p-f_n)^2}}{\sqrt{1-(f_p-f_n)^2}}.$$

Example 11: Let $g(\xi,\eta)$ be a non-negative function for every $(\xi,\eta)\in[0,1]^2$ then consider the $\mathscr{S}$-functional

$$\rho_g(\mathbf{A})=g(P_{FP}(\mathbf{A}),P_{TP}(\mathbf{A}))$$

and the corresponding $\mathscr{F}$-functional is

$$\varrho_g(\mathbb{A})=\max_{\mathbf{A}\in\mathbb{A}}\rho(\mathbf{A})=\max_{\mathbf{A}\in\mathbb{A}}g(P_{FP}(\mathbf{A}),P_{TP}(\mathbf{A})).$$

All the previous examples have a well-defined choice of a function g.

Example 12: The area under the ROC curve is NOT a functional defined as a Riemann integral

$$\varrho_{AUC}(\mathbb{A})=\int_0^1 f_{\mathbb{A}}(p)\,dp.$$

This functional is NOT a $\mathscr{F}$-functional since

$$\varrho_{AUC}(\mathbb{A})\neq\max_{\mathbf{A}\in\mathbb{A}}\rho(\mathbf{A})$$

for any $\mathscr{S}$-functional ρ.

B. case $n>2$

Example 13: **Bayes Cost** (BC) Given error cost values $C_{i,j}\geq 0$, that is, the cost to make the i,j error ($\mathtt{C}_{i,i}=0$) then the Bayes cost for the system $\mathbf{A}$ is

$$\begin{aligned}\rho_{BC}(\mathbf{A}) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathtt{C}_{i,j}P\left(\mathcal{E}_j\right)P_{i|j}(\mathbf{A})\\ &= \mathtt{V}\odot\mathtt{C}\odot\mathtt{P}(\mathbf{A})\end{aligned}$$

when matrix $\mathtt{V} = \mathbf{v} \otimes \mathbf{1}^T$ where

$$\begin{aligned} \mathbf{v} &= (P(\mathcal{E}_1), P(\mathcal{E}_2), \ldots, P(\mathcal{E}_n)) \\ \mathbf{1} &= (1, 1, \ldots, 1) \end{aligned}$$

then

$$\varrho_{BC}(\mathbb{A}) = \max_{\mathbf{A} \in \mathbb{A}} \rho_{BC}(\mathbf{A}).$$

VI. Conclusions

There is a large collection of performance functionals to choose from. To evaluate an ATR system one should consider the performance criteria used. It might come down to analyzing the ROC curve/manifold. In order to evaluate the fusion of multiple system families, one needs to know the performance functional used (see equation (6).)

Acknowledgments

We acknowledge the financial support for this research from the Air Force Research laboratory, Sensors Directorate, Automatic Target Recognition Branch (AFRL/RYAT) at Wright-Patterson Air Force Base, Ohio; the Air Force Office of Scientific Research, Mathematics Directorate (AFOSR/MN) and the Air Combat Command (ACC/DRCA) at Langley Air Force Base, Virginia.

References

[1] S. I. Resnick, *A Probability Path*. Boston MA: Birkhauser, 2005.

[2] C. M. Schubert, "Quantifying correlation and its effects on system performance in classifier fusion," Ph.D. dissertation, AFIT/DS/ENC/05-01 Air Force Institute of Technology, Wright-Patterson AFB OH, 2005.

[3] S. Thorsen and M. Oxley, "A description of competing fusion systems," *Information Fusion Journal*, vol. 7, pp. 346–360, 2006, doi:10.1016/j.inffus.2006.03.004.

[4] J. P. Egan, *Signal Detection Theory and ROC Analysis*. New York: Academic Press, 1975.

[5] D. G. Luenberger, *Optimization by Vector Space Methods*, ser. Wiley Professional Paperback Series. New York: John Wiley and Sons, Inc, 1969.

[6] H. Schubert, *Topology*. Allyn and Bacon, Inc., 1968.

[7] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," HP Laboratories, Palo Alto CA, Technical report, 2004.

[8] B. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochim. Biophys. Acta*, vol. 405, pp. 442–451, 1975.

[9] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, pp. 412–424, 2000.

The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the US Government.